Министерство образования и науки Российской Федерации
Московский государственный технический университет им.Н.Э.Баумана
Кафедра «Системы обработки информации и управления»

ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ

Сборник учебно-методических работ и статей аспирантов и студентов

Выпуск 8

Под редакцией Ю.Н.Филипповича

НОК «CLAIM»
Москва
2006

УДК 007.52
ББК 32.813
И 76

Составитель: Ю.Н.Филиппович
Авторы статей:
А.А.Александрова, Е.М.Буркова, Ван Чэнь, Гуань Ли, И.С.Кесель, С.А.Кесель, А.Б.Кирнарский, Е.П.Коннова, А.А.Кононков, Ле Ба Хонг Минь, И.М.Нейский, Д.А.Никитченко, М.А.Павлова, А.И.Панченко, М.Е. Постников, А.А. Проскурнин, Б.И.Рабинович, Е.В.Ртищева, А.А.Самочетов, А.М.Серебряков. А.В.Сиренко, А.Ю.Филиппович, Анна Филиппович, Ю.Н.Филиппович, Г.А.Черкасова.

И 76 **Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. Выпуск 8** / Сост. и ред. Ю.Н.Филипповича. — М.: НОК «CLAIM», 2006. — 326 с.

Сборник содержит 18 научных статей аспирантов и студентов. Статьи посвящены: исследованию и разработке информационных технологий в области образования, ERP-системам и интеллектуализации бизнес-процессов, компьютерной лексикографии и семиотики, проектированию баз знаний.

Сборник предназначен для преподавателей, аспирантов и студентов специальностей двух направлений: 230100 — Информатика и вычислительная техника, 230200 — Информационные системы.

Содержание

МЕТОДИЧЕСКИЕ РАБОТЫ

Предисловие

В восьмом выпуске сборника представлены научные статьи преподавателей, аспирантов и студентов кафедры «Системы обработки информации и управления (ИУ5) Московского государственного технического университета им. Н.Э.Баумана.

Редакторская статья посвящена рассмотрению информационной технологии поддержки нового класса экспериментальных исследований языкового сознания — когнитивным экспериментам, одной из целей которых является создание лингвокультурного тезауруса (ЛКТ).

Начало исследованиям было положено несколько лет тому назад и связано с научными сообщениями и публикациями в 2002–2004 гг. чл.корр. РАН, д.ф.н., профессора Ю.Н.Караулова. Разработки ЛКТ в 2003–2004 гг. проводились в Московском государственном лингвистическом университете по Федеральной целевой программе «Русский язык» на 2002-2005 годы», а с 2005 по н.в. по гранту РФФИ 05-06-80284 «Языковое сознание нашего современника: когнитивная структура и лингвокультурное содержание».

В статье рассматриваются три аспекта информационной технологии, условно названные «создание», «функционирование» и «существование».

Анализируются состав и принципы концептуальной модели когнитивного эксперимента, лежащего в основе построения ЛКТ. Последовательно рассматривается исходная и расширенная концептуальные модели, а также возможные направления дальнейших ее расширений. Описывается архитектура автоматизированной системы поддержки эксперимента, основные процедуры планирования эксперимента и проектирование автоматизированной системы обработки экспериментальных данных. Подробно приведены описания ввода и первичной обработки данных, состава и структуры базы данных когнитивного эксперимента. Описаны процедуры получения различных проекций базы данных, операции формальной и семантической верификации экспериментальных данных.

Статьи аспирантов и студентов посвящены исследованию и разработке информационных технологий и конкретных систем.

Несколько статей сборника продолжают тему редакторской статьи. Так в статье А.А.Проскурнина рассмотрена автоматизированная система контроля знаний, на основе когнемной модели Ю.Н.Караулова и однопараметрической модели обработки результатов педагогических экспериментов Раша. А.В.Сиренко подробно описывает рабочий вариант базы данных лингвокультурного тезауруса русского языка.

В статьях А.Ю.Филипповича с А.Б.Кирнарским и А.М.Серебряковым рассматриваются вопросы моделирования вербального сознания: практическая реализация web-ориентированных интерактивных ассоциативных экспериментов и пример возможного использования генетических алгоритмов в психолингвистических исследованиях. Данные разработки ведутся по гранту РГНФ № 06-04-03803в «Автоматизированная система научных исследований динамики ассоциативно-вербальной модели языкового сознания русских как индикатора образа России в новейшей истории и современности» (руководитель А.Ю.Филиппович).

Статьи И.С.Кеселя и Е.В.Ртищевой посвящены разработкам обучающих систем. Они непосредственно связаны с учебными курсами, которые ведутся для студентов кафедры ИУ-5.

В сборнике представлено несколько статей-обзоров, раскрывающих предмет будущих исследований и разработок их авторов: А.А.Александровой А.А.Кононкова, И.М.Нейского, М.А.Павловой, А.А.Самочетова.

Традиционными для сборника являются публикации посвященные интеллектуализации процессов переработки информации и принятия решений для практических систем. К ним относятся работы Е.М.Бурковой, Е.П.Конновой, Д.А.Никитченко, М.Е.Постникова, Б.И.Рабиновича.

Статья Ван Чэня и Гуань Ли содержит фрагмент возможного будущего русско-китайского словаря метафор информационных технологий.

В статье Ле ба Хонг Миня описан один из алгоритмов системы русско-вьетнамского машинного перевода, разработка которой им ведется в рамках магистерсгого исследования.

Статья Анны Филиппович посвящена исследованиям особенностей и разработке элементов шрифтов кон. XVIII– нач. XIX вв., использованных при печати многих научных изданий этого исторического периода. Работа выполняется в рамках гранта РГНФ № 06-04-12412в «Интегрированная инструментальная информационно-программная среда для автоматизации исследований Словаря академии российской 1789–1794 гг.».

В сборник включены две методические разработки — рекомендации для студентов к выполнению лабораторных работ по дисциплине «Защита информации в АСОИУ» и курсового проекта по дисциплине «Семиотика информационных технологий».

Благодарю за помощь в подготовке и выпуске сборника Г.А.Черкасову и А.Л.Волкова.

Ю.Н.Филиппович

Редакторская статья

Ю.Н.Филиппович

АСПЕКТЫ ИНФОРМАЦИОННОЙ ТЕХНОЛОГИИ КОГНИТИВНОГО ЭКСПЕРИМЕНТА СОЗДАНИЯ ЛИНГВОКУЛЬТУРНОГО ТЕЗАУРУСА

Введение

Статья посвящена рассмотрению одной из компонент проекта РФФИ «Языковое сознание нашего современника: когнитивная структура и лингвокультурное содержание» (РФФИ 05-06-80284). Проект направлен на решение фундаментальной проблемы, касающейся комплекса наук о человеке, — исследование структуры, содержания и функционирования языкового сознания. В рамках проблемы поставлена конкретная задача — моделирование языкового сознания на основе анализа, не изучавшегося ранее знакопорождающего режима его работы, в противовес хорошо изученному смыслопорождающему режиму. Теоретические положения предложенного подхода включают рассмотрение взаимоотношений семантического и когнитивного уровней анализа, обоснование введения элементарной единицы знания, разработку процедур перехода от семантического уровня на когнитивный: от языковых единиц — к единицам знания. Экспериментальная часть строится как массовый психолингвистический эксперимент с носителями языка-культуры, в котором испытуемым предлагается по некоторому заданному смыслу установить знак, отвечающий этому смыслу. Материалы этого уникального эксперимента подвергаются специальной лингвокогнитивной обработке, в результате чего получается совокупность когнитивных единиц, которые формируются в базу знаний современного усредненного носителя русского языка и культуры. Работа с базой знаний позволяет установить метрику пространства языкового сознания и его

содержание, характеризующее реальный и виртуальный миры, в которых живет современный человек. [Караулов 2003а-б, 2004а-в]

В проводимом научном исследовании используются два понимания информационной технологии: «широкое» и «узкое». «Широкое» понимание информационной технологии, как способа существования мыслящей материи, используется в суждениях при моделировании вербального сознания носителя русского языка-культуры. «Узкое» понимание информационной технологии, как автоматизированной компьютерной системы поддержки деятельности исследователей, используется при практической реализации экспериментальных исследований. Связь этих пониманий находит свое выражение в компьютерной форме существования информационной технологии лингвокультурного тезауруса (ИТ ЛКТ) в виде информационно-программной модели вербальной картины мира носителя языка культуры.

В связи с этим отметим две особенности ИТ ЛКТ, как лингвистического научного исследования (эксперимента): а) в его основу положены принципы лингвистического конструирования; б) форма существования ЛКТ и информационная технология его получения в основе своей лексикографические.

Лингвистическое конструирование как понятие было введено в научный оборот в 70-х годах XX века. Наиболее полно оно представлено в работе [Караулов 1981] и понимается как особое направление исследовательской деятельности, целью которого является построение новых лингвистических объектов с заданными свойствами. “Лингвистическое конструирование — это совокупность обобщенных способов и приемов компиляции и комбинирования “образцов решения проблем”, экстраполяции уже имеющихся, готовых теоретических и практических результатов, полученных в разных областях лингвистики и их прямого или эвристического использования для преодоления трудностей и решения проблем, возникающих в тех или других областях при построении новых лингвистических объектов” [Караулов1981. С.16].

Лексикографическая технология сегодня очень распространена практически во всех областях научной деятельности. Из инструментария языкознания, навыков и умений конкретной группы ученых, она сегодня превратилась в необходимую и обязательную составляющую любого научного исследования: во-первых, повсеместной практикой стало представление результатов исследований, экспериментов и вообще освоения некой предметной области реального мира в лексикографической форме, в виде словаря, справочника, энциклопедии и т.п.; во-вторых, результат научного исследования стал представляться одновре-

менно в двух формах — в виде основного текстового описания (представления знания) и дополнительного лексикографического метаописания (представления метазнания).

Необходимость включения в учебные и научные печатные, а теперь и электронные издания (а именно в таких формах чаще представляются результаты научных исследований или экспериментов) так называемого «научного аппарата» является стандартным требованием. Современный «научный аппарат издания» — это, прежде всего, различные указатели слов и понятий, глоссарии, словари (определений, частотные, переводные и др.) и фрагменты тезаурусов (понятийных, поисковых). В совокупности эти лексикографические компоненты «научного аппарата издания» придают результату научного исследования новые дополнительные дидактические и металогические свойства.

Применение лексикографического инструментария в научном исследовании как основного или дополнительного требует определенных ресурсных затрат, которые в общем случае только частично могут быть компенсированы достижением новых качественных показателей результата исследования.

Современный лексикографический инструментарий научных исследований и экспериментов — это компьютерные методы и средства лексикографической технологии, предназначенные для поддержки проведения исследований, и обеспечения существования его результатов.

Резюмируя приведенные суждения, выделим три основных аспекта ИТ ЛКТ: 1) аспект *создания* ЛКТ — разработка средств поддержки когнитивного эксперимента и оформление его результатов; 2) аспект *функционирования* ЛКТ — конструирование модели (моделирование) вербального сознания носителей языка культуры; 3) аспект *существования* ЛКТ — измерение параметров (характеристик) ЛКТ и установление возможных форм его существования.

Эти три аспекта являются отражением трех составных частей проводимого исследования: эмпирического эксперимента, теоретического анализа его результатов и создание форм их представления. Рассмотрим более подробно названные аспекты.

Аспект «*Создание*» ИТ ЛКТ

Информационная технология создания ЛКТ русского языка представляет собой совокупность методов и приемов использования современного программного обеспечения ЭВМ для автоматизации научных исследований. Она реализована как интеграция типовых инструментальных информационных технологий и специализированной лингвистической тех-

нологии. Сама методика интеграции (последовательность использования функциональных средств поддержки методов обработки данных, представленных в применяемых компьютерных программах) традиционна и содержит этапы: 1) ввода и первичной обработки исходных данных; 2) формирование базы данных исследования; 3) получение обобщающих и выборочных данных; 4) лингвистическое конструирование – получение частотных словников, индексов и т.п. [Филиппович Ю. 2002 в]

Концептуальная модель когнитивного эксперимента

Сообразно уже приведенному «широкому» и «узкому» пониманию информационной технологии, когнитивный эксперимент будем понимать как некоторый конкретный вариант ее реализации.

Тогда в «широком» понимании когнитивный эксперимент — это конкретная информационная технология восприятия, осознания и представления реального мира субъектом в заранее определенной форме. Такой формой и объектом эксперимента являются данные результатов языковой игры «Кроссворд».

В «узком» понимании когнитивный эксперимент — автоматизированная компьютерная система поддержки конкретной деятельности субъектов-экспериментаторов. Такой деятельностью является сбор, обработка и представление данных эксперимента. Класс подобных систем получил название «систем извлечения знаний».

Определим когнитивный эксперимент как информационную технологию извлечения знаний о реальном мире, имеющихся у людей носителей языка культуры.

Традиционно при разработке информационных технологий первый этап проектирования называют концептуальным, подчеркивая тем самым то, что основное внимание на данном этапе проектирования сосредотачивается на тех понятиях, которые определяют существо предметной области. При этом концепты вводятся императивно и декларативно [Филиппович, 2001а, б].

Исходная концептуальная модель

Определим исходную концептуальную модель когнитивного эксперимента в следующем составе: субъект эксперимента — экспериментатор; объекты эксперимента — кроссворды, источники кроссвордов, карточка когнемы, картотека когнем; процессы эксперимента — поиск кроссворда, формирование когнем, сбор данных.

Экспериментатором в когнитивном эксперименте является Ю.Н.Караулов автор идеи эксперимента как способа создания ЛКТ и представления вербального знания носителя языка культуры в форме

связанных между собой пятикомпонентных отношений (пентаграмм), получивших название когнем.

Кроссворд — распространенная языковая игра, состоящая в нахождении однословных ответов на вопросы, при известном количестве букв в слове, а также, по мере увеличения количества ответов, при некоторых известных буквах слова. Игра заканчивается, если разгаданы (найдены) слова-ответы на все вопросы. Во время игры заполняется графическая форма кроссворда — специальная таблица, слова-ответы записываются по строкам и столбцам таблицы, в одной ячейке таблицы записывается одна буква слова. Буквы в ячейках на пересечении строк и столбцов принадлежат обоим словам и являются одинаковыми.

Источниками кроссвордов в исходной модели эксперимента являются подшивки газет за некоторый период.

Карточка когнемы — это типового размера (библиографическая карточка) лист плотной бумаги с основными и дополнительными рукописными записями экспериментатора. Основные записи – это следующие компоненты пентаграммы (когнемы): 1) формула смысла, 2) способ представления смысла, 3) слово-знак, 4) область референции, 5) функция. Дополнительная запись представляет собой некоторый комментарий или замечание исследователя, а в некоторых случаях возможные варианты значений одной или нескольких основных записей. Записи на карточке «формула смысла» и «слово-знак», представляют собой выписки из кроссворда вопроса и ответа соответственно. Все остальные записи внесены в карточку экспериментатором.

Картотека когнем является упорядоченным собранием карточек, размещенных в ящиках. Количество карточек в картотеке более 10 тысяч.

Поиск кроссворда — это процесс, состоящий из процедур и операций выбора источника кроссворда, который состоял из двух номеров газет, содержащих вопросы кроссворда (один номер) и ответы (другой номер).

Формирование когнем является наиболее сложным процессом информационной технологии эксперимента и представляет собой формальное переписывание на карточку вопросов и ответов кроссворда («формула смысла» и «слово-знак») и неформализованную интеллектуальную деятельность экспериментатора, внешнее проявление которой состоит в занесении основных («способ представления смысла», «область референции» и «функция») и дополнительных записей на карточку.

Сбор данных в основном представляет собой следующие процедуры: а) упорядочение рукописных карточек по значению записи «область референции», б) разделение упорядоченных групп карточек специальной карточкой-разделителем и в) размещение их в ящики хранения.

Расширенная концептуальная модель

Причинами расширения концептуальной модели когнитивного эксперимента являются ресурсные ограничения, связанные с необходимостью и возможностями его развития, а также предварительный анализ полученных данных. Расширенная модель по существу является моделью автоматизированного эксперимента.

Определим расширенную концептуальную модель когнитивного эксперимента в следующем составе: субъекты эксперимента — экспериментатор, разработчик, анкетеры; объекты эксперимента — кроссворды, источники кроссвордов, карточка когнемы, картотека когнем, анкета, текстовый файл картотеки, база данных когнем, документация эксперимента; процессы эксперимента — поиск кроссворда, формирование когнемы, сбор данных, ввод и первичная обработка данных, создание базы данных.

Разработчики — это участники эксперимента, задачами которых являются: разработка средств автоматизации эксперимента; разработка и реализация автоматизированной информационной технологии сбора, обработки, представления данных эксперимента и последующего их анализа. Количество разработчиков в расширенной модели — три человека.

Анкетеры — участники эксперимента, появление которых оказалось следствием необходимости увеличения объема базы данных ЛКТ. В качестве анкетеров были привлечены студенты пятого курса МГТУ им.Н.Э.Баумана, выполнявшие задание на самостоятельную работу при изучении дисциплин «Семиотика информационных технологий» и «Лингвистическое обеспечение автоматизированных систем обработки информации и управления». Общее количество анкетеров составило 36 человек[1].

[1] **Задание анкетеров**, которое выполняли студенты получило название «Исследование когнитивного тезауруса». При его выполнении решались следующие **задачи**: 1) подбор и подготовка материалов для исследования; 2) выявление (установление названия) референтной области предмета вопроса; 3) определение способа выявления референтной области предмета вопроса; 4) определение функции когнитивных единиц тезауруса. В качестве **формы отчетности** студентам предлагалось составить таблицу когнитивных единиц тезауруса. Студентам-анкетерам даны были следующие *методические указания* ***по выполнению задач:***

Анкета — форма отчета анкетера о выполнении задания, представляющая собой текстовый файл таблицы, строки которой являются когнемами, а столбцы — ее компонентами: 1) формула смысла, 2) способ представления смысла, 3) слово-знак, 4) область референции, 5) функция. Кроме этого, каждая анкета содержала сведения об источнике кроссворда.

Текстовый файл картотеки — это текстовый файл сортированной таблицы, строки которой являются когнемами, а столбцы — ее компонентами.

База данных когнем — совокупность компьютерных файлов, содержащих данные когнитивного эксперимента, созданная средствами СУБД Borland Paradox v.5.0.

Документация эксперимента — совокупность методических материалов эксперимента.

Процессы расширенной концептуальной модели эксперимента хотя и обозначены также как в исходной модели, но имеют несколько иное содержание. Так процесс *Поиск кроссворда* для анкетеров является расширенным и направлен на другие источники — публикации в специальных сборниках и в Интернет. Использование в качестве источника кроссвордов Интернет существенно сокращает затраты времени на формирование когнем, в части переписывания вопросов и ответов. Фактически процесс *Формирование когнем* в этом случае представляет собой последовательность операций по «скачиванию» с сайта источника кроссворда текста вопросов и ответов и последующему его форматированию (приведению к табличной форме). Иначе выглядит и *Сбор данных*. Его содержание — это объединение файлов анкет (объединение таблиц когнем), полученных анкетером.

Задача 1. В качестве материалов для исследования когнитивного тезауруса носителя русского языка следует использовать опубликованные в печатной форме кроссворды (вопросы и ответы на них). Объем экспериментального материала: минимальное количество кроссвордов – 5; минимальное общее количество вопросов – 200. Сформировать таблицу когнитивных единиц тезауруса, состоящую из семи полей: <номер кроссворда>, <номер вопроса>, <вопрос>, <ответ>, <область>, <способ>, <функция>. Подобранные пары <вопрос= «формула смысла»>-<ответ=знак (слово)> необходимо записать в таблицу. Необходимо указать источник материалов.

Задачи 2-4. Для успешного решения задач 2-4 необходимо ознакомиться с методикой построения когнитивных единиц тезауруса чл.корр РАН Ю.Н.Караулова (изучить лекционные материалы – прочитать статьи). После ознакомления с методикой последовательно выполнить задачи 2, 3, 4. Результаты выполнения задач занести в таблицу когнитивных единиц тезауруса. В файле *Отношения фигур знания.doc* представлены возможные наименования референтных областей, способов их выявления и функции.. Результаты представляются в печатной и электронной (в виде файла MS Word) формах.

Ввод и первичная обработка данных сложный процесс, состоящий из процедур и операций клавиатурного ввода рукописных карточек когнем, формирования файлов таблиц когнем, группировки и объединении файлов, вычитки и редактирования текстовых полей таблиц, сортировки таблиц и др.

Возможные направления дальнейших расширений концептуальной модели

Основной причиной возможных дальнейших расширений могут явиться обнаруженные особенности собранных данных, а также результаты их анализа. Назовем несколько возможных направлений.

1. Включение в концептуальную модель эксперимента таких объектов и процессов как: *Источники данных для составления кроссвордов*, *Средства поддержки составления кроссвордов*, *Составление кроссворда* и др. Основанием для данного направления является давние исследования и работы в области «автоматического» составления кроссвордов. В качестве примера можно привести публикацию 1986 года [Каганов 1986]. Основой для «автоматического» составления кроссворда является словарь определений, составленный по каким-то правилам отбора лексики из уже существующих словарей. Важным является как определение множества словарей-источников, так и правил, лежащих в основе составления кроссворда: тематических преференций, языковых форм, частотных и других количественных характеристик слов.

2. Увеличение объема эксперимента (количества когнем) и усложнение технологии сбора и обработки данных с целью обеспечения их репрезентативности и устранения различных статистических «перекосов». Представительность данных, собранных в результате проведения когнитивного эксперимента, для многих потенциальных выводов по результатам их анализа является важным аргументом, в связи с этим увеличение базы эксперимента объективно положительный фактор. Вместе с тем следует отметить: 1) материал исследования «игровой», т.е. разгадывание кроссворда ориентировано на «приятное потешное времяпрепровождение» в течение ограниченного времени — «часа»; 2) собранный материал очень вероятно будет подчиняться гиперболическим частотным законам. Эти два обстоятельства позволяют предположить быстрое исчерпание слов-знаков в когнемах, на основе которых строится лингвокультурный тезаурус. В связи с этим полнота некоторых его областей окажется недостаточной. Необходимым может оказаться многоступенчатый, гнездовой по концептосферам и расслоенный по референтным областям направленный доотбор данных.

3. Включение в концептуальную модель эксперимента объектов и процессов, представляющих результаты исследований: *Проекции базы данных, Лингвистические процессоры, Печатные и Электронные версии лингвокультурного тезауруса* и др.

4. Включение в концептуальную модель эксперимента *Баз данных словарей* (толковых, терминологических, переводных, синонимов, омонимов и т.п.), *Базы данных русского ассоциативного словаря.*

Автоматизированная система поддержки эксперимента

Одним из наиболее распространенных подходов к автоматизации проведения экспериментов является подход, основанный на двух традициях: а) выделять в эксперименте его временные стадии и б) создавать автоматизированную систему поддержки эксперимента перманентно (непрерывно) по мере его развития во времени.

Основными стадиями когнитивного эксперимента являются подготовка, проведение, анализ результатов и завершение. Рассмотрим некоторые наиболее важные компоненты выделенных стадий эксперимента.

Планирование эксперимента

Планирование когнитивного эксперимента осуществляется на стадии его подготовки и состоит в выработке его целей, исследовательских гипотез, конкретных задач, критериев достижения целей и качества решения задач, а также оценки ресурсных затрат.

Главной целью когнитивного эксперимента является создание русского лингвокультурного тезауруса на основе когнем. Данная цель является корневой вершиной «дерева целей и задач эксперимента», достижение которой означает завершение эксперимента. В связи с этим следует заметить: во-первых, так поставленная цель эксперимента достижима только частично, т.е. только в пределах ресурсных ограничений; во-вторых, достигнутая цель эксперимента, его результат — лингвокультурный тезаурус — будет являться конкретным эмпирическим вариантом (экспериментальным образцом) теоретически возможных тезаурусов; в-третьих, лингвокультурный тезаурус, создаваемый в результате эксперимента, по определению являющийся моделью вербального сознания носителя русского языка культуры, в практическом своем воплощении является некоторым «действующим макетом», на котором возможными являются исследования и разработки фундаментального и прикладного характера.

Приведенные замечания требуют формирования сложной системы связанных *критериев* для оценки достижения главной цели и подцелей эксперимента. Основными критериями будут являться:

Во-первых, *репрезентативность* собранных данных. В связи с этим, особенностью когнитивного эксперимента является его отнесенность к классу выборочных статистических экспериментов и необходимость разработки соответствующих процедур и операций.

Во-вторых, «*полнота»* представления теоретически возможного тезауруса. Данный критерий комплексный и включает логико-лингвистические и лингво-статистические оценки полноты представления компонент когнем (формул смысла, способов, знаков, функций и референтных областей), степени их связанности и структурности.

В-третьих, «*сходимость»* процесса макетирования вербального сознания. Та особенность когнитивного эксперимента, что в его основе лежит языковая игра типа «кроссворд», необходимо требует исчисления темпа извлечения новых данных (знаний) из экспериментального материала. Критерий сходимости косвенно позволяет оценить полноту словника эксперимента, функциональность, способность формирующегося в процессе эксперимента тезауруса как действующего макета вербального сознания носителя языка культуры.

Проектирование автоматизированной системы

Автоматизированная система поддержки эксперимента относится к классу систем обработки информации и управления, проектирование функциональности которых для различных приложений направлено на следующие основные цели: преобразование исходных данных к виду, позволяющему их компьютерное хранение и обработку (функция «*input*»); организованное компьютерное хранение (функция «*memory*»); компьютерную обработку, состоящую в реорганизации хранения и преобразовании к виду, позволяющему их представить в форме отличной от исходной (функция «*transform=interpret*»); преобразование обработанных данных к виду удобному для дальнейшего использования (функция «*output*»). При совместной реализации перечисленных функций, возникает необходимость координации их во времени (функция «*control*»).

Исходными данными эксперимента являются: карточки когнем, организованные в картотеку; данные кроссвордов в электронной форме; документация эксперимента (методические материалы, инструкции, программные документы, распечатки хранимых и обработанных данных и др.).

Преобразование исходных данных к виду, позволяющему их компьютерное хранение и обработку является целью информационной технологии *ввода и первичной обработки данных*. Особенность ее проектирования для эксперимента состоит в определении таких конечных тексто-

вых форм, которые могли бы использоваться на этапе создания баз данных и одновременно для автоматической и «ручной» их верификации.

Хранение экспериментальных данных организовано в виде компьютерной *базы данных,* проектирование которой для задач этого класса состоит в применении хорошо известных профессиональных умений и навыков работы с программными средствами типа СУБД. Объем экспериментальных данных (до 100 тыс. когнем-записей), их структура (до 10 полей в записи) и принимаемые значения (все поля текстовые до 256 символов) позволяют создать базу данных эксперимента с одинаковой эффективностью практически во всех известных СУБД.

Для компьютерной обработки экспериментальных данных когнитивного эксперимента могут использоваться как типовые информационные технологии, так и специальные лингвистические, а также могут быть разработаны специализированные алгоритмы и программы непосредственно ориентированные на цели и задачи эксперимента. Все эти средства являются составными частями *программного обеспечения эксперимента.*

Типовыми инструментальными информационными технологиями обработки символьной информации и соответствующими программными изделиями, которые использовались для создания ЛКТ, являются: технология обработки текстов (MS Word), технология электронных таблиц (MS Excel), технология баз данных (Borland Paradox v.5.0).

MS Word и MS Excel использовались для ввода исходных данных, их группировки, орфографического контроля, подсчета некоторых количественных параметров, сортировки и представления в печатной форме для визуального контроля и анализа.

СУБД Borland Paradox v.5.0 использовалась для создания и ведения базы данных исследования, формирования выборочных данных по запросам, сортировки данных и их компаративного анализа.

Специальная лингвистическая технология использовалась для получения словников и была реализована с помощью программы Andrew Tools [Филиппович А. 2002].

Специализированных программных средств для обработки материалов эксперимента на первых его стадиях не разрабатывалось. Однако очевидна потребность в них для обработки данных на стадиях их анализа. К числу таких программ следует отнести: лингвистические процессоры когнайзера, работающие в активном и пассивном режимах; программы обеспечивающие взаимодействие базы данных эксперимента с базами данных словарей; программы обеспечивающие создание и использование печатных и электронных форм существования ЛКТ.

Особенностью *форм представления результатов когнитивного эксперимента* и процессов их проектирования является ориентация на пользователя «непрограммиста». Результаты эксперимента предназначены, прежде всего, для ученых гуманитарных специальностей, для которых традиционные лексикографические формы являются предпочтительными. В связи с этим наибольшего эффекта от последующего использования обработанных экспериментальных данных можно достигнуть, представляя их в виде проекций базы данных. Проекции — это словники и комбинированные сортированные списки значений полей базы данных. При их формировании важно соблюдать сложившиеся традиционные требования к печатным изданиям словников, конкордансов, словарей и т.д.

Словники и списки должны отвечать требованиям наглядности и удобства внесения изменений, т.е. одновременно являться и «входными» и «выходными» вариантами документов автоматизированной технологии обработки данных эксперимента. Фактически первый вариант представления результатов эксперимента — это материал для верификации собранных данных, а последний — материал для анализа и последующих исследований.

Ввод и первичная обработка данных

Исходными данными для создания ЛКТ являлись рукописные карточки, содержащие пять основных и одну дополнительную записи исследователя.

Формула смысла – данная запись представляет собой одно предложение (очень редко - несколько), включающее 5-10 словоформ.

Способ представления смысла – практически всегда это авторская сокращенная запись полного слова. Для одинаковых слов используются несколько вариантов сокращений. Разных способов – 47.

1.	дескрипция
2.	дефиниция
3.	перифраза
4.	фрейм
5.	прецедентный текст
6.	множество
7.	метафора
8.	суждение
9.	синоним
10.	словосочетание

11	смена кода
12	импликация
13	фразеол. единица
14	символ
15	омоним
16	деталь биографии
17	гипероним
18	игра слов
19	внутренняя форма
20	афоризм

21.	гипоним
22.	поговорка
23.	образ
24.	сравнение
25.	ассоциация
26.	пословица
27.	предикат
28.	антоним
29.	загадка
30.	термин
31.	оксюморон
32.	рифма
33.	непрецедентный текст
34.	псевдоним
35.	шарада
36.	гипербола
37.	марка
38.	аббревиатура
39.	анаграмма
40.	жест
41.	вид-род
42.	пароним
43.	эвфемизм
44.	полисемия
45.	Часть-целое
46.	конверсив
47.	литота

Слово-знак – это слово или слово, заключенное в кавычки. Общее количество разных слов составило 5376.

Область референции или наименование предметной области представляет собой в основном одно слов или короткое словосочетание, в исследовании выделено 122 наименований областей.

1.	авиация
2.	археология
3.	архитектура
4.	астрономия
5.	балет
6.	биология
7.	ботаника
8.	быт
9.	взаимоотношения людей
10.	внешность человека
11.	водоем
12.	война
13.	время
14.	гармония, красота
15.	география
16.	геология
17.	геометрия
18.	город
19.	государственное устройство
20.	движение
21.	деньги
22.	дети
23.	документы
24.	должность
25.	дорога
26.	живопись
27.	жизнь замечательных людей
28.	зоология
29.	игра
30.	изобретательство
31.	инструменты
32.	интеллект
33.	информация
34.	искусство
35.	история
36.	календарь
37.	катастрофа
38.	кино, телевидение

39.	книга
40.	количество
41.	корабль
42.	космонавтика
43.	криминал
44.	лес
45.	литература
46.	лошадь
47.	марка
48.	материал
49.	мед
50.	медицина
51.	мера
52.	места памяти
53.	металл
54.	минералогия
55.	мифология
56.	мода
57.	мораль
58.	Москва
59.	музыка
60.	награда
61.	наркотик
62.	наука
63.	образ жизни
64.	отдых
65.	отношения международные
66.	отношения социально-трудовые
67.	отношения финансовые
68.	отношения экономические
69.	охота
70.	победа
71.	поведение
72.	политика
73.	поп-арт
74.	похожесть
75.	правосудие

76.	праздник
77.	пресса
78.	приметы
79.	природа
80.	производство
81.	промысел
82.	пространство
83.	профессия
84.	психология
85.	равенство
86.	разрушение
87.	река
88.	религия
89.	секс
90.	сельское хозяйство
91.	семья
92.	скульптура
93.	смерть
94.	собака
95.	спорт
96.	строение
97.	строительство
98.	творчество
99.	театр
100.	телевидение
101.	техника
102.	тождество
103.	торговля
104.	транспорт
105.	туризм
106.	успех
107.	физика
108.	физиология
109.	философия
110.	фольклор
111.	фотография
112.	характер человека
113.	химия

114.	хобби
115.	цирк
116.	чувства
117.	шоу
118.	энергетика
119.	этнография
120.	ювелирные изделия
121.	юмор
122.	язык

Запись *функция* представлена двумя сокращениями: ртш. — ретушь, и рцп. — рецепт.

Все собрание карточек представляет собой картотеку, упорядоченную по значениям записи *Область,* наименование которых вынесено на специальную карточку-разделитель.

Ввод картотеки в ЭВМ осуществлялся с использованием стандартной клавиатуры с использованием программы MS Word. Текстовый файл, сформированный после ввода, содержал таблицу, каждая строка которой соответствовала одной введенной карточке. Столбцами таблицы являлись записи карточки, обозначенные как: № — номер карточки в порядке ее введения; *поле1…поле5* — основные записи; *примечание* – дополнительная запись, * — помета оператора ввода данных. Оператор мог сделать следующие пометы при вводе данных: * — есть неопределенность в данных, ** — есть несловесные примечания, *** — не разобрал почерк. Имя сформированного текстового файла определялось, исходя из названия области, и бралось с карточки-разделителя.

Ввод карточек осуществлялся в пять этапов. Общее количество введенных карточек составило более 8000 строк в 156 файлах. В дальнейшем файлы группировались и создавались объединенные таблицы.

Основными операциями с текстовыми файлами являлись:

1. Автоматическая проверка правописания;
2. Автоматическая замена «синонимичных» сокращений;
3. Ручная замена наименований областей.

После выполнения данных операций таблицы преобразовывались в текст с разделителями (символ табуляции) и сохранялись в формате *.txt (текст DOS) для последующего импорта в СУБД Paradox.

База данных когнитивного эксперимента

Основной целью создания базы данных ЛКТ является интегрированное и непротиворечивое хранение данных когнитивного эксперимента. Дополнительно следует выделить цели, достижение которых выходит за рамки эксперимента и возможно в процессе дальнейших исследовательских работ и практического использования достигнутых результатов. К ним относятся: 1) создание полного лингвокультурного словаря-тезауруса русского языка; 2) разработка информационной системы словаря-тезауруса для использования в других научных исследо-

ваниях и учебно-методической работе; 3) интеграция с данными других исследований, в частности с Русским ассоциативным тезаурусом; 4) более глубокие научные исследования языкового сознания носителя языка-культуры и его моделирования в информационных системах для поддержки коммуникативных процессов.

База данных тезауруса сформирована стандартной процедурой импорта текстовых данных СУБД Paradox. В результате ее выполнения была получена таблица, включающая семь полей: 1–5 поля — основные записи карточек, поле 6 – Примечания, поле 7 – пометы оператора ввода. Все поля текстовые, а их размер определился автоматически по максимальной длине соответствующей записи, за исключением поля 3 – Формулы смысла, величина которого была установлена как максимальная – 256 символов.

Используя запросную систему СУБД Paradox, были получены несколько производных таблиц, которые после экспорта в формат текстового файла послужили основой для построения частотного словника словоформ, применяемых для представления формулы смысла (поле 3), словника слов-знаков (поле 2), списка способов задания смысла с указанием частоты их использования (поле 1).

Сформированная база данных потенциально позволяет создание сложной модели данных, объединенной с базой данных Ассоциативного тезауруса, на основе которой могут быть построены компаративные исследования форм активного и пассивного способов существования языкового сознания носителя языка-культуры.

Проекции базы данных

Для представления результатов эксперимента были сформированы два словника: частотный словник словоформ, использованных для задания смысла в фигурах знания (когнемах), словник слов-знаков, частотный список способов задания смысла в фигурах знаний и комбинированные сортированные списки когнем.

Частотный словник был получен следующим образом.

На первом шаге в базе данных было выделено поле 3 (формула смысла) в отдельную таблицу, которая затем экспортирована в текстовый файл с разделителями (символ абзаца).

Далее, второй шаг, используя программу Andrew Tools, был получен частотный словник словоформ в виде таблицы СУБД Paradox, которая также была экспортирована в форму текстового файла с разделителями (символ табуляции).

На третьем шаге были осуществлены замены служебных символов (знаки табуляции и абзаца), в результате которых были сформированы столбцы и строки частотного словника.

Словник слов-знаков был получен путем преобразования экспортированного текстового файла таблицы поля 2 базы данных когнем. В результате серии эквивалентных замен и сортировки списка слов-знаков были выделены слова-знаки, являющиеся именами собственными.

Частотный список способов задания смысла в фигурах знаний был получен путем исполнения запроса типа Calc Count к базе данных когнем. Таблица с результатами запроса была экспортирована в текстовый файл с разделителями (символ табуляции) и представлена в виде таблицы MS Word.

Комбинированные списки когнем. Для удобства «ручной» верификации когнем были осуществлены сортировки полей базы данных и сформированы пять комбинированных списков, которые затем, используя стандартную операцию СУБД Paradox, были экспортированы в формат текстового файла с разделителями (символ табуляции) и выведены в печатном виде.

С целью печатного издания фрагмента базы когнем ЛКТ один из комбинированных списков (сортировка: референтная область, функция, знак, способ, формула смысла) был экспортирован в формат текстового файла, снабжен тремя числовыми показателями и распечатан. Числовые показатели характеризуют: мощность множества когнем отнесенных к референтной области (1), мощности составляющих их подмножеств рецептов (2) и ретуши (3).

Верификация экспериментальных данных

Под верификацией экспериментальных данных понимается сложная процедура, состоящая из следующих операций двух типов: тип 1 — формальный поиск и исправление орфографических и синтаксических ошибок в записи компонент когнем, тип 2 — семантический анализ когнем и исправление (изменение) значений их компонент.

Первый тип операций верификации не требует особого комментария, является необходимым и традиционным для представления текстовых данных, введенных клавиатурным способом. Он направлен, прежде всего, на устранение ошибок ввода, возникших из-за неразборчивости рукописного текста некоторых карточек с записями когнем, и «опечаток». Он состоит в автоматизированном орфографическом контроле, реализованном в текстовом процессоре MS Word, и последующем исправлении найденных ошибок правописания. Назовем эту верификацию «формальной».

Представляя второй тип операций верификации нужно сделать следующие замечания:

Во-первых, при проведении эксперимента по исходной концептуальной модели, изменения значений компонент практически отсутство-

вали. Небольшое их количество было связано только с ошибками ввода (например, неразборчивостью скорописного написания слов «ретушь» и «рецепт», *ртш.* и *рцп.* соответственно). Верификационные операции этого типа также можно считать «формальными».

Во-вторых, потребность в изменении значений компонент когнем возникла в основном при проведении эксперимента по расширенной модели. В эксперименте приняли участие анкетеры, которые формировали когнемы по установленным правилам (это правила выполнения самостоятельной работы). Их знаний для правильного выбора «способа задания формулы смысла» оказалось недостаточно, кроме этого, они имели возможность произвольного соотнесения интенсионала когнемы с наименованием референтной области и типом функции. В связи с этими обстоятельствами реализации расширенной модели эксперимента и потребовалась дополнительная, назовем ее «семантическая», верификация.

Основными задачами «семантической» верификации компонент когнемы «способ» и «референтная область» являются поиск и исправление неправильных (некорректных, противоречивых) решений выбора из списка их возможных значений.

Для компоненты когнемы «функция» были установлены следующие особенности. Одна часть анкетеров определяла значение этой компоненты исходя из собственного знания/незнания ответа на вопрос кроссворда, другая — исходя из необходимости для себя знания/незнания или известности им людей, для которых знание ответа на вопрос кроссворда является основным (обязательным).

Изменения значений компонент когнем при «семантической» верификации вносились и в случае обнаружения недобросовестного выполнения задания анкетером (подтасовки результатов, компиляции данных и путаницы терминов «рецепт» и «ретушь»).

Аспект «*Функционирование*» ИТ ЛКТ

Аспект «*Функционирование*» — это моделирование вербального сознания носителя языка культуры в компьютерной среде. Результат моделирования — представление ЛКТ в виде двухкомпонентной интеллектуальной системы, состоящей из базы знаний и интерфейса взаимодействия с пользователем-исследователем. База знаний ЛКТ может состоять из двух компонент: базы данных и лингвистического процессора.

База данных ЛКТ, созданная в результате ввода данных когнитивного эксперимента в рамках исходной концептуальной модели, уже на втором этапе эксперимента, когда были привлечены к сбору данных

несколько десятков анкетеров, существенно возросла. Фактически это происходило не симультанно, а по мере получения данных (сдачи заданий студентами). Иначе, уже при такой технологии проведения эксперимента, при фиксации «порции» новых данных и времени их поступления, оказалось возможным рассматривать базу данных с двух сторон: *синхронного (статического) и диахронного (динамического) когнитивного экспериментов.* В понятие диахронии (динамики) можно вкладывать два смысла: во-первых, связать ее с очередной порцией знания (когнем), это следует рассматривать как псевдовременную характеристику, которую, однако, можно использовать для расчета многих статистических характеристик эксперимента и ЛКТ (ср., например, исследования РАС в работах Г.А.Черкасовой); во-вторых, можно действительно фиксировать время, например год опубликования кроссворда, с целью последующего использования этой информации для анализа изменения содержания ЛКТ.

Лингвистический процессор ЛКТ в отличие от базы данных в настоящее время еще не создан, но в его составе можно представить следующие три сопроцессора: коммуникативный (связной), синтагматический и парадигматический конструкторы. Каждый из сопроцессоров может иметь свою модульную структуру.

Коммуникативный лингвистический процессор может состоять из следующих модулей: взаимодействия с другими лексикографическими базами данных, верификации, синтеза и распознавания когнем, экстенсионального и интенсионального расширения базы данных, когнайзера.

Модуль взаимодействия с другими лексикографическими базами данных обеспечивает работу модулей связного процессора и предназначен для осуществления обмена данными с лексикографическими системами, к числу которых следует отнести электронные словари, терминологические тезаурусы и идеографические словари. Среди словарей (их баз данных), которые могут использоваться в исследовании: а) специально созданные словарные базы — синонимов, антонимов, сокращений; б) существующие, доступ к которым определен их создателями — дефиниций, терминов, ассоциаций, метафор, идиом, пословиц, афоризмов (прецедентных текстов), личных имен и т.д.

Модуль верификации предназначен для решения двух задач: автоматизированной проверки выбора значения способа представления формулы смысла в когнеме и для расширения количества вариантов различных способов связывания формулы смысла со знаком.

Модули синтеза и распознавания когнем реализуют алгоритмы построения и разгадывания кроссвордов. Основное их назначение автома-

тизированная проверка правильности отношений «знак» ←→ «формула смысла», а также генерация множественных отношений между ними.

Модули экстенсионального и интенсионального расширения базы данных эксперимента предназначены для ее неэмпирического расширения. Это можно сделать за счет автоматической генерации экстенсионала на основе процедур синтеза и верификации когнем, и/или включения в интенсионал омонимичных форм знака. На основе такого расширения возможными оказываются также и оценки полноты экспериментальных данных.

Модуль когнайзера в отличие от предыдущих трех модулей работает на основе двух баз данных: ассоциативного и когнитивного экспериментов. Он является одним из основных, а его реализация в составе лингвистического процессора преследует три цели: первая связана с автоматизацией процесса «осознания» как перехода от «формулы смысла» к «знаку» на основе РАС; вторая — с «восхождением» к наименованию референтной области (концепту), т.е. переходом от экстенсионала к интенсионалу, от {«знака» ←→ «способа» ←→ «формулы смысла»} к {«референтной области»←→ «функции»}; третья — с построением интегрированной информационно-программной системы, которую условно можно назвать «ассоциативно-когнитивный эксперимент». Эти цели могут быть достигнуты на основе алгоритмизации активного и пассивного режимов работы когнайзера.

Алгоритмизация активного режима работы когнайзера состоит в формальном описании уже реализованного ассоциативного эксперимента, сведения о котором представлены в Русском ассоциативном словаре-тезаурусе (РАС), как интеллектуального агента, получающего на вход слово-стимул и генерирующего на выходе слово-реакцию. Суть алгоритма генерации реакции состоит в «прямом» и «обратном» просмотре пар {стимул←→реакция} и выборе одной из них. Основной проблемой при этом следует считать задание критерия (правила) его завершения в тех случаях, когда должна быть выбрана одна реакция из некоторого их множества. В работах Г.А.Черкасовой представлены многие основания для разработки такого алгоритма, проведены исследования-измерения РАС, однако алгоритм еще не разработан.

В работах Ю.Н.Караулова приведено «естественно-языковое описание» алгоритма моделирования пассивного режима работы когнайзера на основе РАС. В описании можно выделить процедуры четырех типов: а) переходы от «реакции» к «реакции» при известном «стимуле», т.е. внутри одной словарной статьи прямого РАС; б) переходы от «стимула» к «стимулу» при известной «реакции», т.е. внутри одной словарной ста-

тьи обратного РАС; в) произвольные переходы от «реакции» к «стимулу» и обратно по всему множеству фактических (экспериментально зафиксированных) и виртуальных (теоретически возможных) цепочек {{стимул←→реакция…{стимул←→реакция}}; г) формирование пропозиций. Формальное описание процедур а) и б) не представляет сложности, а для описания процедуры типа в) необходимо сформулировать правила выбора «направления» движения по цепочке и «смены» цепочки. Также как и в случае с алгоритмизацией активного режима работы когнайзера необходим формальный критерий-правило завершения алгоритма. В качестве этих правил-критериев могут быть рассмотрены длины цепочек, а также эквивалентность пропозициональных форм цепочки и «формулы смысла». При этом должен быть найден способ формирования пропозиции для «формулы смысла». Все же основной проблемой алгоритмизации пассивного режима работы когнайзера следует считать нахождение метода априорного задания имени референтной области (концепта), точнее выбора одного из двух типов рассуждений при логическом выводе конечного элемента цепочки — индуктивных или дедуктивных.

Синтагматический лингвистический конструктор. Идея этого лингвистического процессора подробно изложена в [Филиппович 2002б]. Интерпретация идеи применительно к задаче создания лингвокультурного тезауруса следующая. Основными синтагмами базы данных когнем являются символы, слова и предложения; производными — словосочетания и тексты (формальное объединение некоторого количества или всех «формул смысла»). Синтагматический процессор выполняет функции предредактирования текстовых данных эксперимента и создания основных синтагматических конструктивов: словников (полных и частичных, прямых и обратных, частотных) и словоуказателей (индексов, связывающих синтагмы с логическими и физическими структурными единицами собранных данных, таких как когнема, кроссворд, источник, анкетер и др.). В основе информационно-программной реализации синтагматического лингвистического конструктора лежит формальное (теоретико-множественное, автоматное) описание синтагм. База данных ЛКТ, представленная в форме совокупности синтагматических конструктивов, по своей сути является синтагматической моделью ЛКТ, на основе которой можно оценить, например, лексическое богатство языка культуры, особенности подъязыка вопросов языковой игры «кроссворд» и др. Особой функцией синтагматического конструктора является функция комбинирования синтагм и основных синтагматических конструктивов, в результате которой получаются производные

конструктивы (комбинаторные перестановки = сортировки и выборки = проекции).

Парадигматический лингвистический конструктор в понимании [Филиппович 2002б] добавляет в описание синтагматического конструктива некоторую (формальную или неформальную) интерпретацию. Неформальная интерпретация, как правило, представляется в виде научного комментария в естественно-языковой форме. Формальная интерпретация — это, чаще всего, некоторая процедура или алгоритм (программа) преобразования синтагм, например, автоматическая лемматизация — приведение словоформ к основной форме, в общем случае замена одной синтагмы на другую. Формальная интерпретация когнем может быть выполнена с использованием, например, методов компонентного, дистрибутивно-статистического или частотно-семантического анализа [Филиппович 2002б]. В результате между знаками когнем будут установлены отношения, которые можно интерпретировать как семантические поля и ареалы. Основное назначение так полученных парадигматических конструктивов вспомогательное и состоит в использовании при формировании словарных статей тезауруса, а также при автоматической селекции потенциально возможных связей всех синтагм со всеми.

Аспект «*Существование*» ИТ ЛКТ

Третьим аспектом ИТ ЛКТ является аспект «*Существование*» — это форма реализации ИТ ЛКТ как информационно-программного изделия, используемого в исследовательских целях и его характеристики.

Как объект реального мира ЛКТ обладает набором свойств, которые могут быть измерены. Все измерения мы условно разделили на четыре группы: статистические, структурные, процессные и ресурсные

Лингво-статистические измерения (частотный анализ). Эта группа измерений позволяет оценить мощности множеств элементов, составляющих ЛКТ. База знаний лингвокультурного тезауруса русского языка, сформированная в ходе экспериментов и обработки их результатов, составляет ~18470 когнем. Разных «формул смысла» ~18470, на основе ~94380 словоупотреблений (~35400 разных слов). Количество областей референции ~130, количество разных слов-знаков ~13500, количество способов задания «формулы смысла» 47. В числе других измеряемых характеристик следует отметить частотные распределения словоформ в «формулах смысла» и используемых способов их задания.

Лингво-структурные измерения. Эти измерения также «привязаны» к отдельным составляющим когнем. Они пока не проводились, но воз-

можны оценки двух составляющих: «формулы смысла» и «референтной области». Для «формулы смысла» возможным является, например, синтаксический анализ предложений. Для «референтной области» могут быть: выделены типы используемых структур (деревья, сети, цепочки, кольца); проведен анализ мощности деревьев; определена степень связанности когнем (вхождение знаков в различные «референтные области»); определены мощности тезаурусных статей и др.

Лингво-процессные измерения ЛКТ направлены на получение оценок временной и емкостной сложности алгоритмов обработки базы данных. Они носят вспомогательный (служебный) характер и позволяют оценить продолжительность и затраты памяти на обработку экспериментальных данных, а также время ожидания пользователя-исследователя при диалоговом взаимодействии с системой поддержки исследований. Эти измерения касаются практически всех программных компонент ЛКТ, однако особенно важно их знание для процедур, реализуемых синтагматическим и парадигматическим конструкторами.

Лингво-ресурсные измерения нацелены на определение величины материальных затрат на проведение исследований ЛКТ и их трудоемкости.

Формы существования ЛКТ — печатная и электронная.

Печатное (книжное) издание ЛКТ может быть двух типов: а) монография со значительным количеством примеров и комментариев к ним и б) идеографический словарь-тезаурус. Основным недостатком печатного издания словаря-тезауруса является неизбежная необходимость повторения словарного материала при попытке обеспечить тезаурусную многовходовость. Альтернативным вариантом этим двум типам изданий может оказаться вариант одновходового словаря, например, алфавитного расположения когнем, последовательно упорядоченных (отсортированных) по концептосферам, «референтным областям» (концептам), функции, знаку и способу. При этом для удобства пользования возможно применение различных шрифтов и разметки. Другим альтернативным вариантом может оказаться алфавитный список знаков и соответствующих им «формул смысла» с добавленными индексами: индекс знаков рецептов и ретушей, индекс знаков и способов, индекс распределения знаков по концептосферам и концептам. Во всех случаях выбор варианта в основном определяется экономико-технологическими параметрами издания, в итоге сводящихся к количеству печатных листов.

Электронное издание может оказаться более удобным для пользователя. Оно может быть разработано в виде некоторой инструментальной среды, позволяющей не только пользоваться уже полученными резуль-

татами исследований, но и проводить новые. Такая информационно-программная среда может быть локальной и размещена на CD/DVD, или быть оформлена в виде Интернет-версии.

Оптимальным может оказаться комбинированное издание: одновходовый словарь и CD/DVD-версия инструментальной среды создания и использования ЛКТ.

Заключение

В настоящее время отлаженная база данных ЛКТ содержит две части: 1 этап — Картотека 2001-2004 гг. и 2 этап — кроссворды студентов 2004 г.

	1 этап Картотека	*2 этап 12.2004*	*3 этап 12.2005 прогноз*
введено когнем (карточек)	7955	17122	~5000
получилось уникальных	6835	13570	~3800
разных знаков	5352	9242	~2600
новых знаков	5352	8125	~1100
всего разных знаков в базе	5352	13477	~14600

Грубый подсчет показывает, что на первом этапе каждая тысяча карточек давала 670 новых знаков, на втором — 470, на третьем можно предположить — 220. Явно просматривается гиперболическая зависимость между переменными, такая же, как и в ассоциативном эксперименте (в исследованиях Г.А.Черкасовой), и в частотном с текстами (множество данных различных авторов, начиная с Ципфа).

В связи с этим возникает гипотеза-предположение:

Возможно, гиперболический закон проявляет себя в любых вербальных (языковых) экспериментах вообще, или даже семиотических? Может это «закон когниции» для субстрата?

Разворачивая данное предположение в конкретные когнитивные эксперименты можно говорить о сравнительно легко реализуемых технологиях, по типу описанной в статье — а) когнитивный эксперимент с иноязычной (не русской) лексикой; б) полиязычный когнитивный эксперимент; в) когнитивный эксперимент с предметной лексикой (особая технология извлечения предметных знаний, тестирование или контроль профессиональных компонент знаний) и др.

Более неопределенными выглядят технологии когнитивных экспериментов с «семиотическими объектами», например с элементами гра-

фических изображений, будь то рукописные или печатные графемы, или даже виземы жестомимических и артикуляционных образов.

С осторожностью следует отнестись к возможному математическому самообману — подумать о тривиальном статистическом результате обработки большого объема данных как о некой физической константе языкового сознания. Поэтому важным является применение в качестве механизма интерпретации результатов экспериментов формальных моделей различного типа: логико-лингвистических, алгебраических, нейро-сетевых, генетических и др.

Не последнюю роль в когнитивных экспериментах может играть и эффективные информационно-программные средства как компоненты автоматизированных систем научных исследований. Удачные пользовательские интерфейсы для исследователей языкового сознания, предназначенные для управления экспериментами и визуализации их результатов, способствуют нахождению рациональных объяснений его функционирования.

Литература

Караулов 1981 *Караулов Ю.Н.* Лингвистическое конструирование и тезаурус литературного языка. М., Наука, 1981

Караулов 2003а *Караулов Ю.Н.* Единицы языка и единицы знания. // Теория и практика лингвистического анализа текстов СМИ в судебных экспертизах и информационных спорах: Материалы научно-практического семинара. Ч. 2. М., 2003 а.

Караулов 2003б *Караулов Ю.Н.* Языковое сознание: пассивный и активный режим работы // Межкультурная коммуникация и перевод: Материалы межвузовской научной конференции 31 января 2003 г. М., 2003 б.

Караулов 2004а *Караулов Ю.Н.* Концептография языковой картины мира. Статья 1. Первый этап ”восхождения” к образу мира: от элементарных фигур знания к предметно-референтным областям культуры // Проблемы прикладной лингвистики. Вып. 2. М., 2004

Караулов 2004б *Караулов Ю.Н.* Концептография языковой картины мира. Статья 2. Референтные области,

концепты и концептосферы. (Второй этап “восхождения” – от областей к концептам). // Языковое сознание: теоретические и прикладные аспекты. Москва – Барнаул, 2004

Караулов 2004в *Караулов Ю.Н.* Основы лингвокультурного тезауруса русского языка. // Русское слово в русском мире. Калуга, Изд. “Эйдос”, 2004

РАС 2002 Русский ассоциативный словарь. В 2 т. Т. 1. От стимула к реакции: Ок. 7000 стимулов / Ю.Н.Караулов, Г.А.Черкасова, Н.В.Уфимцева, Ю.А.Сорокин, Е.Ф.Тарасов. – М.: «Астрель»: ООО «АСТ», 2002. – 784 с.

Филиппович 2001б *Филиппович Ю.Н.* Семиотическая концепция интеграции информационных технологий // Scripta linguisticae applicatae. Проблемы прикладной лингвистики –2001. Сборник статей / Отв. ред. А.И.Новиков. М.: «Азбуковник». 2001.

Филиппович 2002а *Филиппович Ю.Н., Черкасова Г.А., Дельфт Д.* Ассоциации информационных технологий: эксперимент на русском и французском языках. / Серия «Компьютерная лингвистика». Вступ. статья Н.В.Уфимцевой. М.: МГУП, 2002. 304 с.

Филиппович 2002б *Филиппович Ю.Н., Прохоров А.В.* Семантика информационных технологий: опыты словарно-тезаурусного описания. / Серия «Компьютерная лингвистика». Вступ. Статья А.И.Новикова. М.: МГУП, 2002. 368 с.

Статьи аспирантов и студентов

А.А.Александрова

ЖЕСТОМИМИЧЕСКИЕ АВТОМАТИЗИРОВАННЫЕ СИСТЕМЫ

Много лет в сурдопедагогике решается вопрос о необходимости использования языка жестов при обучении глухих. Традиционной является система «чистого устного (орального) метода обучения», когда основное внимание сурдопедагогов направлено на произносительную сторону речи глухих. Однако в 1980 году в странах Западной Европы и США появляется новое направление в обучении глухих – билингвистический подход. Этот подход в обучении глухих предусматривает использование национального языка (в письменной, устной и дактильной форме) и национального жестового языка, выступающих в качестве равноправных «партнеров» в процессе коммуникации между глухими и слышащими преподавателями, учащимися, родителями. Появлению и развитию нового направления сурдопедагогики способствовали следующие факторы:

1. Фундаментальные исследования национальных жестовых языков, проведенные У. Стоку в США, Г.Л. Зайцевой в России, М. Бреннон в Великобритании, показали, что жестовые языки – это сложные лингвистические системы со своей грамматикой, морфологией, лексикой.

2. Недовольство результатами устного метода обучения и метода тотальной коммуникации. Первый метод скорее приемлем для лиц с ощутимыми остатками слуха. Второй метод, по сути, сводился к сопровождению речи преподавателя жестами.

3. Изменились взгляды общества на проблему инвалидности и нетрудоспособности: люди с физическими, умственными или сенсорными дефектами, психическими заболеваниями должны иметь те же права и обязанности, что и другие члены общества [3, 6].

Многие проблемы методического и организационного характера, связанные с использованием билингвистического подхода в обучении глухих, предстоит еще решить, однако накопленный мировой опыт в этой области позволяет судить о перспективности данного направления в сурдопедагоги-

ке. Использование жестового языка позволяет устранить коммуникационные барьеры, создает доверительные отношения между преподавателями и учащимися (далеко не сразу глухие дети и подростки могут передавать свои мысли словами [4]), обеспечивает эмоционально окрашенное обучение. Это дает возможность ускорить передачу учебной информации, увеличить ее объемы и, следовательно, расширить кругозор глухого учащегося.

Принимая во внимание то, что жестовый язык – это сложная, высокоразвитая лингвистическая система, все, кто заинтересован в знании жестового языка (глухие, родители глухих детей, преподаватели спецшкол, переводчики и т.д.), нуждаются в существующем разнообразии средств поддержки словесного языка, в том числе и в автоматизированных системах, оснащенных соответствующим программно-лингвистическим обеспечением (так называемые, жестомимические и артикуляционные автоматизированные системы или просто жестомимические системы). По функциональному назначению можно выделить следующие группы таких систем: мультимедийные словари, системы машинного перевода, мультимедийные обучающие системы, конструкторы жестов, программы автоматического распознавания жестов, а также комплексные системы, в которых объединены функции двух или более перечисленных выше видов систем.

Ниже представлен ряд автоматизированных систем (как отечественных, так и зарубежных), которые были отобраны для демонстрации существующих возможностей, предоставляемых жестомимическими системами, в целом.

Online-словарь Британского жестового языка (BSL), размещенный на официальном сайте Центра исследования проблем жестового языка и глухих Бристольского университета [7]. На настоящий момент словарь содержит 4807 жестов. Каждая словарная статья состоит из заголовочного слова (словесное описание жеста), видеоклипа и примеров употребления данного жеста. На сайте предлагаются следующие виды поиска жестового обозначения слова: индексированный поиск, тематический поиск, поиск по заданному слову. Для просмотра видеороликов пользователю предлагается установить соответствующие проигрыватели. Кроме этого предоставляется возможность задать качество просматриваемого видеоклипа.

Мультимедийный словарь Американского жестового языка (ASL), разработанный компанией Vcom3D [9]. Данный словарь входит в серию программных продуктов Sign Smith® и использует технологию ASL Animations. Для демонстрации жестов используются трехмерные динамические изображения, позволяющие пользователю просматривать жест под нужным углом на требуемом расстоянии. Скорость демонстрации по желанию пользователя может быть изменена. Кроме этого, пользователю предоставляет-

ся возможность выбрать облик анимированного персонажа (в настоящий момент разрабатывается библиотека виртуальных образов). Словарь содержит 500 жестов. Словарная статья состоит из заголовочного слова, анимации, указания части речи и примера употребления данного жеста (данный пример можно просмотреть как на ASL, так и на калькирующем английском языке). В данном словаре поддерживаются следующие виды поиска жестового обозначения слова: индексированный поиск, тематический поиск (всего 20 категорий) и поиск по заданному слову.

«Программно-аппаратный комплекс жестовой речи», разработанный в Институте социальной реабилитации Новосибирского государственного технического университета. Данный комплекс предназначен для компьютерного сурдоперевода с русского языка на видеофильм калькирующей жестовой речи. Ввод русского текста осуществляется либо с клавиатуры, либо из файла, либо с голоса. Стратегия перевода русского текста на дактильную азбуку и язык жестов направлена на наиболее близкий по смыслу перевод и использует аппарат теории нечетких множеств. Если анализируемое слово или словосочетание содержится в базе данных видеожестов, то оно переводится без изменений. В случае отсутствия – система пытается выбрать наиболее близкое по смыслу слово или словосочетание в словарной базе, просматривая синонимы и антонимы и исходя из максимума функции принадлежности. Дактильный перевод применяется в последнюю очередь, когда не остается других возможностей, причем, возможен как побуквенный перевод, так и перевод по слогам. Имеется возможность привязки к определенному слову (идентификатору) какого-либо текста с последующим переводом данного текста по заданному идентификатору. Для обеспечения плавности результирующего видеофильма используются промежуточные жесты. Для повышения эффективности работы программного комплекса был введен многопотоковый режим: в один поток включены операции показа видеоролика, во второй – операции поиска, анализа, обработки и т.д. [5].

Программный продукт Sign Smith Studio, разработанный компанией Vcom3D [9]. Данный продукт использует технологию ASL Animations. Позволяет автоматически преобразовывать текстовую информацию на английском языке в калькирующую жестовую речь (Signed English), сопровождающуюся устной речью. Тем пользователям, которые знакомы с грамматикой ASL, предоставляется возможность автоматизированного перевода исходного текста на ASL, при этом пользователь может задать выражение лица, направление взгляда, положение головы, сопровождение жестикуляции беззвучным шевелением губ демонстратора. Облик анимированного персонажа также может быть выбран пользователем. Словарная база сис-

темы содержит 2500 жестов. Имеется возможность автоматического перевода более чем 10000 английских слов. Sign Smith Studio позволяет сохранять созданные анимационные ролики во внутреннем формате программы, в виде файлов html и в видео-формате.

Можно отметить еще одну систему машинного перевода текстовой информации в калькирующую жестовую речь. Это *программа Mimehand II* фирмы Hitachi, Ltd. [8].

Среди работ по автоматическому распознаванию жестов, можно выделить *проект Artemis* во Франции [10]. В данном проекте ставится задача распознавания дактильных знаков французского языка, но не жестов, передающих понятия (одной из причин этого является отсутствие полного лингвистического описания Французского жестового языка).

Наиболее интересной разработкой в области конструирования жестов является *редактор жестов Sign Builder* компании Vcom3D [9]. При помощи Sign Builder пользователь может создавать специализированные жесты, означивающие технические, медицинские и пр. термины, конструировать жесты не только для ASL, но и для других национальных жестовых языков (British Sign Language и т.д.). Данный редактор использует технологию Inverse Kinematics: при создании пользователем жеста (выборе формы кисти и положения руки виртуального персонажа) программа автоматически подбирает естественное положение кистевого, локтевого и плечевого суставов. В настоящее время данная система находится в стадии разработки.

Наконец, к комплексным жестомимическим системам можно отнести отечественный *проект «Фонд жестовой речи»* (в настоящее время данный проект находится в стадии разработки) [2]. Данная система будет представлять собой online-переводчик текстовой информации в жестовую разговорную и калькирующую речь, для работы со словарем жестов пользователю будет предоставлен online-редактор жестов, банк данных жестового языка будет включать в себя базы данных различных национальных жестовых языков.

Анализ приведенных выше систем позволил сформировать следующий перечень наиболее интересных возможностей, предоставляемых современными жестомимическими системами.

− Использование в качестве способа представления жестовой информации трехмерных динамических изображений. Такой способ представления жестов особенно полезен в тех случаях, когда пользователь плохо или совсем не знаком с жестовым языком.

− Развитие жестомимических систем, таких как online-словари и системы online-обучения. Такие системы позволяют получать удаленным

пользователям необходимую информацию о жесте того или иного национального жестового языка и организовывать дистанционное обучение.

− Автоматическое преобразование системами машинного перевода текстовой и аудио информации в калькирующую жестовую речь с последующей возможностью внедрения полученного ролика в html-файл.

− Автоматизированное преобразование системами машинного перевода текстовой информации в разговорную жестовую речь с последующей возможностью внедрения полученного ролика в html-файл.

− Возможность создания жестов различных национальных жестовых языков при помощи программ-редакторов жестов.

Наряду с несомненной пользой, которую принесло развитие и использование жестомимических систем, многие из них обладают рядом существенных недостатков, которые в дальнейшем еще предстоит исправить.

1. Большинство жестомимических программ для представления жестов используют двухмерные динамические изображения (видеофрагменты и анимацию), которые в отличие от трехмерных динамических изображений не позволяют просматривать жест с нужной стороны. Ввиду специфики мышления глухих, такие программы как, например, словари, использующие в качестве способа представления жеста видеофрагменты, имеют еще и тот недостаток, что могут пополняться только с помощью демонстраторов жестов, которые работали над первой версией программы. Таким образом, наиболее эффективным способом представления жестовой информации является использование трехмерных анимированных персонажей. Данные персонажи должны иметь человеческий облик, т.к. в ряде случаев у глухих вызывает отторжение наделение каких-либо персонажей (например, животных) способностями, несвойственными им в реальной жизни.

2. Отсутствие в обучающих системах средств редактирования жестов. Таким образом, пользователь не может создать собственное жестовое высказывание и сравнить его с эталонным вариантом.

3. Современные программы машинного перевода не предусматривают функции автоматического перевода текстовой или аудио информации на жестовый разговорный язык, что в первую очередь связано с отсутствием полного лингвистического описания национальных жестовых языков, а также с проблемой передачи смысла в словесном и жестовом языках. В настоящее время ведутся работы по сопоставлению процессов передачи смысла в жестовом и словесном языке [2]. Результаты этих работ, возможно, могут быть полезны при разработке программ машинного перевода с одного словесного языка на другой, а также при создании Semantic Web.

4. Отсутствие в современных жестомимических системах такого вида поиска, как поиск жеста.

5. Современные словари жестового языка построены по принципу слово-жест, что не позволяет включать в них образцы безэквивалентной лексики и фразеологии жестового языка. В настоящий момент разрабатываются подходы к составлению словарей нового типа, исходящие из значения жеста.

6. Отсутствие online-редакторов жестов, что не позволяет удаленным пользователям пополнять существующие словарные базы национальных жестовых языков.

Каждый из этих недостатков представляет собой серьезную научную и практическую задачу. Решение данных задач откроет новые перспективы в области развития жестомимических автоматизированных систем, поисковых систем, систем машинного перевода, а также систем распознавания речи и образов.

Литература

1. *Базоев В.З.* Человек из мира тишины / В.З. Базоев, В.А. Паленный. – М.: ИКЦ «Академкнига», 2002. – 815 с.: ил.
2. *Воскресенский А.Л.* Фонд жестового языка [Электронный ресурс] // Официальный сайт компании Deria Grapfics. – Режим доступа: http://www.deria.ru/our_sci_invent_GLF.php.
3. Декларация о правах инвалидов [Электронный ресурс] // Резолюция Генеральной Ассамблеи ООН A/RES/3447 (XXX). – Режим доступа: http://www.un.org/russian/documen/declarat/disabled.htm.
4. *Зайцева Г.Л.* Жестовая речь. Дактилология : учеб. для студентов вузов / Г.Л. Зайцева. – М.: «Владос», 2004. – 191 с.: ил.
5. Официальный сайт Института социальной реабилитации НГТУ [Электронный ресурс]. – Режим доступа: http://www.nstu.ru/isr.
6. Стандартные правила обеспечения равных возможностей инвалидов [Электронный ресурс] // Резолюция Генеральной Ассамблеи ООН A/RES /48/96. – Режим доступа: http://www.un.org/documents/ga/res/48/a48r096.htm.
7. Centre for Deaf Studies of Bristol University [Электронный ресурс]. – Режим доступа: http://www.bris.ac.uk/deaf/.
8. Official website of Hitachi Europe, Ltd. [Электронный ресурс]. – Режим доступа: http://www.hitachi-eu.com.
9. Official website of Vcom3D, Inc. [Электронный ресурс]. – Режим доступа: http://vcom3d.com.
10. Project Artemis [Электронный ресурс]. – Режим доступа: http://www-artemis.int-evry.fr.

Е.М.Буркова

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ОБРАБОТКИ СМЕТ И ЛИМИТОВ: КОМПОНЕНТА РАЗНЕСЕНИЯ ДОГОВОРОВ НА СТАТЬИ РАСХОДОВ[2]

Введение

Проблема автоматизации работы высших учебных заведений и их отделов существует уже достаточно давно. Современный объем информации не позволяет вести все, и даже часть операций только в бумажном формате. Однако, с появлением персональных компьютеров, которые стали доступны не только крупным промышленным учреждениям, но и домашним пользователям, появились и специальные программы, предназначенные для упрощения таких операций как бухгалтерский учет, документооборот и других.

Большинство предприятий уходит от бумажного документооборота, переходя на автоматизированные системы управления предприятием. В подобной ситуации резко возросла потребность в системах поиска и анализа данных и повышения скорости первичной обработки данных.

Целью разработки, описанию которой посвящена статья, является повышение эффективности работы управления планирования и финансирования МГТУ им. Н.Э.Баумана. Ускорение составления смет, работа с лимитами, уменьшение количества ошибок, принятие более качественных управленческих решений, рациональная организация работы управления планирования и финансирования.

Анализ видов планирования

Виды планирования различают по срокам и по виду деятельности.

Планирование по срокам

В зависимости от сроков планирование разделяется на оперативное (до года), краткосрочное (один год), среднесрочное (три-пять лет), долгосрочное (свыше пяти лет).

---

[2] Научный руководитель работы к.т.н., доцент Г.И.Ревунков. Статья передана для публикации в декабре 2004 г.

Планирование осуществляется с соблюдением конкретных принципов, т.е. правил формирования, обоснования и организации разработки плановых документов. Основными из них являются: научность, социальная направленность, повышение эффективности производства, пропорциональность и сбалансированность, приоритетность, согласование кратко-, средне- и долгосрочных целей.

Принцип научности реализуется в том, что плановые документы должны разрабатываться с использованием законов общественного развития, анализа тенденций и прогнозов социально-экономического функционирования общества, НТП.

Принцип социальной направленности планирования подразумевает первоочередной учет интересов человека и общества, направленность планов на удовлетворение их потребностей.

Принцип повышения эффективности производства требует, чтобы результаты выполнения планов достигались максимальной экономией труда и ресурсов на единицу продукции. Важны при планировании такие показатели эффективности, как рост производительности труда, снижение материало- и энергоемкости, повышение фондоотдачи, рентабельности и т.д.

Принцип пропорциональности и сбалансированности заключается в том, что рост производства должен сопровождаться улучшением структуры хозяйства, уменьшением доли отсталых производств, расширением применения высоких технологий и др. При этом должны увеличиваться объемы продукции, имеющей максимальный спрос на мировом рынке, главным образом, товаров с большей долей добавленной стоимости, и сокращаться сырьевые поставки другим странам.

Принцип согласования краткосрочных и долгосрочных целей по существу является продолжением принципа пропорциональности, но здесь главное внимание уделено учету целей при планировании на временных отрезках. Задаваемые результаты планов при этом должны согласовываться со стратегическими целями предприятия.

Планирование по виду деятельности

Стратегическое планирование - это деятельность верхнего уровня. Она отражается в идеях, концепциях, задачах, подходах. Стратегическое планирование формулирует широкие идеи и цели, развивает стратегии (определяет пути и средства достижения целей). Благодаря стратегическому планированию определяется перспектива развития учреждения, разрабатывается концептуальная основа для принятия кардинальных решений в части будущих рынков, продукции, структуры, прибыльно-

сти и профиля риска банка. Но стратегический план не отличается подробностью. В нем даются ответы на вопросы: "кто?", "что?" и "как?"

Роль стратегического планирования неоднозначна в разной среде. Если условия мало меняются, то роль его незначительна. Достаточно внести небольшие коррективы, и прежний успех в деятельности обеспечен.

Тактическое планирование ориентируется на конкретные мероприятия, представляет собой второй уровень планирования. Оно реализуется в форме конкретного плана действий (мероприятий) для достижения конкретной цели и является поддержкой стратегического плана.

На третьем уровне планирования осуществляется *финансовое планирование* и разработка бюджетов, где определяются финансовые показатели для конкретизации целей, стратегий и заданий плана. Эти показатели служат надежным средством контроля деятельности предприятия в предстоящем году.

Стратегии реализуются в оперативных планах. Оперативный план — документ, цель которого - обеспечить общее понимание задач учреждения, стратегии и тактики для выполнения этих задач, а также определить объемы, качество и структуру ресурсов, выделяемых для этого.

Если стратегии разрабатываются на относительно долговременную перспективу, то оперативные планы составляются обычно на предстоящий год. [1][3]

Оперативный план устанавливает границы количественных и качественных задач, конкретизируя их для ВУЗа в целом и для каждого его подразделения. Этот документ определяет взаимосвязи и взаимозависимости в рамках ВУЗа и создает представление о его будущем. Оперативный план дает характеристику деятельности ВУЗа при распределении бюджетных средств, предусматривает меры по расходу средств и выявлению остатков средств, устанавливает рамки для разработки системы оплаты труда и стимулирования труда.

Для полноценной и эффективной работы и развития университета необходимы все виды планирования.

В управлении планирования и финансирования, в частности используется долгосрочное, краткосрочное и оперативное планирование, что обеспечивает его наибольшую эффективность работы в составе структурных подразделений.

---

[3] Данная статья посвящена описанию видов и способов экономического планирования. В статье рассматривается планирование на примере банка, но также содержится общая информация о планировании, которая и была использована в данной работе.

В системе автоматизации составления и обработки плановых смет по бюджету и внебюджету в управлении планирования и финансирования МГТУ им. Н.Э.Баумана используется по временному критерию – краткосрочный и оперативный анализ, а по виду деятельности – финансовое планирование.

Анализ видов смет

Смета — почти бизнес-план, но со своими особенностями. Существует несколько видов смет. Рассмотрим подробнее их основные виды.

Строительные сметы

Строительная смета – представляет собой оценку стоимости строительно-ремонтных работ. Строительная смета поможет оценить:

- сколько материалов будет потрачено на строительство (ремонт);
- каков объем работ;
- сколько времени потребуется на проведение работ;
- сколько это должно стоить.

При составлении строительной сметы учитывают сметные нормы существующей сметно-нормативной базы. Существует порядок составления сметных норм и единичных расценок:

- государственные элементные сметные нормы — ГЭСН;
- федеральные единичные расценки для базового района — ФЕР;
- нормы накладных расходов;
- нормы сметной прибыли (плановые накопления);
- нормы затрат на временные здания и сооружения;
- нормы затрат на удорожание работ в зимнее время;
- методические указания по включению прочих затрат в главу 9 сводных сметных расчетов и смет. Остальное делается на местах под методическим руководством Госстроя России.

Строительные сметы могут быть составлены базисно-индексационным методом и ресурсным методом — это калькулирование сметной стоимости каждого строительного процесса по статьям затрат (заработная плата рабочих, эксплуатация машин, материалы, накладные расходы). [2][4]

Сметы бюджетных учреждений

Бюджетное учреждение может вести учет самостоятельно или обслуживаться централизованной бухгалтерией. Но и в первом, и во втором случаях для обеспечения своей деятельности учреждение составля-

[4] В статье дается определение строительной сметы, ее назначение, виды строительных смет по методам их составления.

ет индивидуальные сметы и планы ассигнований по каждой бюджетной программе (функции), которую она будет выполнять, например, для предоставления медпомощи населению, для контроля за поступлением налогов и т.п.

Чтобы бюджетное учреждение выполняло свои функции, определенные его Положением или Уставом, ему в первую очередь необходим план, в котором должны быть определены доходы, которые оно может получить на протяжении бюджетного периода и соответствующие расходы, которые будут осуществлены на содержание учреждения, для осуществления им своих полномочий. Так, городскому совету необходимо выполнять свои полномочия относительно осуществления соответствующего контроля за поступлением налогов и местных сборов (содержание автостоянок, заработная плата контролеров, которые собирают сбор за парковку, контроль за полнотой поступления выручки и т.п.). Управление культуры планирует, какие культурно-массовые мероприятия будут проведены на протяжении года, сколько на это пойдет средств, и какая выручка от них поступит в бюджет за соответствующий период.

Смета бюджетных учреждений — основной плановый документ, который предоставляет полномочия бюджетному учреждению относительно получения доходов и осуществления расходов, определяет объем и направление средств для выполнения бюджетным учреждением своих функций и достижения целей, определенных на год в соответствии с бюджетными назначениями.

Бухгалтер бюджетного учреждения имеет непосредственное отношение к смете, как к ее планированию (предоставляя отчетные данные за предшествующий период финансовому подразделению), так и к контролю за ее выполнением, ведь бухгалтер отвечает за все финансовые операции, которые будут осуществлены бюджетным учреждением на протяжении бюджетного периода.

Каждое учреждение, выполняя свои функции, должно иметь на это соответствующее разрешение, закрепленное законом, и право получать и распоряжаться средствами в определенных объемах, то есть иметь определенные полномочия. В свою очередь бухгалтеру необходимо знать:

- какие средства должны поступить в бюджет;
- сколько можно использовать на содержание учреждения;
- как их правильно использовать.

Перед началом бюджетного периода всеми учреждениями на следующий бюджетный год, на который планируются расходы, составляются сметы. Составление смет необходимо т.к.:

Во-первых, в соответствии с государственным классификатором управленческой документации баланс выполнения сметы расходов и отчет о его выполнении являются финансовой документацией, которую подает бухгалтер.

Во-вторых, согласно квалификационной характеристике главного бухгалтера именно он принимает участие в подготовке и представлении других видов периодической отчетности, которые предусматривают подпись главного бухгалтера. Сметы и планы ассигнований подписываются руководителем учреждения и руководителем его финансового подразделения или главным (старшим) бухгалтером. В том случае, если в учреждении не предусмотрена по штату должность начальника финансового подразделения (например, сельских, поселковых, городских советах, территориальных центрах собеза и т.д.), составление и утверждение сметы ложится на плечи главного бухгалтера. Рассмотрим структуру сметы более подробно. Любая смета состоит из двух частей: I часть "ДОХОДЫ"; II часть "РАСХОДЫ". Доходы и расходы состоят из общего и специального фондов.

К общему фонду сметы бюджетного учреждения можно отнести средства, которые поступают от уплаты налогов, сборов, предоставления услуг и от вышестоящего учреждения для содержания учреждения и расходуются согласно определенным видам *основной* деятельности.

К специальному фонду сметы относятся средства, которые поступают от уплаты налогов, сборов, предоставления услуг, которые предусмотрены законом о государственном бюджете, и расходуются за счет этих поступлений на *конкретную цель*, например, для создания государственных запасов и резервов.

Поэтому для создания специального фонда в составе местного бюджета должно быть выдано решение соответствующего совета о создании такого фонда, в котором перечисляются конкретные источники поступлений и расходы, которые будут осуществляться за счет средств этого фонда. Данное решение принимается с соблюдением закона о Госбюджете России.

Чтобы сгруппировать доходы, расходы и финансирование бюджета по экономическим признакам, функциональной деятельности, организационным управлениям, соответственно действующему законодательству и международным стандартам, Минфином РФ разработана *бюджетная классификация.*

Бюджетная классификация имеет такие составные части:

1) классификация доходов бюджета;

2) классификация расходов (в том числе кредитования за вычетом погашения) бюджета;

3) классификация финансирования бюджета;

4) классификация долга.

Теперь рассмотрим более подробно, из чего состоят доходы.

Доходы - это средства, полученные учреждением от:

1. Предоставления платных услуг;

2. Выполнения работ;

3. Реализации продукции;

4. Продажи необоротных активов (кроме зданий и сооружений);

5. Суммы излишков материалов, выявленных во время проведения инвентаризации;

6. Списания дебиторской задолженности;

7. Средств на выполнение отдельных поручений;

8. Другие собственные поступления: средства, которые фактически поступили от вышестоящего учреждения;

9. Средств, которые поступили непосредственно на имя учреждения (например, благотворительная помощь), субвенции, полученные из бюджета другого уровня.

Кроме доходов органы власти и органы местного самоуправления также привлекают средства путем взятия долговых обязательств. Доходы и долги бюджета считаются *поступлениями в бюджет.*

Расходы - это затраты, которые осуществляет учреждение для своего содержания и выполнения своих функций. В т. ч. расходы по неуплаченным счетам кредиторов, по начисленной и не выплаченной заработной плате и стипендиям и т.п., одним словом - кредиторская задолженность. То есть под расходами понимают государственные платежи, которые не подлежат возвращению. *Расходы делятся на текущие и капитальные.* Они бывают отплатными, т.е осуществленными в обмен на товар или услугу, или неоплатными (односторонними), например, такими, как заработная плата, питание, медикаменты и т.п. Под капитальными расходами понимают платежи с целью приобретения предметов, оснащения и т.п. с сроком службы больше года, запасов товаров, земли, нематериальных активов или неоплатные платежи, которые передаются получателям с целью приобретения ими активов, или компенсации потерь, связанных с разрушением или повреждением основных фондов. К текущим расходам относятся платежи, которые осуществляются на про-

тяжении года на неотложные потребности учреждения (выплата заработной платы, расходы на командировку, отопление и т.п.). [3][5]

Разрабатываемое ПО посвящено работе со сметами бюджетных организации, поэтому они рассмотрены так подробно. Так же в ПО идет работа с лимитами.

Лимиты – это фактически те же сметы, но они поступают в Управление планирования и финансирования из Федерального агентства по образованию. После определения лимитов Управление начинает работу с ними. Происходит их распределение по подразделениям и на нужды ВУЗа.

Обзор и сравнение аналогов разрабатываемой системы

В настоящее время существует достаточно большое количество разработок, посвященных автоматизации бюджетных организаций и подразделений ВУЗов.

В управлении планирования и финансирования в самом начале перевода деятельности в сторону автоматизации учет велся в Microsoft Excel. Далее была предпринята попытка перевести все на Microsoft Access, однако она не увенчалась успехом. Причин тому несколько:

- автоматизация представляла сбой отдельные модули не связанные между собой.
- не было возможности проследить исполнение договора по бухгалтерии.
- сотрудники подразделений приносили сметы только на бумажном носителе и их приходилось перенабирать.

Затем в качестве аналога была рассмотрена программа 1С, которая тоже не увенчалась успехом. Причин тому тоже несколько:

- отсутствие готовой конфигурации, подходящей для работы управления планирования и финансирования
- достаточно высокая стоимость разработки конфигурации «с нуля» (в то время)
- специфический язык программирования в конфигураторе

[5] В данной работе говорится об учете в бюджетных учреждениях. Рассматриваются особенности составления смет в бюджетных учреждения. Подробно рассматривается структура бюджетной сметы, из чего она состоит. Что входит в статьи дохода и расхода.

- постоянно возникающие неудобства в работе, требующие перестройки конфигурации (причина тому не система 1С, а специфика работы УПФ МГТУ)

Также в качестве аналогов были рассмотрены решения по автоматизации ВУЗа, представленные на выставке SOFTOOL 2003 г. в секции «ИТ в образовании»:

Для системы "Университет" в качестве основы выбрана программная платформа R/3 компании SAP AG. Предлагает данное решение ООО "РЕДЛАБ ЛТД".

Система состоит из следующих модулей:

- Управление учебным процессом.
- Управление финансами и материальными ресурсами.
- Планирование и управление бюджетом.
- Управление персоналом и оргменеджмент.
- Учет труда и зарплата.
- Управление документами.
- Управление исследованиями и грантами.
- Управление документами.

Указанные модули настраиваются для каждого вуза в полном комплекте или в соответствии с потребностями вуза. Но данная система имеет ряд существенных недостатков: Во-первых, высокая стоимость. Во-вторых, сложность внедрения, более высокие технические требования, сложность в обучении. В-третьих, дорогое сопровождение и высокая стоимость сотрудников, поддерживающих решения на данной платформе.

Сведем аналоги в итоговую таблицу.

	Вес	**Excel (до 2000г)**	**ИПО Смета (Access)**	**Конфигурация 1С**	**ИС Университет (SAP R/3)**	**СиЛ+(SQL Server)**
Стоимость (затраты на разработку)	0,184	5	5	2	1	4
Временные затраты на разработку (внедрение)	0,092	5	4	3	2	4
Совместимость с существующим ПО	0,066	2	3	3	4	5
Скорость внесения изменений	0,053	4	4	3	3	4

Производительность работы за системой	0,105	2	3	4	5	5
Удобство работы за системой	0,066	3	4	3	4	5
Наличие поискового механизма (автоматическое предложение статей расходов после введения названия договора)	0,105	1	2	2	4	5
Требования к аппаратному обеспечению	0,112	5	5	3	4	4
Стоимость поддержки	0,098	1	3	4	5	3
Надежность хранения данных	0,119	2	3	4	5	5
ОЦЕНКИ		3,113	3,698	3,033	3,533	4,363

Оценки даны по пятибалльной шкале, причем «5» — наилучшая оценка, «1» — наихудшая. Кроме того, каждая оценка имеет свой вес от 0 до 1. Из таблицы видно, что наилучшим решением является система «СиЛ+» при условии, что она будет удовлетворять всем критериям качества. Кроме того, как уже отмечалось выше, при принятии решения о среде разработки немаловажную роль играет опыт разработчика и существующее программное обеспечение в организации и у партнеров. В этом отношении предпочтение следует отдать СУБД Access и SQL Server.

Планируемая реализация компоненты поиска релевантных статей

Как правило, сметы по внебюджету состоят из отдельных договоров, которые составляет экономист. Договор по смете это фактически та же смета (см. п. 2.2).

Также как и смета, договор имеет доходную часть и расходную. Существует большое число документов, в которых говорится, какие расходы можно отнести к соответствующим статьям.[4]

При внесении договора в базу данных, можно ускорить этот процесс путем автоматического предложения статей расхода после внесения пользователем названия договора. Это значительно сократит время обработки, т. к. экономисту не нужно будет просматривать все документы и правила отнесения расходов на статьи. Для этого в автоматизирован-

ной системе работы со сметами и лимитами предусмотрен поисковый механизм, который будет рассмотрен ниже.

*Замечание***:** данная функция является подсказкой, окончательный же выбор делает пользователь.

Для реализации этой функции необходимо:

1. Разработать обобщенный алгоритм поиска релевантных статей по названию договора.
2. Определить структуру БЗ и БД для ее хранения.
3. Разработать алгоритм разбора запроса.
4. Разработать интерфейс для заполнения БЗ.

Анализ и выбор алгоритма поиска релевантной информации

Сначала дадим определение релевантности. Найденный по запросу документ может иметь отношение к запросу, т. е. содержать нужную (искомую) информацию, а может и не иметь к запросу никакого отношения. В первом случае документ называется релевантным (по-английски relevant – «относящийся к делу»), во втором – нерелевантным, или шумовым. [5]

Булев поиск. Опирается на использование инвертированного индекса ключевых слов, т. е. таблицы, в которой для каждого ключевого слова перечисляются все документы, где оно встречается. Главным достоинством этого алгоритма является возможность связывания слов запроса логическими операциями, например, он позволяет осуществить поиск по запросу «кофе или чай» и получить в результате объединение множеств документов, содержащих слова «кофе» и «чай». К недостаткам этого алгоритма следует отнести невозможность определения релевантности запросу полученной выборки документов и, как следствие, невозможность ее сортировки.

Булев поиск:

« » (пробел) или «&» — логическое «И» в пределах предложения;

«&&» — логическое «И» в пределах документа;

«|» — логическое «ИЛИ»;

«~» — логическое «И НЕ» в пределах предложения;

«~~» — логическое «И НЕ» в пределах документа. [6][6]

Контекстный поиск по ключевым словам. Средства контекстного поиска позволяют искать документы по содержащимся в них словам и фразам, которые могут объединяться логическими операциями. Резуль-

[6] В статье рассмотрены вопросы извлечения информации из документов различного формата (DOC, TXT, RTF, HTML), а так же алгоритмы поиска и индексации. Говорится о построении таблиц индекса, эффективной организации словаря, построении интерфейса поисковой системы.

таты поиска ранжируются по релевантности на основе частоты встречаемости слов запроса в найденных документах и во всей коллекции в целом.

Для обеспечения высокой скорости поиска по коллекции документов предварительно создается индекс, в котором для каждого слова устанавливаются ссылки на все документы, в которых оно встречалось. Однако самым большим недостатком подобной схемы является то, что необходимо выбрать правильный набор терминов. Кроме того, ввиду ограничения на размер индекса, документы индексируется не полностью, а лишь самыми «тяжелыми» терминами, которые не всегда точно отражают тематику документа.

Контекстный поиск объективно не может обеспечить ни точности ответа на запрос в виду наличия омонимии русского языка и отсутствие учета контекста, ни полноты ответа ввиду неоднозначности терминологии, наличии большого ряда синонимов, особенностей морфологии и т.д.

Контекстный поиск:
«( )» — группирование слов;
«/(n m)» — расстояние в словах: «-» — назад, «+» — вперед;
« «» » — поиск фразы;
«&&/(n m)» — расстояние в предложениях: «-» — назад, «+» — вперед.

Нечеткий поиск. Технология нечеткого поиска позволяет расширять запрос близкими по написанию словами, содержащимися в коллекции документов, по которым ведется поиск. Алгоритм нечеткого поиска способен найти все лексикографически близкие слова, отличающиеся заменами, пропусками и вставками символов. Нечеткий поиск целесообразно применять при поиске слов с опечатками, а также в тех случаях, когда возникают сомнения в правильном написании — фамилии, названия организации и т.п.

Существует несколько подходов для реализации нечеткого поиска: классические алгоритмы основанных на вычислении расстояния между строками (К- несовпадений и К- различий Ландау-Вишкина); алгоритмы ассоциативного поиска похожих слов, основанные на анализе цепочек символов слова; алгоритмы, основанные на различных возможностях написания слов (применяется для английского языка).

Развитие данных алгоритмов следует связывать с применением при анализе морфологических особенностей языка и различных шаблонов построения слов, учет статистической информации о частоте встречи

словоформы в документах, разработке и исследовании различных хэш-функций и алгоритмов ассоциативного поиска. [7][7]

Семантический поиск. Семантическая сеть — это множество понятий (слов и словосочетаний), связанных между собой. В семантическую сеть включаются наиболее часто встречающиеся слова текста, которые несут основную смысловую нагрузку. Для каждого понятия формируется набор ассоциативных (смысловых) связей, т.е. список других понятий, в сочетании с которыми оно встречалось в предложениях текста.

Семантическая сеть применяется на двух этапах поиска. Во-первых, после ввода запроса, входящие в него слова дополняются связанными с ними по смыслу словами (синонимами, вариантами написания, аббревиатурами и т.п.). Это позволяет находить и те документы, в которых фигурирующая в запросе идея выражена иначе.

В некоторых приложениях, для построения семантической сети используются несколько категорий отношения – по типу «вид-род», «общее-частное» и т.д.

Несмотря на то, что этот подход является скорее инструментом аналитика, чем поисковым механизмом, используемые в нем концепции анализа текста может служить сильным инструментом для повышения релевантности поиска информации в полнотекстовых документах. Дальнейшее развитие данного подхода тесно связано с задачей представления знаний в электронном виде и задачах автоматического анализа текста.

Перспективным развитием данного подхода является отказ от векторной модели представления образа документов, т.е. представления поискового образа документа в виде вектора, размерностью словаря системы с 1 или 0 соответствующими наличию или отсутствию термина в документе, и переход к семантической модели. Данный подход должен базироваться на алгоритмах построения семантических сетей текста и сохранять информацию о документе именно в виде сети смысла. Запрос и поиск информации в этом случае может быть выражен в терминах нечеткой логики и теории графов. Запрос пользователя будет представлять собой семантическую модель знаний пользователя в интересующей его предметной области. База данных будет хранить объективную семантическую модель предметной области (а точнее модель, которая будет составлена на основании коллекции документов). Задача

[7] В работе рассмотрены основные концепции построения информационных поисковых систем и сферы их применения. Проанализированы существующие подходы к решению задач поиска и предложено их улучшение, основанное на использовании семантических сетей при поиске информации в полнотекстовых базах данных.

поиска будет являться задачей нахождения нечеткого вхождения семантической сети знаний пользователя в семантическую сеть документа. В случае нахождения подобного изоморфизма документ считается релевантным запросу. Степень релевантности может определяться степенью нечеткого вхождения одного гиперграфа в другой.[7][8][9][8]

Средства повышения точности и полноты поиска. Повышение качества поиска следует связывать с применением тезауруса для расширения запроса синонимичными значениями и с учетом морфологии языка, для поиска, содержащего все грамматические формы слов запроса.

Применение морфологического анализа языка запроса и документа позволяет решать задачи определения всех грамматических характеристик слова, приводить различные грамматические формы слова к нормальной форме, получать все грамматические формы слова.

Использование синтаксического анализа текста на русском языке позволяет решить задачи грамматического разбора предложения с построением дерева синтактико-семантических зависимостей между его словами, выделения понятий предложения с определением их синтаксических и семантических ролей, генерация канонической формы понятий с использованием тезауруса. По результатам анализа может быть получена следующая информация: все слова с указанием части речи и синтаксической роли в предложении; все слова, синтаксически подчиненные выбранному слову, с указанием типа синтактико-семантической связи; все понятия текста, соответствующие выбранному слову, в канонической форме.

Использование тезауруса в информационно-поисковых системах призвано повысить качество анализа текста и полноту поиска информации. Так, отсутствие полноценного тезауруса русского языка в поисковом модуле не позволяет задействовать все возможности расширения запроса при поиске, как то: расширение слов запроса синонимичными, более общими или более частными, родственными по смыслу понятиями.

Применение комбинации тезаурус + морфологический словарь позволяет увеличивать количество слов в запросе на несколько порядков и повысить релевантность найденной информации за счет более интеллектуального подхода к запросу. [7]

[8] Статья посвящена иллюстрации основных возможностей технологии анализа и визуализации содержания текста на основе семантических сетей, которая предназначена, прежде всего, для решения экспертных задач, связанных с поиском информации в больших массивах документов.

Дополнительные требования. Сегодня к поисковым системам предъявляется ряд дополнительных требований, таких как построение запроса на естественном языке, поиск информации не только по формально заданным терминам, но и автоматический анализ, и расширение запроса, возможность создания сложных запросов с итеративным и интерактивным уточнением его параметров, интеллектуальное ранжирование выдаваемой информации.

Для поиска информации на основании запроса на естественном языке предварительно необходимо построить модель знаний пользователя о предметной области. Эта модель строится на основании вводимого пользователем текста запроса с использованием алгоритмов семантико-синтаксического анализа. Построение подобной модели позволяет более точно указать, что именно ищет пользователь.

Для расширения границ поиска с целью охвата всей предметной области служат различные алгоритмы, связанные с применением тезауруса. Знания о языке в целом и предметной области в частности представляются в виде фиксированной (а иногда динамичной) семантической сети с выделенными отношениями между понятиями (терминами). Используя знания о тезаурусе предметной области, на основании построенной по запросу пользователя модели знаний, можно сформировать расширенный запрос, включающий в себя ряд терминов, отсутствующих в запросе, однако принадлежащих синонимическим рядам и предметной области, указанной в запросе. Применение подобного подхода позволяет существенно повысить полноту поиска.

После осуществления поиска выполняется ранжирование списка найденных документов. Ранжирование осуществляется путем оценки релевантности документов запросу пользователя. Способы оценки релевантности, основаны на сравнении количества совпадающих слов в запросе и на странице с учетом местонахождения слов. Этот подход не всегда дает хороший результат, особенно в случае если запрос был сложным и содержал многозначные понятия. В этом случае целесообразно применять оценку релевантности запроса с использованием аппарата нечеткой логики для сравнения семантических сетей запроса и документа. Подобный подход оценивает содержимое не по формальному наличию искомых слов из запроса, а по смысловому содержанию страницы и его соответствию смысловому содержанию запроса. [10][9] [11][10]

[9] В статье рассмотрены поисковые системы в исторической перспективе, алгоритмы поиска, лингвистические приемы, проблемы качества ранжирования.

[10] В статье описан необходимый минимум возможностей для работы с русским языком, без которого поисковая машина просто не может функционировать, затем сделан

Выбор обобщенного алгоритма. В разрабатываемом модуле для наиболее эффективного поиска нужных статей сначала используем булев поиск, отсекая слова, которых нет в таблице слов. Затем используем семантический поиск, при помощи которого отбираются релевантные документы, на основании того, что у каждого слова стоит свой коэффициент релевантности, чем он выше, тем слово или словосочетание наиболее подходит. В итоге общий коэффициент релевантности отражает все найденные релевантные слова и словосочетания.

Семантическая сеть словаря русского языка включает в себя около 40 тысяч семантических групп в базовом варианте, однако в разрабатываемом модуле будет использовано гораздо меньшее количество семантических групп, т.е., мы имеем ограниченную тему для поиска. Использование семантической сети позволяет пользователю просто ввести поисковый запрос на естественном языке, предоставив системе самой искать все документы, контекст которых совпадает с контекстом запроса. Используемые технологии позволяют распознать слово в любой грамматической форме.

База знаний и словарь

База знаний (БЗ) - совокупность программных средств, обеспечивающих поиск, хранение, преобразование и запись в памяти ЭВМ сложно структурированных информационных единиц (знаний), где знания — совокупность сведений, образующих целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете, проблеме и т.д. [5]

Для того чтобы логический вывод, выполняемый в БЗ, давал правдоподобные результаты, необходимо, чтобы БЗ удовлетворяла следующим требованиям:

1. БЗ должна быть построена на основе знаний эксперта – профессионала в конкретной предметной области;
2. БЗ должна быть полной, то есть предусматривала возможность получения правдоподобного ответа в конкретной предметной области;
3. БЗ должна быть непротиворечивой, то есть на заданные вопросы она должна давать непротиворечивые ответы;
4. в ходе экспертного опроса должны использоваться методы, естественные для системы обработки информации. Ответы эксперта

переход к стандартному набору поисковых средств, предназначенных для работы с русским языком, а в конце кратко описана близкая к идеалу поисковая машина.

необходимо проверять на непротиворечивость.[12]

База знаний, содержащая информацию о статьях расхода, представляет собой набор связанных таблиц реляционной базы данных, реализованной средствами СУБД Microsoft Access. Пополнение и редактирование БЗ осуществляется средствами СУБД.

Эффективная организация словаря. Одна из самых важных и трудных проблем индексации текстов связана с созданием и пополнением словаря ключевых слов. Главная сложность ее заключается в том, что для эффективной работы системы необходимо рассматривать только базовые словоформы ключевых слов. При этом индексация даст лучшие результаты при обработке текстов, так как, например, неважно, в каком падеже встретилось в тексте слово.

Существует два подхода к решению этой проблемы. Первый состоит в реализации механизма выделения базовых словоформ из всех слов языка, и ему отвечает словарь небольшого размера. При индексации документа или обработке запроса происходит выделение из текста документа или запроса базовых словоформ всех слов. Очевидно, что механизм выделения для большинства языков крайне не прост, и это усложняет применение данного метода. При использовании подобного подхода велика также вероятность образования ошибочных словоформ.

Для выделения базовых словоформ существует ряд алгоритмов. Простейшим из них является метод отсечения всех знакомых системе приставок и суффиксов/окончаний. Из более продвинутых методов стоит упомянуть прием, называемый «квадратом Джозефа Гринберга», который предложил в графическом представлении размещать по двум противоположным сторонам квадрата (или ромба) английские слова, например sleeps и eats, по двум другим противоположным сторонам — слова sleeping и eating. То же можно записать в виде пропорции *sleeps:eats = sleeping:eating*. Анализ этой конструкции позволяет выделить две базовые словоформы *sleep u eat* и окончания *s u ing*.

Американский языковед З. Харрис предложил вероятностный подход к решению проблемы выделения базовой словоформы, граница которой определяется по наибольшему числу совпадений букв в сравниваемых словоформах.

Другим часто используемым алгоритмом выделения базовых словоформ является метод Портера, разработанный специально для поисковых систем и изначально применявшийся для английского языка. Идея алгоритма состоит в создании для каждого используемого языка правил отсечения окончаний (в широком смысле) слов. Каждое из них имеет вид:

(условие) S1->S2, где *S1* — окончание слова, которое необходимо удалить, *S2* — новое окончание слова. Условия, при которых применяется данное правило, являются логическими функциями следующих событий: **S* — базовая словоформа заканчивается на S (здесь S — любая буква); **v** — базовая словоформа содержит гласную; **d* — базовая словоформа заканчивается на двойную согласную и т. д.

В качестве примера приведем следующие правила, построенные по этой схеме:

*SSES->ES # caresses->caress IES->I # ponies->poni (*v*) ED-> # plastered->plaster (*v*) ING-> # sing->sing*

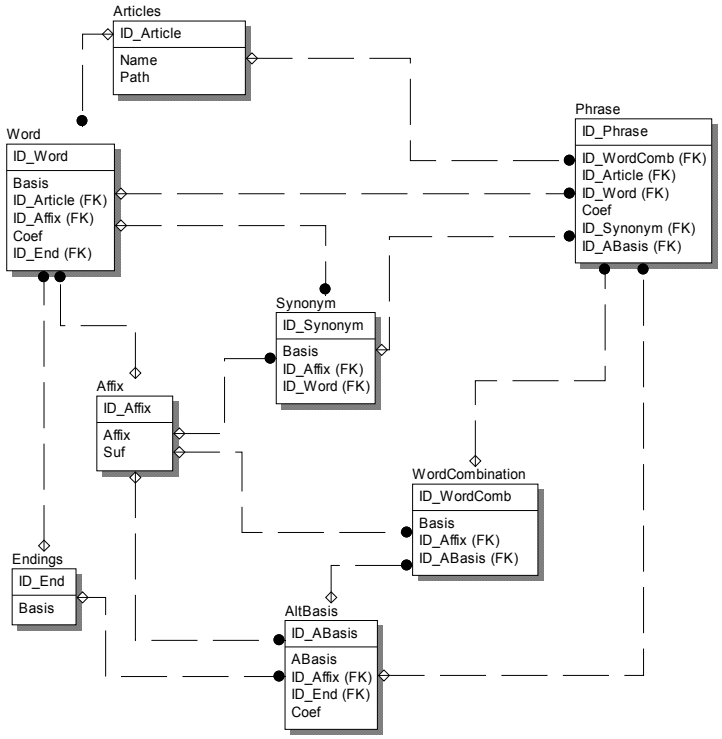
Существуют другие подходы к решению проблемы словаря. Например, создается фиксированный набор слов языка с указанием для каждого слова его базовой словоформы. Существенным недостатком данного метода является большой объем словаря, в котором приходится хранить не только базовые словоформы, а все используемые слова. Однако это позволяет построить более точное соответствие между словами и их базовыми словоформами. Кроме того, появляется возможность ограничить словарь некоторым набором базовых слов, созданным при его инициализации. Во время индексации можно игнорировать неизвестные слова (в случае, если исходный словарь уже велик). Это позволит отсечь большое количество «мусора» — ошибочных и служебных слов, появление которых неизбежно. Наряду с этим могут быть пропущены слова, важные для индексации, преимущественно специальные выражения и термины предметных областей. Чтобы избежать этого, предлагается добавлять новые слова, но без учета образования от них словоформ. Такой подход дает особенно хорошие результаты при добавлении в словарь аббревиатур, не требующих образования от них дополнительных словоформ.

Еще одна проблема индексирования связана с выявлением и удалением из текста так называемых стоп-слов. Они не несут смысловой нагрузки в текущей предметной области, и для эффективной работы системы их следует удалять при индексировании. Как правило, стоп-словами являются предлоги, союзы, артикли, вводные слова и т. п. Они очень часто встречаются в документах, но малоинформативны. Для их удаления можно либо использовать отдельный словарь стоп-слов, либо считать все слова с высокими частотами встречаемости в базе данных текстов стоп-словами и удалять их при индексировании.

Итак, при выделении слов будем использовать процедурный метод обработки лексем. Каждое слово разделяется на основу и аффикс (окончание и, возможно, суффикс). Словарь содержит только основы слов

вместе со ссылками на соответствующие строки в таблице возможных аффиксов. Основной критерий при разбиении слова на основу и аффикс - основа должна оставаться неизменной во всех возможных словоформах данного слова. Поскольку большое количество слов русского языка имеет одни и те же аффиксы, то суммарный объем словаря основ и словаря аффиксов оказывается значительно меньше, чем объем полного словаря всех словоформ, используемого в декларативных методах. [6],[12][11]

В итоге получаем следующую схему БД:



Алгоритм разбора запроса

Входными данными является название договора, которое вводит пользователь. Название вводится в естественно-языковой форме.

[11] Рассматривается использование возможностей SQL-запросов к реляционным базам данных при осуществлении поиска. А также вопрос о том, что следует хранить в базе данных, а точнее, в полях таблицы базы данных.

Выходными – набор слов и словоформ, входящих в название.

Результатом работы ОЕЯ — процессора является массив значений параметров выделяемых объектов, к которому в блоке трансляции во внутренний язык многократно применяются правила комбинации слов. В блоке трансляции в язык SQL путем сравнения запроса с шаблонами и применения правил для обработки исключений запрос во внутреннем языке представляется инструкциями языка SQL.

Задачу преобразования естественно-языкового предложения (или текста) в некоторый набор семантических структур, являющихся формальным представлением "смысла" исходного предложения или текста, выполняет лингвистический процессор (ЛП). Цель такого преобразования - сформировать исходные данные для работы стандартных поисковых механизмов СУБД. Поэтому нужно чтобы запрос, сформулированный на ОЕЯ, мог быть приведен к языку SQL (рис. 1).

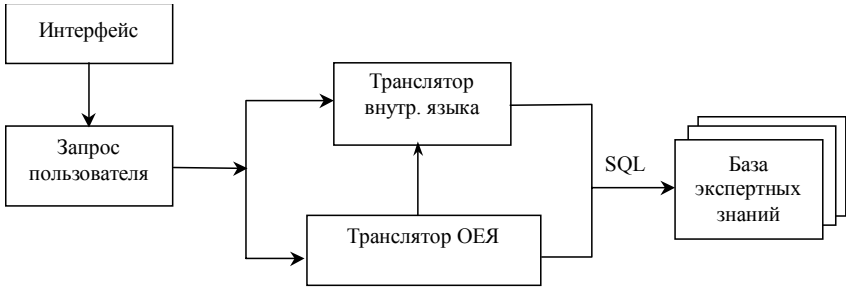


Рис. 1 Схема преобразования запроса пользователя

В настоящее время принято употреблять пробелы между словами, но выделение слов – не такая простая задача, как кажется.

Обычно словом считается последовательность букв «от разделителя до разделителя». Разделителями считаются пробелы, знаки препинания и прочие неалфавитные символы. Однако имеются символы, которые, не будучи буквами, не являются и разделителями в полном смысле этого слова. К ним относятся, в частности, апостроф «‘» и дефис «-». Действительно, иногда дефис разделяет части слова (кто-нибудь, сине-зеленый), а иногда – самостоятельные слова (мать-одиночка). Апостроф, особенно в английском языке, может являться самостоятельной частью слова, а может использоваться просто в качестве кавычек. [14][12]

[12] В статье описано решение задачи преобразования естественно-языкового предложения (или текста) в некоторый набор семантических структур, являющихся формальным

Итак, при разборе запроса был создан следующий алгоритм. Предложение разбивается на слова, с учетом всех разделителей. Из слов выделяется основа, которая ищется в таблице словоформ, если в словоформах такого слова нет — значит, выбрасываем его из запроса. Также выкидываются, так называемые стоп-слова и слова паразиты.

Разрабатывать списки шумовых слов не нужно, т. к. поиск будет вестись только по известным словам.

Интерфейс для заполнения БЗ

Процесс диалога можно определить как обмен информацией между вычислительной системой и пользователем, вводимый с помощью интерактивного терминала и по определенным правилам. Он может рассматриваться как оболочка, включающая все процессы, входящие в систему, по выполнению определенных знаний. Процессы ввода-вывода обеспечивают обмен на самом верхнем уровне, на этом уровне диалоговый процесс должен правильно интерпретировать каждое слово. Задача диалогового процесса — определить задание, которое пользователь возлагает на систему. Важны правила ведения диалога с компьютером. Разработка интеллектуального интерфейса направлена на повышение согласования физических возможностей системы и человека, согласования стиля диалога с потребностями и представлениями человека. Отделение процессов диалога от процессов выполнения задания позволяет использовать один и тот же процесс для выполнения различных заданий пользователя путем включения его в разные процессы диалога. Диалог можно классифицировать с учетом формата входных сообщений и гибкости, позволяющей пользователю вводить входные сообщения, когда ему это необходимо. Входное сообщение позволяет: выбрать режимы диалога, выбрать нужный процесс выполнения задания, ввести данные для выполнения задания. Существуют различные виды диалогов. Диалог, управляемый системой, более удобен тем, что он лучше подстраивается под пользователя, но при этом имеет больше ограничений, чем диалог, управляемый пользователем. Сообщения на ограниченном естественном языке наиболее часто используются для построения диалогов, управляемых системой. Однако, при определенном подходе, использование их для построения диалогов, управляемых пользователем, так же дает неплохие результаты. При использовании ограниченного естественного языка распознается лишь небольшое количество слов и в каж-

представлением "смысла" исходного предложения или текста с целью сформирования исходных данные для работы стандартных поисковых механизмов СУБД.

дый конкретный момент входное сообщение состоит из очень небольшого числа слов. Так как эти слова могут обозначать конкретные задания или являться данными (в том числе и числовое значение), их можно использовать как для контроля, так и для ввода данных. Цель использования естественного языка – естественное, наиболее удобное для пользователя ведение диалога с системой, как с человеком. Такая система будет реагировать на любую фразу или синтаксическую конструкцию, понятную человеку. В системах управляемых пользователем диалог наиболее гибок и более естественен для пользователя, но в тоже время более сложен для создания его. Используя ограниченный, а в особенности неограниченный естественный язык, в определенной предметной области, наиболее удобен обоюдный способ ведения диалога. В диалоговых системах специфика предметной области, как правило, накладывает сильные ограничения на состав используемых понятий, на способы интерпретации понятий, на лексику языка и т.д., что значительно упрощает задачу обработки естественного языка. Поэтому и появляется возможность говорить об использовании неограниченного естественного языка. Сложность методов анализа и синтеза естественного языка зависит не только от языка общения, но и от языка, используемого для представления знаний. На начальном этапе диалога язык общения может быть несколько более жестко формализован определенным набором запросов системы и множеством различных ответов пользователя. Обработка сообщений пользователя сводится к анализу входных сообщений. В этих условиях задача синтеза может сводиться к генерации подготовленных заранее, в определенных границах, вопросов, а задача анализа – к обработке слов, словосочетаний и предложений требующей проведения семантического, синтаксического анализов.

Интеллектуальный интерфейс (ИИн) можно подразделить на внутренний и внешний. При внутреннем ИИн организуется взаимосвязь между интеллектуальными устройствами и (или) интеллектуальными программами. Внешний ИИн – это интеллектуальный человеко-машинный интерфейс, т.е. комплекс интеллектуальных программных средств, обеспечивающих взаимодействие пользователя с системой. Рассмотрим некоторые вопросы внешнего ИИн. Сложность решаемых задач в процессе общения пользователя с ИИн, приводит к необходимости выделения в структуре ИИн диалоговой компоненты. Функции диалоговой компоненты способствуют осуществлению анализа (семантического, синтаксического) естественного языка. Назначение диалоговой компоненты заключается в приемлемой организации обоюдного диалога "пользователь – интеллектуальный интерфейс", в обозначении функций

участников общения в ходе совместного решения задачи. В том числе организация возможности, например, пополнения баз данных и (или) баз знаний для наиболее качественного (и вообще возможности) проведения семантического и синтаксического анализов естественного языка. [15]

В качестве базы данных был взят готовый тезаурус по экономике. Формат файла базы данных DBF. Работа с этим форматом возможна при помощи Microsoft Access, после осуществления связи с базой данных из MS Access. Его объема достаточно для работы.

Литература

1. *Бронникова Т.С., Чернявский А.Г.* Стратегия маркетинга, планирование и контроль. Маркетинг. Учебное пособие [online]. 1999 г. [20.10.04], http://www.aup.ru/books/m49/14.htm
2. *Степанов И.С.* Экономика строительства. – Юрай-Издат, 2004. – 620 с.
3. Смета бюджетного учреждения. Школа бухгалтера [online]. Март 2003 г. №10/2003 [26.10.04], http://www.dtkt.com/school/rus/2003/10/10sc6.html#1
4. Приказ о порядке распределения расходов по соответствующим предметным статьям и подстатьям экономической классификации расходов бюджетов РФ (в редакции приказов Минфина РФ от 20.05.03 г. №45н, от 12.08.03 г. №73н, от 11.12.03 г. №115н))
5. *Аверкин А.Н., Гаазе-Рапопорт М.Г., Поспелов Д.А.* Толковый словарь по искусственному интеллекту. Российская ассоциация по искусственному интеллекту [online]. [20.10.04]. www.raai.org/library/tolk/aivoc.html
6. *Аграновский А. В., Арутюнян Р. Э.* Индексация массивов документов. Издательство «Открытые системы. Мир ПК» [online]. 11 июня 2003 г. №06/2003 [26.10.04], http://www.osp.ru/pcworld/2003/06/049.htm
7. *Андриенко Е.В.* Концепции поиска адекватной информации в полнотекстовых базах данных. Электронный журнал «Перспективные информационные технологии и интеллектуальные системы» [online]. №03/2003 [26.10.04], http://pitis.tsure.ru/files15/13.pdf
8. Промышленная информационно-поисковая система Convera RetrievalWare. Аналитико-стратегические технологии ОДЕОН [online]. [26.10.04] www.odeon-ast.ru/ru/products/rware.php
9. *Ермаков А.Е., Плешко В.В.*Семантическая сеть текста в задачах аналитика. Публикации [online]. 2002 г. [23.10.04],

http://www.metric.ru/publications.asp?ob_no=308&#pic1

10. *Сегалович И.В.* Как работают поисковые системы. Международная конференция «Диалог» [online]. 2003 г. [23.10.04], www.dialog-21.ru/direction_fulltext.asp?dir_id=15539
11. *Игорь Ашманов.* Национальные особенности поисковых систем. Издательство «Открытые системы. Мир ПК» [online]. 19 января 2000 г. №01/2000 [26.10.04], www.osp.ru/school/2000/01/012.htm
12. Построение базы нечетких знаний продукционного типа на основе определения состояний объекта. Электронный журнал «Перспективные информационные технологии и интеллектуальные системы» [online]. №03/2003 [26.10.04], http://pitis.tsure.ru/files2/p29.htm
13. *Моисеенко С.И., Майстренко А.* Релевантность полнотекстового поиска: подход на основе построения терминологической базы документов. Каталог аналитических статей [online]. 2001 г. [22.10.04]. www.citforum.ru/database/articles/rel_search.shtml
14. *Комарцова Л.Г.* Проблемы разработки пользовательских интерфейсов с базами экспертных знаний в интеллектуальных системах. Международная конференция «Диалог» [online]. 2003 г. [21.10.04], http://www.dialog-21.ru/Archive/2003/Komartsova.htm
15. *Блувштейн Д. В.* К вопросу о создании интеллектуального интерфейса с организацией обработки запроса на естественном языке. Международная конференция «Диалог» [online]. 2004 г. [23.10.04], http://www.dialog-21.ru/Archive/2004/Bluvshtein.htm

Ван Чэнь, Гуань Ли

КИТАЙСКИЕ МЕТАФОРЫ ИНФОРМАТИКИ

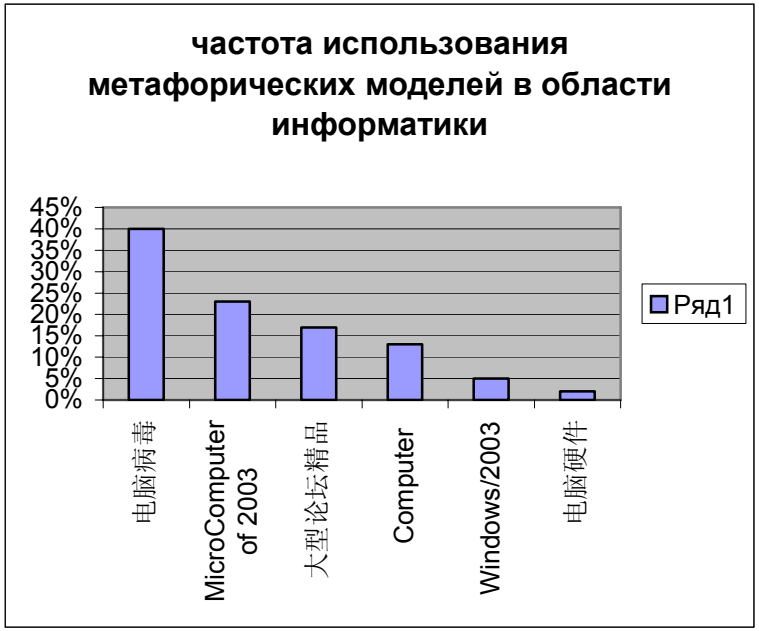
Анализ китайских компьютерных изданий

Программам автоматического перевода (АП) пока не удается осуществлять перевод на должном уровне. Программы не подстраиваются под предметную область, и поэтому плохо переводят технические тексты, но технические тексты сложны не только своей терминологией, часто в них встречаются метафоры. Метафоры, встречающиеся в иностранном техническом тексте, бывают двух видов: метафоры, сходные с метафорами в русском языке, и метафоры иностранного языка. Особенно тяжело понять метафоры иностранного языка. Сам компьютер не идентифицирует метафору, поэтому, если у программы будет подгружаемый словарь метафор, который при переводе будет выдавать либо объяснение метафоры, либо аналог данной метафоры в русском языке, либо, если ни то, ни иное не требуется, просто правильный перевод, то качество перевода может немного улучшиться, а программа сможет расширить круг пользователей.

Метафора — а) в широком смысле — всякое иносказание, образное выражение понятия; б) употребление слова или выражения в переносном смысле.

В технической литературе метафора имеет очень большое значение: делает высказывание образным и доступным. Каждая метафора – определенный образ, который, с одной стороны упрощает восприятие текста, с другой стороны, если это метафора иностранного языка, еще больше запутывает читателя.

Чтобы выявить особенности метафор в предметной области ИВТ, были проанализированы статьи из компьютерных журналов на китайском языке. Не все журналы используют метафоры в большом количестве, обычно это журналы, рассчитанные на профессионалов, журналы для пользователей, напротив пестрят метафорами.



Как видно из представленной диаграммы чаще всего метафорические модели встречаются в журнале «MicroComputer of 2003». В этом журнале метафоры используются в статьях разного вида, как в статьях, посвященных новинкам в области ИВТ, так и в разделах по ремонту ПК, наверное, именно этим объясняется такая популярность «MicroComputer of 2003», читатели которого, обладают совершенно разными уровнями знаний ПК.

Стоит заметить, что в текстах, темами которых являются компьютерные вирусы, метафорические модели употребляются в большом количестве.

Основными метафорическими моделями в области ПК являются: животное, война, организм, транспортное средство, действие, водное пространство, болезнь. Если рассматривать все метафорические модели в целом, то получается, что самые употребляемые метафорические модели — это одушевленные существа.

Но, поскольку мы также рассматриваем метафорические модели иностранного языка, проведем их классификацию отдельно. Среди метафор китайского языка, есть те, что дословно переводятся схожей ме-

тафорой в русском языке, но имеют совершенно иное значение. Вот один из примеров: китайское выражение 菜鸟(指电脑初学者) дословно переводиться как *огородная птица* (компьютерный профан). Многие метафорические модели китайского языка не имеют аналогов в русском, но по их переводу можно догадаться о значении.

Рабочие материалы для словарных статей

Материалы для построения словарных статей Словаря китайских метафор информатики и вычислительной техники имеют следующую структуру:

<№> – <**Китайская метафора (троп)**> – <*Русский перевод*> – <Пояснение> – <Китайский контекст. Сведения об источнике. Русский перевод контекста>

1 — **网虫** — *сетевой червь* — Человек, который очень часто сидит в Интернете — 我是超级**网虫**，什么样的困难也不能阻止我上网。http://happy.enet.com.cn/html/20041119110001.html Я супер **сетевой червь,** ничто не может помешать мне сидеть в Интернете.

2 — **菜鸟** — *огородная птица* — Человек, который очень мало знаёт о компьютере — 她是**菜鸟**，她不知道怎样安装 Windows. http://article.pchome.net/2004/09/13/26384.htm Она **огородная птица**, она не знаёт, как установить систему Windows.

3 — **大虾** — *большая креветка* — Человек, который очень хорошо знаёт всё о компьютере —
如果我有与电脑有关的问题，我就问他，因为他是**大虾**。http://bbs.zol.com.cn/new/TopicView_new2.php?nId=344323&board_id=21 Если у меня появятся вопросы по компьютеру, я спрошу его, потому что он **большая креветка.**

4 — **上网冲浪** — *сёрфинг по Интернету* — Работать в Интернете — **上网冲浪**之前，掌握一些网络术语是必须的。http://www.pep.com.cn/200406/ca479609.htm Перед тем, как устраивать **сёрфинг по Интернету**, нужно овладеть компьютерными терминами.

5 — **被人黑了** — *был чёрный* — Hacker взломал мой компьютер — 我想我大概也**被人黑了**，我决定格式化硬盘重新安装操作系统. http://db.kingsoft.com/c/2004/12/17/161170.shtml Я думаю, мой ком-

пьютер тоже **был черным**, я решил отформатировать жесткий диск и заново установить систему.

6 — **闪客** — *блёсткий гость* — Человек, который делает flash-анимацию — 他要做网站，需要**闪客**的帮助。 http://www.blogchina.com/idea2/flasher/ Он хочет сделать сайт, нужна помощь **блесткого гостя**.

8 — **猫** — *кошка* — Модем — 我是超级网虫，已经用坏了3个**猫**。 http://publish.it168.com/2004/1201/20041201505801.shtml?cPositionCode=141 Я супер сетевой червь**,** уже сломал три **кошки**.

9 — **枪手** — *ружейная рука* — Человек, который помогает писать программу — 现在有很多大学生找**枪手**， 帮助自己写论文。 http://www.54youth.com.cn/gb/paper247/1/class024700002/hwz189886.htm Сейчас много студентов ищут **ружейной руки**, чтобы помочь себе писать дипломную работу.

10 — **上网** — *сидеть в Интернете* — Работать в Интернете, пользоваться Интернет — ADSL高速宽带网在使用过程中，有时还不如原来的窄带MODEM(调制解调器)**上网**速度快。 http://it.gz163.cn/app/2004-12-20/19534.htm Иногда **сидеть в Интернет** через ADSL еще медленнее, чем через обычный модем.

11 — **发烧友** — *горячий человек (человек-жара, болельщик)* — Человек, приобретающий самые новые и хорошие компьютерные комплектующие — 一般**发烧友**3个月换一台新电脑。 http://www.pconline.com.cn/diy/salon/tjpz/10000/4/0410/481339.html Обычно **горячий человек** каждые три месяца меняет новый компьютер.

12 — **网民** — *сетевой человек (человек в сети)* — Человек, который находится в Интернете — 目前我国上网用户达到7950万，其中男性**网民**占60.4%，女性**网民**占39.6%。 http://tech.sina.com.cn/i/w/2004-04-07/1414345533.shtml Сейчас в Китае пользуются Интернетом уже 795 миллионов человек, среде них мужских **сетевых человек** 60,4%, женских **сетевых человек** 39.6%.

13 — **飞盘** — *отлетающий диск* — Ошибка при записи CD — 现在买CDR盘，即使**飞盘**了，也可以拿来换新的. http://www.ezit.com.cn/pub/article/c1585_a156756_p1.html Сейчас при покупке CDR **отлетающий диск** можно обменять на новый.

14 — **D盘** — *D диск* — Нелицензионный диск, “левый” диск — 在北京卖**D盘**的地方越来越少了。

http://game.xaonline.com/ShowBody.asp?SubID=LTJX&MsgID=270 В Пекине все меньше продаётся **D дисков**.

15 — **网游** — *сетевая игра* — Сетевая игра — **网游**制造了中国最富的人… http://news.xinhuanet.com/fortune/2004-12/15/content_2336081.htm **Сетевая игра** сделала самым богатым человека в Китае…

16 — **黑客** — *чёрный гость (хакер)* — Человек, незаконно проникший в чужой компьютер — **黑客**是以制造病毒并传播病毒、破坏网络安全为目标的。 http://www.h2k2.com/jianjie.asp **Чёрные гости** занимаются тем, что создают и распространяют вирусы.

17 — **防火墙** — *брандмауер* — Брандмауер Firewall — 为了安全的需要，我们大多数是在机器上安装有杀病毒软件或软件**防火墙** http://content.sina.com/news/64/62/7646296_1_gb.html В целях безопасности многие из нас установили на компьютеры антивирусы и **брандмауеры.**

18 — **超频** — *увеличение частоты (разгон)* — Повышение частоты CPU — 赛扬CPU很容易**超频**。 http://info.it.hc360.com/HTML/001/001/002/61183.htm Процессоры Celeron легко **разгонять**.

19 — **死机** — *мёртвая машина,* — Компьютер не может продолжать работу — 电脑容易**死机**，往往是由于硬件过热造成的。 http://myhard.yesky.com/DIY/360853119166316544/20041220/1890939.shtml Компьютер легко становится **мертвой машиной**, так как часто перегревается жесткий диск.

20 — **U盘** — *и-диск* — Flash диск — 有**U盘**，就不需要软盘了。 http://digi.163.com/2004w12/12769/2004w12_1103273588573.html Если есть **U-диск**, то дискеты не нужны.

21 — **网络对战** — *сетевая война* — Играть в войну по Интернету — 通过互联网和电脑，即使是不同操作系统的人们，也将可以实现**网络对战**。 http://www.pcpop.com/game/04/12/54653.shtml По Интернету можно играть в **сетевую войну** даже людям с разными системами на компьютерах.

22 — **灌水** — *полив* — Рассылка ненужной компьютерной информации — **灌水**是网络邮件的缺点。 http://tech.china.com/zh_cn/news/humor/10002547/20041216/12017567.html Рассылка **полива** — это один из недостатков почты по Интернету.

23 — **顶** — *толкание (поддержка, оценка)* — Поддерживать чью-либо идею, мнение в Интернете — 对于我的看法他**顶**，可见他同意了。 http://forum.runsky.com/bbs/viewthread.php?tid=121852&fpage=2 Он **толкал** мою идею, и значит он согласен.

24 — **贴帖子** — *наклеить записку* — Высказать свое мнение в Интернете — 我们一起写文章一起**贴帖子**，享受着网络带给我们的轻松愉快的时光. http://www.jiaodong.net/wenyu/text/sfyc.asp?No=50698 Мы вместе пишем статью, **наклеим записку**, насладимся добрым временем в Интернете.

25 — **网恋** — *сетевая любовь* — Выражать свою любовь через Интернет — 今年7月初，他们在网络游戏中相识并开始了**网恋** http://www.donews.com/donews/article/7/71538.html В начале июля этого года, они познакомились в сетевой игре и завязали **сетевую любовь.**

26 — **后门** — *задняя дверь* — Слабое место в компьютерных программах — 病毒一旦侵入用户电脑后，会在上面开设**后门**，黑客们可以远程控制这些中毒电脑 http://www.fj.xinhuanet.com/jjpd/2004-12/21/content_3428904.htm Однажды проникнув в компьютер пользователя через **заднюю дверь**, вирус позволяет хакеру управлять им издалека.

27 — **补丁** — *заплатка (защита)* — Закрыть слабое место в компьютерных программах — 微软发布的两个**补丁**是Windows XP SP2的 http://tech.163.com/2004w12/12772/2004w12_1103506351488.html Microsoft предоставляет два Windows XP SP2 в качестве **заплатки**.

28 — **网管** — *сетевой заведующий* — Сетевой администратор — **网管**的任务是保障网络的正常用作。 http://www.e-works.net.cn/ewk2004/ewkArticles/413/Article26926.htm Задачей **сетевого заведующего** является поддержка нормальной работы сети.

29 — **升级** — *повышение* — Upgrade , обновить — 只要在网上定购Inspiron 600m，就可以免费将屏幕**升级**到SXGA液晶显示屏 http://www.enet.com.cn/emk/inforcenter/A20041221374181.html В Интернете достаточно купить компьютер Inspiron 600m, и можно бесплатно **повысить** монитор до ЖК-монитора.

30 — **版竹** — *бамбук страницы* — Администратор сайта или страницы — **版竹**应定期发表一些高质量的帖子。 http://www.xf18.com/cgi-bin/bbs/topic.cgi?forum=4&topic=3&show=0 **Бамбук страницы** своевременно выкладывает на сайте полезную информацию.

31 — **网友** — *сетевой друг* — Друг в Интернете — 欢迎胡彦斌作客TOM访谈，先跟我们的**网友**打个招呼吧。 http://sports.tom.com/5093/5110/20041221-492695.html К нам присоединился “TOM”, давайте сначала поприветствуем наших **сетевых друзей.**

32 — **前台** — *сцена (рабочий стол)* — Рабочий стол — 新的微软操作系统Windows XP,**前台**做的很酷。 http://stock.bbn.com.cn/sharemodule/info_dtl_more.aspx?InfoId=963254 Новая операционная система Microsoft Windows XP, сделает **рабочий стол** очень крутым.

33 — **后台** — *за сценой* — Внутренняя часть компьютерной программы — 微软操作系统Windows XP的**后台**是很复杂的。 http://www.cbinews.com/inc/showcontent.jsp?articleid=11983 **Внутренняя часть** Microsoft оперативной системы Windows XP очень сложна.

34 — **激打** — *лазерный удар* — Печатать на лазерном принтере (Лазерный принтер) — 激**光打**印机的价格不菲，如今热销的低端入门型**激打**都在1500元左右 http://www.cnetnews.com.cn/news/hardwares/story/0,3800055190,39331310,00.htm **Лазерный принтер** обычно очень дорогой, но сейчас самая хорошая продажа по **лазерным принтерам** стоит только 1500 юаней.

35 — **墨打** — *чернильный удар* — Печатать на струйном принтере (струйный принтер) — 以往，**墨打**占据了大部份市场，但近年激光打有後来居上之势 http://news.pack.net.cn/newscenter/xzyc/2004-09/2004093011152554.shtml Раньше рынок в основном был наполнен **чернильными принтерами**, а сейчас все больше и больше лазерных принтеров.

36 — **E书** — *E книга* — E книга — 中国最大的电子图书库，提供数万本电子图书（**E书**）完全免费下载！ http://book.httpcn.com/search/ Самая большая китайская база данных в Интернете составит несколько тысяч **E-книг**, для бесплатного скачивания.

37 — **黑屏** — *чёрный экран* — Зависание компьютера — 游戏任何时候按Select+Strat+L1+R1+L2+R2复位后**黑屏** http://game.china.com/zh_cn/guide/tvguide/11005322/20041125/11981640.html В этой игре нажимают Select+Strat+L1+R1+L2+R2, после того как получается **чёрный экран.**

38 — **蓝屏** — *синий экран* — Зависание компьютера — 一般在CPU执行比较繁重的任务时，系统会突然出现**蓝屏** http://www.pconline.com.cn/diy/salon/cncd/study_cpu/0412/518972.html

Когда обычно CPU выполняет сложную задачу, система выводит **синий экран.**

39 — **超刻** — *негабаритное нанесение (превышение объема)* — Превышение объема записи на CD — 普通CD刻录机无法刻录700MB以上的碟片，不过听说市场上部分刻录机能较好地进行**超刻.** http://www.ezit.com.cn/pub/article/c1585_a161443_p1.html Обычно при записи на CD ROM нельзя записать CD－диск объемом больше 700MB, но сейчас на рынке продаются CD ROM, на которые можно сделать запись с **превышением объема.**

40 — **低格** — *низкое форматирование* — Низкое форматирование — 每次读写时硬盘都会出现“嘎嘎”的响声，朋友建议我**低格**它 http://it.enorth.com.cn/system/2004/11/30/000914370.shtml В случае, когда каждый раз при работе с жестким диском раздается скрежет, приятель посоветовал мне выполнить его **низкое форматирование.**

41 — **磁盘清理** — *вычистить диск* — Очистка диска от ненужных файлов — Windows操作系统内自带的**磁盘清理**工具能够清除电脑中的临时文件。 http://www1.people.com.cn/GB/it/1069/3023255.html Windows позволяет **вычистить диск** от неиспользуемых файлов.

42 — **网名** — *сетевое имя* — Сетевой имя — **网名**能很好的隐藏自己。 http://www.northeast.com.cn/kjnews/80200412210160.htm Под **сетевым именем** очень удобно маскироваться.

43 — **主页** — *главная страница* — Главная страница — 情你们在我们的网站**主页**上查找我们公司的电话。 http://www.pconline.com.cn/pcedu/tuijian/network/design/0412/517819.html Посмотрите, пожалуйста, телефон нашей фирмы на **главной странице** сайта.

44 — **搜索引擎** — *поисковый мотор (поисковая система)* — Поисковая система — 面对信息泛滥，有人发明了**搜索引擎**，能够抓出想要的网页信息。 http://it.sohu.com/20041214/n223484202.shtml Представлено так много информации, что кто-то выдумал **поисковый мотор,** чтобы можно найти интересную ему информацию.

45 — **远程控制** — *телеобработка (удаленный доступ)* — Удаленный доступ через Интернет — 黑客们可以**远程控制**这些中毒电脑，利用其发送垃圾邮件或攻击网站 http://www.fj.xinhuanet.com/jjpd/2004-12/21/content_3428904.htm Хакеры занимаются **удаленным доступом,** чтобы распространяют вирусы.

47 — **打包** — *упаковка* — Упаковать много файлов в один — 可以把很多文件**打包**到一个文件。 http://www.pep.com.cn/200406/ca479044.htm Можно **упаковать** много файлов в один.

48 — **压缩文件** — *сжимает файл (архиватор)* — Сжатие объема файла — 为了节省空间，在互联网中大量使用**压缩文件.** http://tech.163.com/2004w11/12747/2004w11_1101354426981.html Для экономии пространства мы можем применять **сжатие файлов.**

49 — **格机** — *формат* — Форматирование компьютера — 安装新系统之前，要**格机**… http://bbs.verycd.com/forum/lofiversion/index.php/t118722.html Перед установкой новой системы надо провести **форматирование**…

50 — **机子中毒** — *машина отравлена* — Компьютер заражен вирусом — 求救：**机子中毒**．每隔一段时间自动关机． http://pfw.sky.net.cn/bbs/viewthread.php?tid=8795&sid=wCoHaTEK Внимание, **машина отравлена** и будет выключена.

51 — **掉线** — *срывается линия (обрыв связи)* — Обрыв соединения — ADSL**掉线**涉及到多方面的问题，包括线路故障、ADSL Modem故障、网卡故障等等。 http://telecom.chinabyte.com/NetCom/218441117551558656/20041202/1882960.shtml **Обрыв связи** ADSL касается многих проблем, которые включают отказ линии, ADSL Modem, сетевой карты и т.д.

52 — **上线** — *подключиться к сети* — Войти в сеть, подключиться — 如果不**上线**下载防火墙的补丁，病毒随时可以进入你的电脑。 http://games.sina.com.cn/newgames/2004-11-29/170663716.shtml Если не удается **подключиться к сети,** и скачать эту защиту брандмауера, то твой компьютер заражен вирусом.

53 — **离线** — *отключиться* — Покинуть сеть, отключиться — 对网虫来说**离线**时，总是很痛。 http://flea.zol.com.cn/showinfo.php?id=1048947 Для сетевого червя **отключиться** настоящее несчастье.

54 — **软件包** — *сумка софтвера* — Программный пакет — 目前许多**软件包**（如mathcad，mathematic，maple）都能完成数学里各种各样的运算 http://www.cbe21.com/subject/maths/html/040202/2004_12/20041220_100199.html Сейчас множество **программных пакетов** (например: mathcad，mathematic，maple) могут выполнять разные математические вычисления.

55 — **主机** — *главная машина* — Системный блок — 我的**主机**太小，没有地方安装第二个硬盘。 http://inews.cnnb.com.cn:81/web/cnnb_asp/newscenter/gaojian.asp?id=25705 У меня очень маленький **системный блок**, некуда установить второй жесткий диск.

56 — **隐身** — *невидимка* — Другие пользователи не видят — 现在越来越多的QQ用户都选择了**隐身**上线 http://article.pchome.net/2004/11/12/30105.htm Сейчас все больше пользователей QQ выбирают **невидимку** в сети.

57 — **下线** — *снижается линия* — Выйти из сети — 由于用户太多，服务器无法承受，造成正常在线用户的**下线**。http://www.huawei.com.cn/news/tech/198/1653.shtml Сервер может не выдержать много пользователей, тогда нормальные пользователи **выйдут из сети.**

58 — **在线** — *на линии* — Online — 我天天都**在线**。 http://tech.sina.com.cn/i/2004-12-21/1850481043.shtml Я всегда **на линии.**

59 — **聊天室** — *комната для переговоров (чат)* — Место на сайте, где можно общаться — 最近我男朋友迷上了网路**聊天室**，他几乎把所有时间都花在那上面。 http://she.xicn.net/item/2004/12/21/91872.html Недавно друг увлекся **чатом**, он почти все время тратит там.

60 — **重启** — *повторится отпирание (перезагрузка)* — Перезагрузка — 早在DOS时代就有不少病毒能够自动**重启**你的计算机 http://it.com.cn/f/edu/0412/21/60801.htm Раньше в DOS было много вирусов, способных автоматически **перезагрузить** ваш компьютер.

61 — **开机** — *открыть машину (включить)* — Запустить компьютер — 最近我的ADSL Modem经常出现问题，有时**开机**能够正常工作 http://myhard.yesky.com/ServerIndex/771319046750863 36/20041222/1891733.shtml В последнее время у меня проблемы с ADSL, иногда невозможно нормально работать после **включения**.

62 — **关机** — *закрыть машину (выключить)* — Погасить компьютер — 消防人员提醒市民，电脑**关机**后，还应切断电源。http://www.zhoushan.cn/xwzx/zsxw/t20041219_154568.htm После **выключения** компьютера, надо отключить электричество.

63 — **集成的板子** — *интегрированная плата* — Интегрированная материнская плата — 我想要用英特的CPU和**集成的板子**配电脑! http://bbs.cpcw.com/archiver/?fid-94-page-67.html Я хочу компьютер с процессором Intel **и интегрированной платой.**

64 — **潜水员** — *водолаз* — Пользователь скрытый в чате — 本来我是**潜水员**来的，但是又想为大家伙做点贡献，所以我就浮出

水面回复一下 http://sh.focus.cn/msgview/10336/20021068.html Ведь я **водолаз**, но тоже хочу сделать вклад для всех, поэтому я всплыву обратно.

65 — **漫游** — *скитание (роуминг)* — Использование телефона вне зоны обслуживания — DoCoMo将提供具有国际**漫游**功能新款3G手机 http://www.pconline.com.cn/news/comm/0412/516582.html DoCoMo будет выставлять 3G телефон с международным **роумингом.**

66 — **奔腾** — *кипение (пентиум)* — Процессор фирмы Intel — IBM**奔腾**M Dothan机型也跌破万元了！ http://digi.163.com/2004w12/12755/2004w12_1102052315332.html IBM Компьютер **Пентиум** M Dothan уже стоит поменьше 10000 юаней.

67 — **速龙** — *быстроходный дракон (атлон)* — Процессор фирмы AMD — **速龙**64处理器获得了媒体的一致好评。 http://game.dayoo.com/content.php?id=26310 Процессор **Атлон** 64 получит всеобщее признание.

68 — **毒龙** — *ядовитый дракон (дюрон)* — Процессор фирмы AMD — 在市场上“**毒龙**”（Duron）处理器比赛扬处理器还便宜。 http://www.blogchina.com/new/display/59418.html На рынке процессор **Дюрон** дешевле процессора Селерон**.**

69 — **赛扬** — *seleron (селерон)* — Процессор фирмы Intel — Prescott核心的P4和**赛扬**D都在整个市场中取得了不错的成绩 http://it.com.cn/f/market/0412/10/57405.htm Процессор P4 Prescott и процессор **Селерон** D имеют очень хороший успех на рынке продаж.

70 — **苹果机** — *яблочная машина (apple)* — Компьютер фирмы Apple — 对于电脑图象爱好者来说拥有一台**苹果机**当然是一个梦想 http://bbs.zol.com.cn/new/TopicView_new2.php?nId=22643&board_id=115 Для компьютерных графических болельщиков возможность получить **Apple** — это мечта.

71 — **风冷** — *холод ветра (кулер)* — Вентилятор — 当时处理器热量比较低，一般铝散热片＋**风冷**就可以满足需要 http://gg.beareyes.com.cn/2/lib/200412/15/20041215067.htm Раньше процессору не было так жарко, поэтому обычного алюминиевого киля + **холода ветра** [радиатора + вентилятора] уже хватает.

72 — **水冷** — *холод воды (водяное охлаждение)* — Водяное охлаждение процессора — **水冷**系统是那些狂热的电脑爱好者们才会用到的东西 http://www.pcpop.com/power/04/12/55928/1.shtml Только супер болельщики используют **водяное охлаждение**.

73 — **蓝牙** — *синий зуб (bluetooth)* — Средство беспроводной связи — 大家都清楚**蓝牙**技术是比红外技术更高级的技术 http://info.home-a.hc360.com/HTML/001/001/004/225229.htm Все знают, что **Bluetooth** техника лучше инфракрасной техники.

74 — **输入法** — *метод ввода букв* — Метод ввода текста — 目前全国流行的**输入法**有10多种 http://www.szrbs.net/news_content.asp?NS_ID=26&NC_ID=348533 Сейчас есть 10 **видов ввода** китайского текста.

75 — **私服** — *собственная служба (сервер)* — Индивидуальный сервер — **私服**成为所有网络游戏运营商的噩梦 http://tech.szonline.net/Channel/content/2004/200412/20041216/355352.html **Индивидуальный сервер** уже стал кошмаром всех эксплуатационщиков сетевой игры.

76 — **液晶** — *жидкокристаллический* — ЖК-монитор — **液晶**价格战越来越激烈，各大厂商奇招叠出。 http://www.hbbear.com/2/lib/200412/22/20041222031.htm Ценовая война **ЖК-мониторов** уже очень ожесточена, каждые фирмы используют много методов.

77 — **纯平** — *плитка (crt-плоский монитор)* — CRT-плоский монитор — 大部分的DIY装机商看到的是17寸**纯平**仍是市场的主流 http://www.cbinews.com/inc/showcontent.jsp?articleid=12866 Большинство DIY коммерсантов думают о 17’**CRT-плоских мониторах**, которые наиболее продаются на рынке.

78 — **球面** — *сфера (выпуклый монитор)* — Выпуклый монитор — 虽说15寸的**球面**显示器已经落伍，但是这样的价格又能强求什么呢！ http://info.office.hc360.com/HTML/001/001/003/8597.htm Пусть 15’**выпуклый монитор** уже отставание, но только такая цена, тоже не плохой выбор!

79 — **联机** — *соединение машин* — Соединение компьютера — 有些爸爸把家里的电脑与公司电脑**联机**，然后就可以呆在家里上班 http://life.beelink.com.cn/20041221/1750261.shtml Некоторые **соединяют домашние компьютеры** с рабочими, потом работают, сидя дома.

80 — **冷启** — *холодное отпирание* — Ctrl+alt+del — 死机以后热启无效，必须从主机**冷启** http://bbs.zol.com.cn/new/TopicView_new2.php?nId=4194&board_id=169 После того как компьютер зависнет, **холодное отпирание** неэффективно, нужно горячее отпирание.

81 — **热启** — *горячее отпирание* — Reset — 机子键盘鼠标**热启**没反应 http://www2.beareyes.com.cn/bbs1/1/2004/01/08_1789278.htm После **горячего отпирания** компьютера клавиатура и мышь реагируют.

82 — **高手** — *гроссмейстер (специалист по компьютерам)* — Специалист по компьютерам — 成为计算机**高手**的道路是很曲折的。 http://www.gaoshou.net/gotop/index.php Путь, идя по которому можно стать **специалистом по компьютерам,** очень труден.

83 — **兼容机** — *сборная машина* — Собранный по желаемой комплектации компьютер — 相对**兼容机**来说，品牌机的稳定性更好一些. http://bbs.zol.com.cn/new/TopicView_new2.php?nId=22311&board_id=115 Относительно **сборной машины,** фирменная машина имеет лучшую устойчивость.

84 — **品牌机** — *фирменная машина* — Фирменный компьютер — 品牌机的核心优势是其售后服务、稳定性和童叟无欺的统一售价等 http://e.chinabyte.com/ServerIndex/7714176157456793 6/20041221/1891432.shtml Центральная доминанта **фирменной машины** — это устойчивость, обслуживание и единая цена.

85 — **激活** — *активация* — Активация софта — 如果你在规定时期内没有**激活**Windows XP，你就不能继续使用它。 http://www.microsoft.com/china/windowsxp/activation.asp Если вы не зарегистрировали **активацию** софта вовремя, вы не сможете продолжать работать в Windows XP.

86 — **千年虫** — *тысячелетний червь* — Компьютер, прошедший проблему 2000 года — 进入新年之后都没有发生异常现象，各部门的电脑系统也没有受到“**千年虫**”的干扰，安全进入2000年。 http://www.zaobao.com/zaobao/special/millennium/pages/bug301299g.html После наступления 2000 нового года, компьютерные системы не пострадали от “**Тысячелетнего червя**”, все преодолели проблему.

87 — **跳线** — *джампер* — Джампер — **跳线**可用于选择对硬盘的物理检测顺序。 http://support.wdc.com/cn/techinfo/general/jumpers.asp **Джампер** используют для выбора режима работы жесткого диск.

88 — **冲击波** — *логическая бомба* — “Волна” — 近日，"**冲击波**"病毒在网上流行传播，给大家造成了很大危害。 http://www.sne.snnu.edu.cn/techsupport/tech_news/blaster.htm Недавно, вирус **“Волна”** распространился в сети, и создал для всех угрозу.

89 — **网关** — *сетевая застава (шлюз)* — Шлюз — **网关**(Gateway)，将两个使用不同协议的网络段连接在一起的设备 http://tech.

sina.com.cn/other/2004-07-09/1757385795.shtml **Шлюз (Gateway),** средство связи сетей.

90 — **DIY机 —** *diy машина* — DIY компьютер — 大部分的**DIY装机**商看到的是17寸纯平仍是市场的主流 http://www.cbinews.com/inc/showcontent.jsp?articleid=12866 Большинство **DIY коммерсантов** думают о 17'CRT-плоских мониторах, которые наиболее продаются на рынке.

91 — **指针 —** *указатель* — Указатель — 函数**指针**是C＋＋最大的优点之一。 http://purec.binghua.com/Article/Class1/Class2/200411/361.html Функция **указателя** — одно из достоинств C++.

92 — **催化剂 —** *катализатор (драйвер)* — Программа для поддержания работы устройства — DirectX V9.0b，游戏的**催化剂**. http://article.pchome.net/2003/12/10/15064.htm DirectX V9.0b игровой **драйвер.**

93 — **保超 —** *гарантийное увеличение частот* — Гарантийное увеличение частот — 今天在市场中发现一批**保超**3.4G的赛扬D。 http://tech.sina.com.cn/h/2004-12-03/1644470044.shtml Сегодня на рынках появились процессоры Селерон D с **гарантийным увеличением частоты** разгона до 3,4G.

95 — **金手指 —** *золотой палец* — Разъём на плате — 我的显卡坏了，可能时**金手指**有问题。 http://www.pcshow.net/article/Articleinfo.jsp?id=212231 У меня не работает видеокарта, неверно сломался **золотой палец.**

97 — **碎片 —** *обломок* — Не используемые кластеры — 经常安装卸载程序，这样很容易在磁盘上会产生**碎片** http://article.pchome.net/2004/05/24/20130.htm Если постоянно устанавливать и удалять программы, то на жестком диске будет много **обломков.**

99 — **最大化 —** *развернуть* — Развернуть окно во весь экран — 让你的IE浏览器自动**最大化** http://www.skycn.com/article/751.html **Развернуть** полностью раскрытое окно IE.

100 — **最小化 —** *свернуть* — Свернуть окно — Outlook Express 会以**最小化**的方式运行。 http://www.glisp.com.cn/feel/pcworld/3113e.html Outlook Express может работать в **свернутом** окне.

101 — **浏览器 —** *прибор просмотра (браузер)* — Браузер — 他们在微软IE**浏览器**中发现了一个普遍性的安全漏洞。 http://content.sina.com/news/63/96/7639612_1_gb.html Они обнаружили слабое место в **браузере** Microsoft Internet Explorer.

102 — **区码** — *региональный код* — Региональный код DVD（всего 6 кодов） — 笔记本电脑的DVD光驱都是有**区码**保护的 http://tech.tom.com/1624/1672/2004129-144563.html DVD-устройства на ноутбуках защищены с помощью **региональных кодов.**

103 — **刷新BIOS** — *обновление bios* — Обновление BIOS — **刷新BIOS**一直是“菜鸟”头痛的事 http://www.softhouse.com.cn/html/200411/20041129131025000024l0.html **Обновление BIOS** — всегда головная боль для “Огородных птиц”.

104 — **优化系统** — *обновление улучшения системы* — Обновление улучшения системы — **优化系统**提高上网速度. http://www.pcworld.cn/99/9947/4735a.asp **Обновление системы** увеличивает скорость работы в Интернете.

105 — **红客** — *красный гость* — Красный хакер — “**红客**”是一种精神，技术是他的能力。 http://www.reder.cn/ **Красный гость** — это дух, техника — его возможность.

106 — **拷贝** — *копия* — Копирование файлов — 快速大批量**拷贝**文件我就要用RichCopy. http://tech.qianlong.com/28/2004/08/13/71@2215942.htm Если нужно **копировать** множество файлов, то я использую программу RichCopy.

107 — **桌面** — *столешница (рабочий стол)* — Рабочий стол — 很多人都喜欢把自己的**桌面**背景换成自己的图片 http://www.enet.com.cn/eschool/inforcenter/A20041220373621.html Многие люди любят изменять свое фото на своём **рабочем столе**.

108 — **水货** — *водный товар (контрафакт)* — Контрафактные компьютеры — 在经济繁荣的同时，非法走私的**水货**市场也随之兴旺起来 http://it.sohu.com/20041210/n223437193.shtml Одновременно с ростом экономика, **водный товар** тоже поступает на рынок.

109 — **行货** — *отраслевой товар (лицензионный товар)* — Лицензионные компьютеры — **行货**与水货之间的差价利润，成为JS以水货充**行货**销售的驱动力。 http://tech.sina.com.cn/digi/2004-12-18/1431479150.shtml Разница цен между контрафактными и **лицензионными** компьютерами велика, и JS продают контрафактные компьютеры.

110 — **爱虫** — *любовный червь* — Название вирус — 5月14日，名为“I LOVE YOU”（**爱虫**）的病毒在全球范围内发作，仅用三天的时间就造成全世界近4500万台电脑感染。 http://www.blogchina.com/new/display/59418.html На 14.05 вирус “I LOVE YOU” (**Любовный**

червь) сыграл в мире, только за 3 дня в мире были заражены вирусом 450 миллионов компьютеров.

112 — **震荡波** — *сотрясательная волна* — Название вируса — 2004年的“毒王”宝座最终被“**震荡波**”病毒夺得。 http://www.yunnan.cn/1898/2004/12/15/230@180708.htm Вирус “**Сотрясательная волна**” признан главным вирусом 2004 года.

113 — **笔记本** — *тетрадь для записи (ноутбук)* — Переносной компьютер — 如今**笔记本**在设计上已经越来越轻薄。 http://info.secu.hc360.com/HTML/001/001/62182.htm В последнее время **ноутбуки** всё легче и тоньше.

114 — **本本** — *тетрадь (ноутбук)* — Переносной компьютер — SONY推出的全球首款最薄的**本本**，它的性能也非常的棒！ http://digi.163.com/2004w12/12773/2004w12_1103614961131.html SONY выдвигает самый тонкий **ноутбук** в мире, и его возможности тоже очень сильны.

115 — **纯水** — *чистая вода (спам)* — Пустая реклама, объявления — 严重制止**纯**水泛滥成灾！ http://www.xf18.com/cgi-bin/bbs/topic.cgi?forum=4&topic=3&show=0 Внимание, нельзя писать **СПАМ** на сайте!

118 — **论坛** — *трибуна (форум)* — Место в Интернете, где обмениваются пониманием — 我对**论坛**有足够的热情，虽然我的文采不是很好,但我有实干的精神 http://club.china.alibaba.com/club/post/view/40_2844565.html Я люблю **форум**, пусть я не могу писать красивый текст, но у меня есть аккуратный дух.

119 — **火线** — *огненная линия* — Средство компьютерной связи — 1394**火线**PCMCIA接口卡为您的笔记本电脑带来革命性的**火线**传输速率(400Mbps/秒) http://www.mobile-touch.com/catalog/product_info.php?manufacturers_id=33&products_id=303 1394 **Огненная линия** PCMCIA даст вашему ноутбуку супер скорость (400Mbps/сек).

120 — **水晶头** — *хрустальная голова* — Разъём, коннектор — 上不了网的原因应该是网线的**水晶头**有问题 http://publish.it168.com/2004/0906/20040906008401.shtml Невозможно подключиться к сети потому, что возникли проблемы с **хрустальной головой.**

121 — **南桥** — *южный мост* — Чипсет —

1.搭载SiS965L南桥的主 板六月上市 作者: — 文志磊 时间: — 2004-05-14 出处: — 极Myhard Материнская плата, которая включается в **южный мост** SiS965L начнет продаваться в начале июня.
2.据消息来源指出，Intel将在今年第二季发布支持USB2.0规范的新款ICH4南桥芯片. (2002-01-03· ·文志磊··天极网硬件频道) По источнику информации, Intel обнародует новую модель ICH4 **южного моста**, который поддерживает порт USB2.0.

123 — **主板** — *главная плата* — Материнская плата — 华硕电脑关于P4P800主板支持PAT的声明. http://tech.tom.com 2003年06月12日来源:厂商提供 Компания АСУС опубликовала что, **главные платы** P4P800 поддерживает PAT.

124 — **镭** — *радий* — Тип видеокарты — ATI显卡: ATI原厂镭9700PRO显卡只卖1150元 www.thethirdmedia.com/pc/200411/20041112122338.shtm - 65k ATI видеокарта **радий** 9700PRO продается только за 1150 юаней.

125 — **硬盘** — *твёрдая тарелка* — Винчестер — 硬盘故障全接触 www.enet.com.cn/eschool/ inforcenter/T20040914343919.html Встречаются все типы аварий ***твердой тарелки***.

126 — **软盘** — *мягкая тарелка* — Дискета — 强力的软盘工具 dev.19xz.com/soft/11449.htm Сильный инструмент для ***мягкой тарелки***.

127 — **总线** — *главный кабель* — шина — 总线的历程与未来之星PCI Express (网易-科技报道) История ***главного кабеля*** и будущей звезды PCI Express.

128 — **5 类线** — *пяти категорийный кабель* — Пятижильный кабель — Cable modem与5类线的比较 http://www.catvshow.com/tech_center/cable_modem/3.htm） Сравните кабельный модем и ***пяти категорийный кабель***

129 — **锅** — *котелок* — Спутниковая антенна — 他告诉记者，前几天刚上了趟县城，买了辆新摩托车、还买了个**卫星锅**（新华网-青藏铁路民工赵世林的除夕夜）。 Он, сказал наш корреспондент, недавно приезжал в уезд, купил новый мотоцикл и спутниковый **котелок**.

130 — **龙心** — *драконовый C P U* — Китайский C P U — **听到的广告语不再是Intel Inside，取而代之的是“龙心inside”……（中青在线- 2024年生活状态大猜想）** Слушаем текст рекламы не Intel inside, a ***драконовый CPU*** inside.

131 — **拨号** — *выпилить номер* — Набрать номер. Набор номера — Windows XP 是微软推出的最新视窗操作系统，功能更强大，集成了PPPoE协议支持，ADSL用户不需要安装任何其他PPPoE拨号软件，直接使用Windows XP 的连接向导就可以建立自己的ADSL虚拟拨号连接。作者：Jackygz－－Windows XP 下设置ADSL拨号连接 WindowsXP является новой ОС компании Microsoft, большую функциональность включает PPPOE протокол. Пользователю не надо запускать никакой программы ***выпиливания номера***, прямо использовать проводник для соединении с Windows XP и можно составить свое индивидуальное наборное номерное соединение с ADSL.

132 — **坏点** — *пропавшая точка* — Дефект — ADI的这家经销商还承诺，这款ADi液晶显示器绝无坏点。http://www.sina.com.cn 2004年12月08日 10:04 PCPOP-电脑时尚 Представительство ADI акцептовали что, у этой модели ЖК монитора вообще нет ***пропавших точек***.

134 — **噪点** — *точка шума* — Точка шума — 富士长焦数码相机S5000自从推出之后就收到用户的关注，但是因为像素较低，画面噪点较多，用户反应并不是很好，因此，富士推出了S5000的升级机型S5500。http://www.sina.com.cn 2004年12月08日 13:59 IT168.com Цифровая фотокамера fujifilm-S5000 получила много оценок покупателей, но у неё мало пиксель (pixel), много ***точек шума***, и мало продаж. Поэтому fujifilm выдвинул новую модель S5500.

139 — **软体** — *мягкое тело* — Софт, ПО — 这款名为Balance的笔记型电脑, 软体部分则使用以Linux为基础的Linspire作业系统 大纪元时报-澳洲版 Balance ноутбук, использует ***софт*** linspire ОС, который является основой linux.

145 — **病毒** — *вирус* — Компьютерный вирус — 有人觉得，病毒是一段程序，而数据文件如.txt、.pcx等格式文件一般不会包含程序，因此不会感染病毒。殊不知像word、excel等数据文件由于包含了可执行码却会被病毒感染，而且，有些病毒可以让硬盘里面的文件全部格式化掉，因此，我们不能忽视数据文件的备份。 http://it.nen.com.cn/79124163859578880/20041221/1574575.shtml Некоторые люди считают, что ***вирус*** представляет собой часть программы, но не файлов данных, например: txt, pcx и т.д. и обычно не включают программы, поэтому не могут подцепить вирус. Но они не знают что, word, excel и другие файлы данных включают макропрограммы, поэтому тоже могут подцепить вирус, а некоторые вирусы могут формировать все

файлы на винчестере. Поэтому нам нельзя игнорировать хранение резервной копии файлов.

146 — **木马** — *кобыла* — Вирус Троян — 2004年木马占全年病毒总数的一半. (http://www.enet.com.cn/eschool/inforcenter/A20041217373069.html) Вирус ***кобыла*** занимал 50% от общей суммы.

147 — **冰河** — *глетчер* — Вирус Глетчер — 例如冰河使用的监听端口是7626 (http://www.ah163.net/cw/show.php?id=21058) Например, вирус ***Глетчер*** использует порт контроля 7626.

148 — **病毒库** — *вирусная база, вирусный склад* — Обновление антивирусов — 金山毒霸每工作日病毒更新，请用户随时做好病毒库的升级工作，确保您系统的安全性。**金山反病毒20041227_周报** Программа антивируса Jinshan duba освежает свою ***вирусную базу*** каждый день. Пользователи, пожалуйста, увеличивайте свои ***вирусные базы*** вовремя, чтобы обеспечить безопасность ОС.

149 — **蠕虫** — *червь* — Вирус Червь — 蠕虫病毒有很多种类. (中国新闻网 - 2004年11月30日) У вируса ***Червя*** много разных типов.

152 — **漏洞** — *дыра* — Дыра — IE又现高危漏洞. 2004年12月22日 09:12 来源：新浪科技 IE опять появились очень опасная ***дыра***.

153 — **汉化** — *китаизация* — Китаизация (тоже, что и русификация) — 这个版本已经可以称得上是完美了。能汉化的地方全部都汉化了。(http://it.com.cn/f/games/0412/20/60530.htm) Эту версию уже можно назвать замечательной. Сделана почти полная ***китаизация***.

154 — **抓图** — *браться чертеж* — Принтскрин — 之所以上面这些图片的大小不一样，这完全是**抓图**软件的问题。 http://www.sport.org.cn/ceg/ceping/2004-11-22/398082.html Причина не одинакового размера всех этих картинок, в полной проблем программе ***взятия чертежа***.

155 — **分区** — *подблок* — Логический диск — 如果已在各主**分区**中独立安装了多套操作系统，可用这个命令来切换到不同的操作系统下。http://content.sina.com/news/64/43/7644327_1_gb.html Если уже установил ОС в каждом ***подблоке***, то можно использовать эту команду для переключения на разные ОС.

156 — **后缀** — *суффикс* — Расширение — "CN"**后缀**网址浏览速度将提高30% 东方网 - 2004年12月21日 Вэб-страница, которая имеет ***суффикс*** с CN, будет повышать скорость на 30%.

157 — **菜单** — *меню* — меню — 我的Windows XP听我的 —定制个性化的XP菜单 (Tom) Мой Windows XP позволяет мне сделать стильное ***меню*** XP.

158 — **磁道** — *магнитная дорога* — Магнитная дорожка — 硬盘是一种磁介质的外部存储设备，在其盘片的每一面上，以转动轴为轴心、以一定的磁密度为间隔的若干同心圆就被划分成磁道(Track)，每个磁道又被划分为若干个扇区(Sector)，数据就按扇区存放在硬盘上。 http://news.onlinedown.net/info/16730-1.htm Винчестер представляет собой магнитное материальное устройство для хранения данных. Каждый диск винчестера имеет по центру вал и разделенные по магнитной плотности несколько концентрических окружностей — ***магнитных дорожек***. Каждая магнитная дорожка разделяется на несколько секторов, все данные сохраняются в секторах.

159 — **扇区** — *веер секторов* — Сектор — 硬盘是一种磁介质的外部存储设备，在其盘片的每一面上，以转动轴为轴心、以一定的磁密度为间隔的若干同心圆就被划分成磁道(Track)，每个磁道又被划分为若干个扇区(Sector)，数据就按扇区存放在硬盘上. http://news.onlinedown.net/info/16730-1.htm Винчестер представляет собой магнитное материальное устройство для хранения данные. Каждый диск винчестера имеет по центру вал и разделенные по магнитной плотности несколько концентрических окружностей — магнитных дорожек. Каждая магнитная дорожка разделяется на ***веер секторов***, все данные сохраняются в секторах.

160 — **磁盘** — *магнитная тарелка* — Дискета — Windows操作系统中如何修复磁盘坏道 中安网 - 2004年12月1日 В ОС Windows как восстановить плохую дорожку на ***дискете***?

161 — **继承** — *наследование* — Наследование — 所谓的面对对象编程，主要是指语言的四个特征：抽象、封装、**继承**性。 世界经理人 - 2004年12月17日 Объектно-ориентированное проектирование указывает на 3 главных свойства языка: абстракция, затаривание и ***наследование***

163 — **指针** — *указатель* — Указатель — C++中，成员**指针**是最为复杂的语法结构。 (eNet硅谷动力 - 2004年12月20日) В языке C++, ***указатель*** имеет самую сложную грамматику.

164 — **短信** — *короткое письмо* — СМС — 去年全国手机短信发送量近20亿． 每日甘肃 В прошлом году во всем Китае количество отправленных ***СМС*** составило почти 200 миллионов.

Литература

1.Филиппович Ю.Н. Метафоры информационных технологий. С предисловием Караулова Ю.Н. – М.: Московский государственный университет печати; 2002г. – 283с. ISBN 5-8122-0279-6.

2.<<微型计算机>>2003年合订本．－北京：人民交通出版社；2004年1月－ 454页．ISBN 7-114-04890-4．<<МикроКомпьютер>> 2003года – экземпляр блоки. – Пекин: Издательство «Народная коммуникация»; 2004г. Яньварь – 454с. ISBN 7-114-04890-4.

3. 现代俄汉双解词典．－外语教学与研究出版社，997年1304页．ISBN 7-5600-0810-0/H.366. Новый русско-китайский словарь. – Издательство «Иностранные языки»; 1997г – 1304с. ISBN 7-5600-0810-0/H.366.

4. 汉俄词典．－商务印书馆出版，1992年1250页．ISBN 7-100-00312-1/H-106. Китайско-русский словарь. – Издательство «Коммерческая печатня»; 1992г.1250с. ISBN 7-100-00312-1/H-106.

5. <<电脑爱好者>>2004年半月刊第13期7月1日出版 «Любитель комньютера» 2004г. пресса дихотомии, 13ая периодика, издание первого Июля

6. <<电脑爱好者－精品文库>>2004年．内蒙古科学技术出版社；2004年5月－ 387页. ISBN 7-5380-1139-0. «Любитель комньютера – эффектные статьи» 2004г. Внутренное монгольское научное техническое издательство; Мая, 2004г. – 387с. ISBN 7-5380-1139-0.

И.С.Кесель

ИНТЕРАКТИВНОЕ УЧЕБНО-МЕТОДИЧЕСКОЕ ПОСОБИЕ ПО ДИСЦИПЛИНЕ «ЗАЩИТА ИНФОРМАЦИИ В АСОИУ»*

Введение

В основе моей работы лежит курс лекций, который читается на кафедре «Системы обработки информации и управления" – МГТУ имени Н.Э Баумана: «Защита информации в АСОИУ». Курс состоит из 10 тем, каждая из которых разделена на несколько подразделов. Основная задача, которую я поставил перед собой при создании проекта, – помощь обучающемуся в освоении учебного курса. Под помощью я понимаю всевозможные функции, направленные на то, чтобы человеку было удобно, приятно и легко изучать данный материал.

Традиционное классическое обучение, практикуемое во всем мире, стало все менее отвечать современным потребностям общества в условиях всеобщей мобильности и глобализации. Если ранее достаточно было один раз получить хорошее образование и в дальнейшем использовать его практически всю жизнь, то в наше время во многих отраслях обновление знаний должно происходить не реже чем раз в пять лет, а в ряде случаев и ежегодно. Но сам процесс обучения в этом случае должен осуществляться без отрыва от работы, при гибком графике и по индивидуальной программе, соответствующей компетенциям конкретного специалиста [1].

Частично эти задачи могут быть решены в рамках заочного или дистанционного обучения, получивших распространение в XX веке. Сразу после их появления практически всеми экспертами была отмечена потеря в качестве (обусловленная пониженным восприятием), что делало его малоэффективным и не воспринималось как достойная замена очному обучению. Но при всех его недостатках оно было доступно для всех желающих. Основным носителем знания при такой форме обучения

---

* В статье представлены материалы конкурсной работы IX научной конференции молодых исследователей «Шаг в будущее, Москва». Консультант — Анна Филиппович

являлась книга (учебник, методическое пособие и т. д.) и редкие встречи с преподавателем, в основном при сдаче контрольных мероприятий. Таким образом, практически отсутствовал диалог «преподаватель — обучаемый», постоянный мониторинг процесса обучения и другие элементы, присущие качественному (хорошему) процессу обучения. Кроме того, данная форма была малоприменима для задач корпоративного обучения, потребность в котором испытывают сейчас почти все организации.[1]

Электронное обучение (E-learning) уже сегодня включает в себя достоинства обеих форм обучения. С одной стороны, предлагая унифицированную услугу вне зависимости от места и времени обучения, с другой – включая интерактивные формы взаимодействия слушателя и преподавателя, а также современный контроль обучения. Поэтому я выбрал именно этот способ для реализации своей задачи.

Чем электронное обучение отличается от других видов? Само то, что оно сделано на компьютере, еще не означает, что оно относится к *E-learning.* Ведь учитель может дать вам электронное пособие вместо обычного учебника, но при этом он будет продолжать заниматься с вами, таким образом, это просто будет еще одним видом очного обучения. Отсюда мы можем сделать определение:

E-learning — полностью автономный курс обучения, не зависящий от человека как проверяющего учителя.

Пути создания программы

При создании электронного курса первый же вопрос, который возникает, – это *как* его сделать и *какие инструменты* при этом использовать. Для себя я выделил три возможных пути:

1. Использовать уже готовые специализированные программы для создания методических пособий.
2. Сделать программу, основывающуюся только на Web- содержимом.
3. Применить известные мне языки программирования.

Готовые программы. AuthorWave. В наше время существует множество программ, сделанных специально для создания курсов интерактивного обучения. Все они имеют схожий интерфейс и практически одинаковые наборы возможностей, функций и свойств. Рассмотрим наиболее популярный пример. Известная фирма Macromedia несколько лет назад выпустила в свет программу под названием *"AuthorWave"*(рис 1.1).

Первоначально спроектированный для создания курсов компьютерной подготовки на дисках, *«AuthorWave»* впоследствии был адаптиро-

ван для создания модулей, которые с помощью собственных плагинов могут проигрываться в Web-браузерах. В программе существует многофункциональный язык, позволяющий автору создавать высоко интерактивные ветвящиеся схемы курсов, но из этого следует, что, чтобы воспользоваться всеми преимуществами, предлагаемыми *«AuthorWave»,* необходима специальная подготовка и практика. Также в программе нет встроенной модели проектирования учебного материала, однако имеется ряд предварительно созданных объектов для различных частей курса, таких как экран входа в систему или вспомогательное меню.[2]

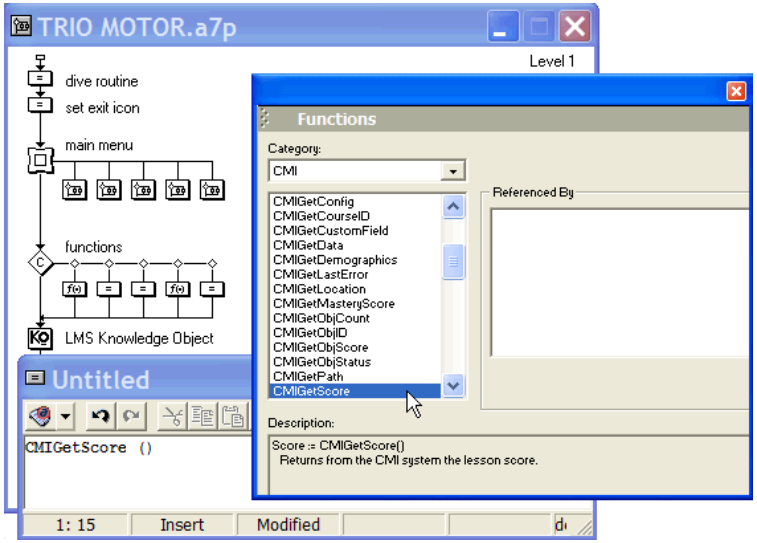


Рисунок 1. Интерфейс программы “AuthorWave”.

Таким образом, изучив *«AuthorWave»,* можно выделить плюсы и минусы данного пути:

Простота в освоении и работе	Каждая такая программа обладает индивидуальным языком, на изучение которого может уйти много времени.
Возможности	«AuthorWave» позволяет создать сложные ветвящиеся схемы курсов. Но при этом такие программы не предоставляют большого места для фантазии автора, то есть они очень шаблонны.

	«AuthorWave» также позволяет создавать различные системы тестов.
Работа с конечным продуктом	Подавляющее число таких инструментов написано на английском языке, что не совпадает с языком курса. Полноценное изучение курса, составленного таким образом, требует наличия дополнительных модулей и плагинов, помимо самого пособия.
Дополнительная информация	Средняя цена такого инструмента колеблется от 100 до 4000 $ («AuthorWave» стоит 2700$).

Использование только HTML- технологий. Несмотря на высокие образовательные цели, некоторые электронные курсы зачастую являются всего лишь специализированными Web-сайтами. Многие из них созданы с помощью авторских программ разработки Web-сайтов. Они представляют собой набор связанных между собой страниц. Этот метод позволяет легко и быстро редактировать страницы изучаемого материала и связывать их.

Однако очень сложно реализовать таким способом, например, простую систему тестирования и оценки знаний обучаемого. Пособие сможет работать только при наличии Web- Браузера на компьютере, и при этом все файлы курса не будут защищены от изменений. То есть пользователь будет иметь свободный доступ к ним.

Простота в освоении и работе	Такой путь не требует никаких дополнительных знаний, помимо простого владения HTML- технологиями.
Возможности	Удобное форматирование текста, но очень маленькие возможности, например, при программировании теста.
Работа с конечным продуктом	Вид конечного продукта не соответствует содержанию курса. Также есть ненадежность, связанная с непостоянностью при работе на разных компьютерах. Однако у программы очень простой и понятный интерфейс.
Дополнительная информация	Доступность (готовая программа не будет требовать установки дополнительных плагинов)

Языки программирования. Delphi 6.0. Язык программирования Delphi — один из самых актуальных языков программирования в мире. На Delphi реализуются как простые программы, так и сложнейшие проекты для ведущих корпораций мира. Delphi - это современная система программирования. Назначение Delphi - быстрая разработка приложений. Delphi построен на Object Pascal, который в свою очередь порожден из Паскаля [3].

Delphi построен на очень понятном и простом в освоении языке, позволяющем быстро воплотить большинство инженерных задумок относительно тестов и управления программой. Все исходные тексты программ хранятся в формате “*.pas” , они открыты для редактирования даже простым «Блокнотом», но, как только программа скомпилирована, она сразу же сохраняется в формате “*.exe” , и это означает, что ваш программный код будет полностью защищен от несанкционированных изменений.

Однако как бы ни был универсален этот язык, все равно обойтись только им не получается. Он очень удобен для создания форм и окон проекта, однако не получается работать с текстом курса. Для того чтобы отформатировать его при помощи Delphi, нужно вставить все 300 с лишним страниц внутрь программного кода, а это означает огромную потерю скорости при работе и загрузке. Следовательно, тут нельзя обойтись без дополнительной помощи текстовых или Web — редакторов.

Итог. Мой выбор.

Второй путь, путь построения программы на HTML – технологиях, не позволит мне создать приемлемую систему оценки знаний обучающегося, но зато с его помощью можно отформатировать текст, сделав его удобным для просмотра и изучения. Третий же путь является полной противоположностью: он не позволяет работать с текстом, зато появится возможность реализовать все идеи относительно режима теста и других вспомогательных функций.

Исходя из этого, я считаю наиболее правильным решением объединить эти два подхода. Таким образом, плюсы одного пути перекрывают минусы второго.

Реализация программы

Поставленные задачи

В самом начале, когда программа находилась еще в стадии задумки, я уже поставил перед собой основные задачи:

1.Удобство. Прежде всего, программа должна быть простой и удобной в обращении. Если человек пользуется ей в первый раз, то он не должен испытывать значительных затруднений при её освоении.

2.Повышение информативности при прочтении. То есть программа должна помогать обучаемому усваивать больше материала за счет экономии времени и сил.

3.Функциональность. Нужно не забывать, что я делаю не просто оболочку для текста, а интерактивное пособие для обучения. То есть помимо простого просмотра и изучения текста должны присутствовать и другие функции, такие как словарь, удобный режим просмотра различных схем и таблиц и т.д.

4.Возможность самоконтроля обучающегося. Этот пункт вытекает из второго: для повышения информативности прочитанного текста обучаемый должен иметь некую возможность проконтролировать себя, понять, что он прочитал и усвоил, а что нуждается в повторном прочтении.

Схема будущего проекта

Начало работы с программой.

Предварительное тестирование. Пользователь проходит тест, который показывает ему уровень его знаний и подсказывает нужные места в курсе для прочтения.

Изучение курса. Режимы, помогающие при изучении: *Список литературы, Формы для работы со схемами, Словарь.*

Второе тестирование. Далее пользователю предлагается второй тест для выявления разделов, которые он изучил плохо. И вновь, программа подскажет ему, что повторить.

Повторное, изучение курса.

Окончательное тестирование.

Завершение работы с программой.

Работа программы

Программа имеет четыре режима работы: «Тест», «Учебник», «Словарь», «Список Литературы». Подробный алгоритм их работы описан ниже.

Для того чтобы реализовать вариант обучения со вступительной промежуточной оценкой знаний, режим «Тест» был разделен на два варианта работы. На каждый из вопросов есть четыре варианта ответа, и только один из них правильный.

Внешний вид самой программы. Программа внешне напоминает студенческую тетрадку в клетку, слева расположены цветные закладки, служащие кнопками для переключения режимов работы.

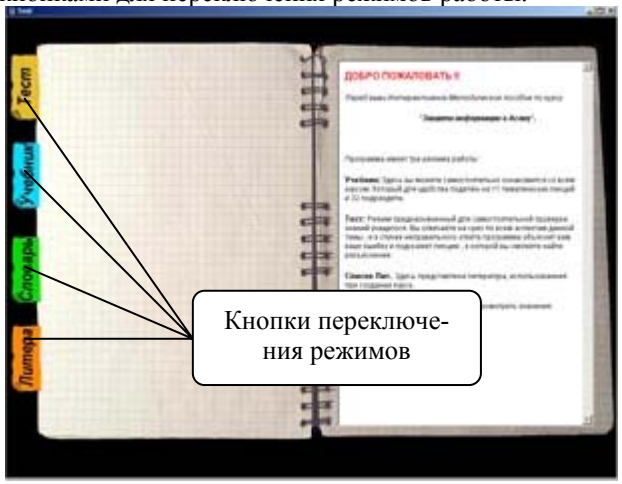


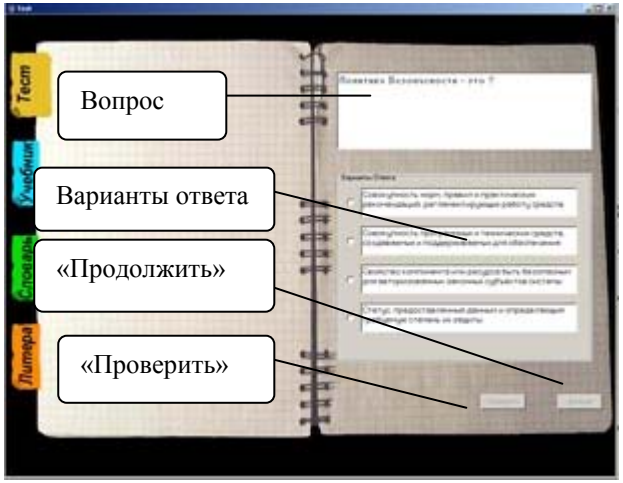
Рисунок 2. Интерфейс самой программы.

На правом и левом листе тетради расположены два рабочих поля. Во время изучения курса в этих полях может отображаться текст курса, список литературы или вопросы в зависимости от того, какой режим выбран.

Режим «Тест».

При работе в данном режиме вверху второго рабочего поля находится вопрос, ниже четыре варианта ответа. Как только один из них выбран, сразу же появляется кнопка проверить. Ее нажатие может привести к двум возможным вариантам: *Ответ правильный; Ответ был дан не верно.*

В первом случае программа сообщит об этом и предложит продолжить тест. Во втором случае на экране высвечивается правильный ответ, показывается содержащая его цитата из текста курса и, наконец, ссылка на это место в режиме учебника. Таким образом, человек может сразу же перейти к изучению (рис.3).



Режимы «Словарь» и «Список литературы».

При переходе в режим Списка литературы (рис 4) в обоих рабочих полях появляется текст, представляющий собой список изданий и их авторов, работы которых были использованы при создании курса.

Рисунок 4. Список литературы.

Внизу есть кнопки перехода назад и вперед, при нажатии которых меняется содержимое рабочих полей.

При работе с режимом «Словарь» алгоритм сохраняется, только вместо списка авторов в рабочие поля загружается список терминов, использованных в курсе.

Режим «Учебник».

При переходе в режим обучения (рис 5) в программе происходят следующие изменения:

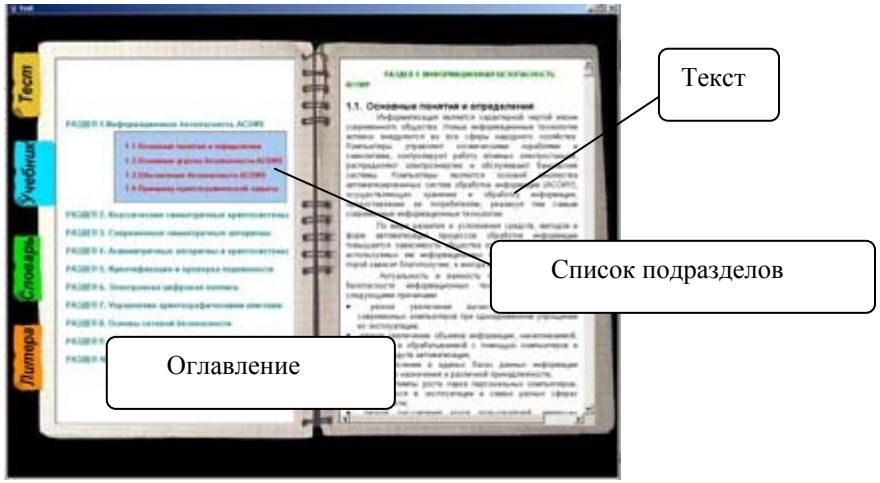


Рисунок 5. Учебник.

В первом рабочем поле появляется оглавление. При нажатии на любую тему её содержание тут же отразится во втором рабочем поле, а также появится маленькое диалоговое окно, в котором вы можете выбрать нужный подраздел. Если в тексте встретятся большие по своему размеру схемы, то при нажатии на них они появятся в отдельном окне. Это позволяет продолжить чтение дальше и не возвращаться назад, чтобы посмотреть рисунок.

Содержание электронного учебного пособия

Курс «Защита информации в АСОИУ».

Новые информационные технологии активно внедряются во все сферы деятельности человека. Компьютеры управляют космическими

кораблями и самолетами, контролируют работу атомных электростанций, распределяют электроэнергию и обслуживают банковские системы. Компьютеры являются основой множества автоматизированных систем обработки информации и управления (АСОИУ), осуществляющих хранение и обработку информации, предоставление ее потребителям, реализуя тем самым современные информационные технологии.

Быстрое развитие этих технологий вносит заметные изменения в нашу жизнь. Все чаще понятие “информация" используется как обозначение специального товара (ресурса), который можно приобрести, продать, обменять. При этом стоимость информации часто превосходит в сотни и тысячи раз стоимость компьютерной системы, в которой она размещается. Поэтому вполне естественно возникает необходимость в защите информации от несанкционированного доступа, умышленного изменения, кражи, уничтожения и других преступных действий.

Проблемы защиты информации привлекают все большее внимание как специалистов в области компьютерных систем и сетей, так и многочисленных пользователей современных компьютерных средств.

Именно этой актуальной теме и посвящен курс «Защита информации в АСОИУ»**,** подробно рассматривающий основные термины и понятия, современные методы и средства защиты информации в компьютерных системах и сетях. В пособии излагаются классические методы шифрования и современные криптографические методы, алгоритмы, протоколы и системы. Описаны методы и средства защиты локальных и корпоративных сетей от удаленных атак через сеть Internet, а также рассмотрена защита информации в электронных платежных системах. Приводятся подробные данные об эффективных отечественных аппаратно-программных средствах криптографической защиты компьютерной информации. Удачным дополнением материала служит перечень вопросов для самопроверки и тестирования, удобные ссылки на соответствующие разделы и обширный список литературы.

Вопросы для тестовых заданий.

Для того чтобы проверить знания обучаемого по всем темам курса, вопросы были составлены для всех разделов. Вопросы основывались на главных темах каждого раздела, а не на частностях. Они строятся на самых важных понятиях, терминах, а так же методах и средствах используемых для защиты информации в АСОИУ.

Например, девятый раздел посвящен Электронным платежным системам (ЭПС), технологиям их использования и защиты. Понятие ЭПС является ключевым, поэтому знать его значение необходимо, чтобы

качественно освоить этот раздел и курс в целом. В результате, этот термин был включен в список вопросов по девятому разделу. Этот же принцип применялся при составлении вопросов и для остальных разделов. *(Таким образом, на данный момент составлен список из 20 вопросов).*

Содержание словаря

В самом курсе дается объяснение очень многим терминам и понятиям в области защиты информации, но в словарь вошли только самые важные слова по этой теме, без которых невозможно качественное освоение курса. Всего было выделено 111 терминов, которые можно разделить на два типа:

1)Слова, значение которых дается в самом тексте курса.

2)Важные слова по данной теме, но не объясненные в тексте.

Это сделано для того, чтобы изучать курс мог не только специалист, хорошо знающий данную тему, но и человек, далекий от неё.

Первый тип слов:

Субъект *– это активный компонент системы, который может стать причиной потока информации от объекта к субъекту или изменения состояния системы*.[4] (определение взято из подраздела *1.1 Основные понятия и определения*).

Второй тип слов:

Федеральная Служба Технического и экспортного контроля (ФСТЭК) *– служба, которая после реорганизации Гостехкомиссии при Президенте РФ в 2003 году взяла на себя ее полномочия.* [5].

В курсе значение ФСТЭК не раскрыто, но цели и задачи этой службы являются важными для понимания всего курса.

Заключение

Таким образом, программа представляет собой полностью интерактивное пособие, имеющее все необходимые функции для удобного и легкого изучения курса «Защита информации в АСОИУ».

Два варианта работы режима «Тест», позволяют оценивать знания обучаемого до, и после изучения. Режим «Учебник» сделан максимально просто и при этом максимально функционально. Расширенный и дополненный словарь тоже

Дальнейшее развитее программы предполагает:

1) Увеличение количества вопросов.

2) Дополнение словаря таким образом, чтобы увеличить количество слов, как первого, так и второго типа.

3) Добавление анимированной заставки.

4) Ввод в программу функции фоновой музыки. Такой, чтобы не отвлекать пользователя, но при этом создавать приятную атмосферу. Предполагается также ввести управление этой музыкой, то есть возможность сделать её громче/тише или вовсе выключить.

5) Ввод другой системы теста, не основанный на выборе варианта ответа.

Литература

1. *Хортон У., Хортон К.* Электронное обучение: инструменты и технологии. // Пер. с англ. – М.: КУДИЦ-ОБРАЗ, 2005.- 640с.

2. www.macromedia.com

3. www.borland.com

4. *Кесель С.А.* Конспект лекций по дисциплине «Защита информации в АСОИУ».

5.*Семенко В.А.* Информационная безопасность: Учебное пособие. 2-е изд., стереот. – М.: МГИУ, 2005. – 215 с.

А.Ю.Филиппович, А.Б.Кирнарский

ПРОВЕДЕНИЕ ИНТЕРАКТИВНЫХ ЛИНГВИСТИЧЕСКИХ АССОЦИАТИВНЫХ ЭКСПЕРИМЕНТОВ В ИНТЕРНЕТ

Лингвистический ассоциативный эксперимент (ЛАЭ) представляет собой опрос некоторого количества людей, как правило, объединенных некой общностью (знаний, сферы, деятельности, языка, места рождения и т.д.) на предмет выявления их ассоциаций (реакций) на определенные стимулы. Информация о респондентах, а также их ассоциации со стимулами вносятся в базу данных, позволяющую формировать статистику и, в дальнейшем, анализировать пары «стимул-реакция» (или обратное соотношение «реакция-стимул») в общем или в разрезе различных групп респондентов.

Цель проведения ЛАЭ – формирование широкой (т.е. содержащий большое количество пар стимул-реакция) и глубокой (т.е. включающей большое количество ассоциаций на каждый стимул) базы данных, на основе которой возможно:

- Понимание ассоциативного ряда усредненного респондента в современный период;
- Построение и анализ ассоциативных взаимосвязей между словами, поиск закономерностей;
- Формирование прямого и обратного ассоциативного словаря.

В дальнейшем результаты ассоциативных экспериментов могут иметь применение в самых разнообразных сферах деятельности человека – от реализации контекстного поиска в сети Интернет, до составления наиболее убедительных рекламных текстов.

Основные этапы классического лингвистического эксперимента это:

- подготовка анкет эксперимента;
- распечатка анкет;
- заполнение анкет группой респондентов;
- ввод анкет в БД;
- анализ полученных данных и формулировка выводов.

Необходимость вручную формировать, распечатывать и вносить в базу данных результаты экспериментов является существенным недостатком при проведении ЛАЭ, так как:

- требует серьезных временных, организаторских и материальных затрат;
- существует возможность ошибки (порча анкет, неразборчивость подчерка, опечатки при вводе);
- накладываются ограничения на удаленность респондентов.

Эти проблемы можно практически полностью устранить, если создать средства проведения интерактивных ЛАЭ в Интернет.

Проведение лингвистических ассоциативных экспериментов в Интернет

Технологии, используемые в сети Интернет, позволяют разработать удобный интерфейс для проведения интерактивные опросов с сохранением результатов в общей базе знаний.

Неоспоримыми преимуществами проведения ЛАЭ через Интернет являются:

- значительное уменьшение временных и материальных затрат;
- возможность прохождения анкетирования одновременно неограниченному числу респондентов;
- возможность одновременного проведения нескольких ЛАЭ;
- значительное уменьшение ошибок;
- увеличение круга респондентов – возможность постоянного развития базы.

Рассмотрим некоторые из этих преимуществ подробнее.

Значительное уменьшение временных и материальных затрат достигается тем, что при проведении эксперимента в Интернет нет необходимости вручную подготавливать анкеты, распечатывать их, организовывать централизованное распространение, заполнение и сбор анкет. Также исключается этап ввода анкет в базу данных специальными операторами – вместо этого, стимулы вводятся в базу данных самим респондентом.

Возможность прохождения анкетирования одновременно неограниченному числу респондентов является очень важным преимуществом – за счет этого можно многократно сократить сроки проведения эксперимента и увеличить его аудиторию.

Значительное уменьшение ошибок достигается за счет отсутствия необходимости распознавать подчерк и вводить данные анкет в базу данных. Кроме того, исключается порча и утеря анкет, неразборчивость подчерка и т.д. В дальнейшем возможно создание алгоритмов фильтрации «мусора» - т.е. единственных и бессмысленных реакций.

Реализация программы по проведению ЛАЭ в Интернет

Для реализации поставленной задачи и создания интерактивного Интернет сайта было принято решение разбить задачу на два этапа:

- Формирование базы данных, хранящей анкеты проведенных экспериментов
- Реализация машины вывода и средств взаимодействия пользователя с базой данных посредствам Интернет сайта.

На первом этапе, после ознакомления с предметной областью были выявлены следующие основные сущности и связи:

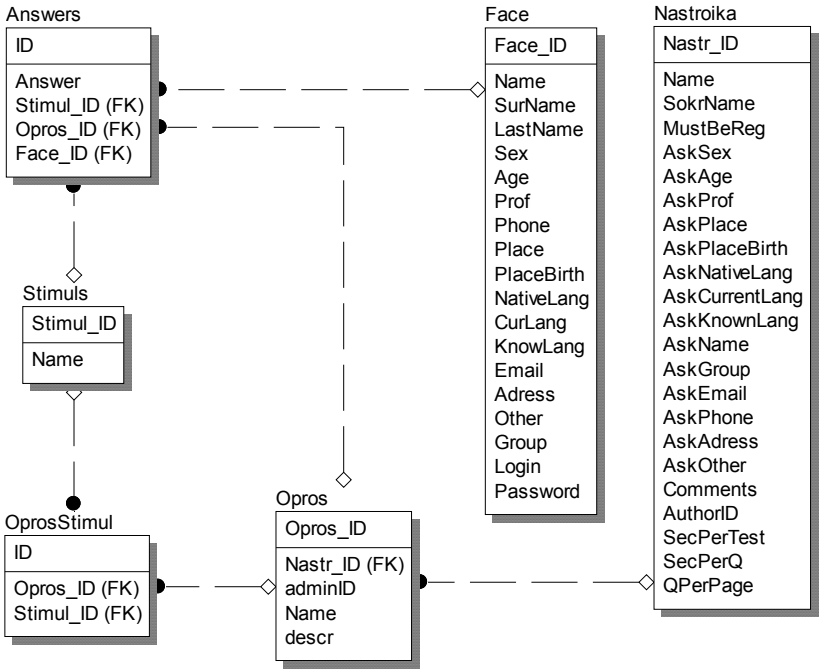


Рис. 1. Инфологическая модель БД ЛАЭ в первой версии программы

В существующей структуре не хватает очень важной компоненты - анкет ассоциативного эксперимента. В настоящей версии пользователю приходится отвечать на все стимулы анкеты, что является большим недостатком и критическим ограничением для проведения крупномасштабного эксперимента.

Для следующих версий программы была создана гораздо более гибкая и адекватная схема данных, которая показана на рисунке:

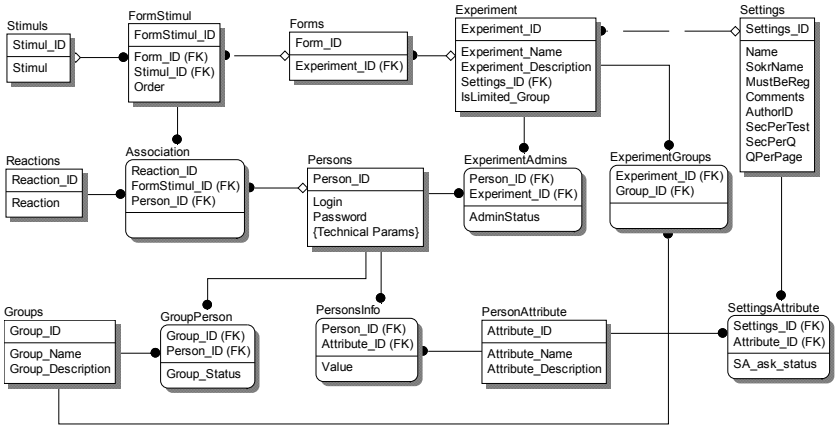


Рис. 2. Даталогическая модель БД ЛАЭ во второй версии программы

В настоящий момент происходит разработка новой версии программы, реализующей именно эту схему данных.

Рассмотрим эти схему данных подробнее. Объектами сущности «Стимул» (Stimuls) являются все стимулы в системе. Через различные соотношения (FormStimul, Forms, Association) стимул связывается с «анкетой» (Forms), «экспериментом» (Experiment) и «реакциями» (Reactions). Сущности, хранящие данные о респондентах, группах респондентов также связаны как с экспериментом, так и с конкретной ассоциацией, что позволяет сохранить связь между конкретным пользователем (группой пользователей) и его парами «стимул-реакция». Что касается настроек параметров эксперимента, в системе предусмотрены сущности «атрибут настройки» (SettingsAttribute и PersonAttibute) и «персональная информация» (PersonsInfo), объекты которых позволят создавать гибкую структуру запрашиваемых у респондента личных данных (напри-

мер, если это необходимо, администратор может запросить у респондента наименование ВУЗа, где тот обучался).

Второй этап заключается в реализации машины вывода и средств взаимодействия пользователя с базой данных посредствам Интернет сайта.

В качестве технологий для реализации Интернет сайта и базы данных были выбраны:

- СУБД Mysql – для управления базой данных;
- Язык программирования PHP – для реализации необходимого функционала
- Язык разметки HTML – для корректного отображения Интернет-сайта в Интернет-броузерах;

Первая версия Интернет-сайта разделена на два раздела:

- пользовательская часть – предназначена непосредственно для проведения эксперимента – ввода реакций на предлагаемые стимулы, а также сохранения личных данных респондента, вывода статистики;
- администраторская часть – предназначена для ввода данных об эксперименте и его настройки.

Общий вид пользовательской части представлен на рисунке:

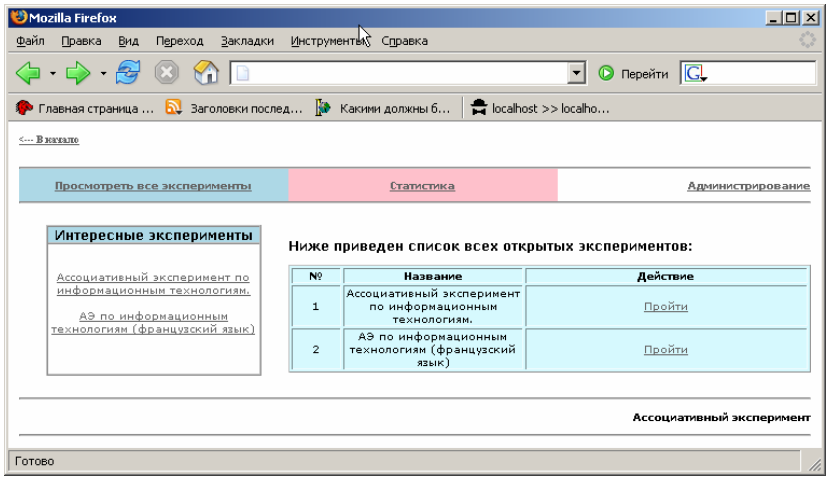


Рис. 3. Общий вид пользовательской части

При прохождении эксперимента, система запрашивает личные данные респондента, в соответствие с настройкой эксперимента:

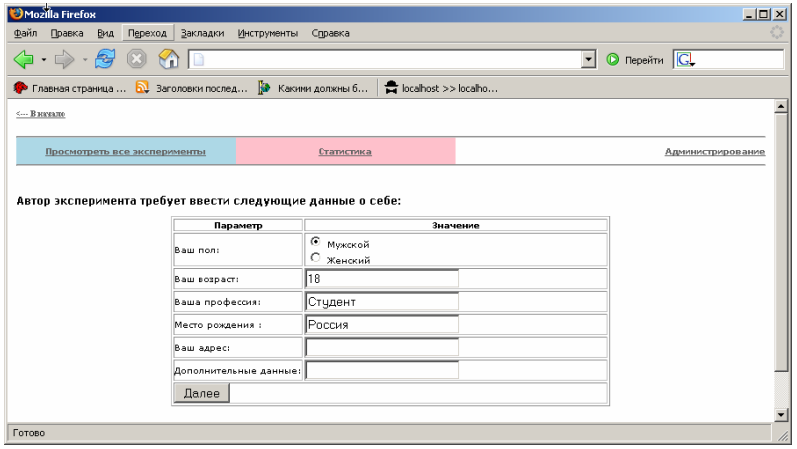


Рис. 4. Ввод личных данных о респонденте

Далее респондент переходит непосредственно к анкете:

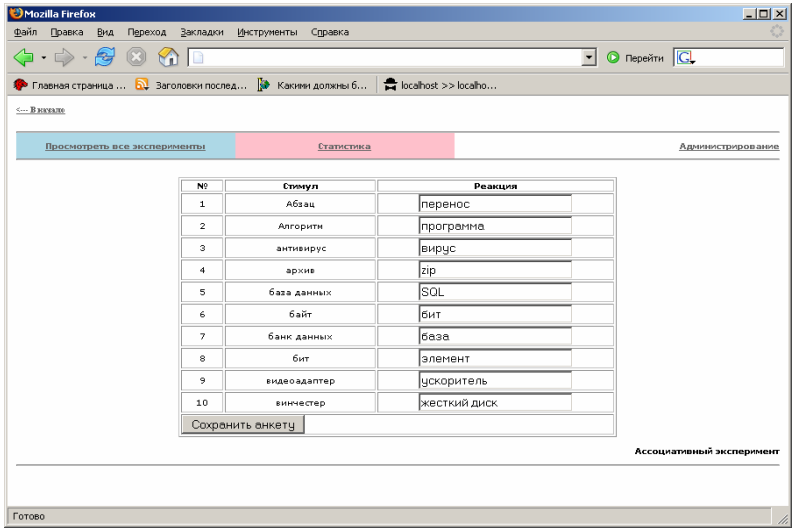


Рис. 5. Заполнение анкеты

Пользователь вводит реакции на все стимулы и нажимает кнопку «Сохранить анкету». После прохождения эксперимента пользователю предлагается просмотреть статистику, в которой отображается относительная чистота реакций на стимулы в эксперименте.

В администраторской части сайта администратор имеет возможность создать эксперимент:

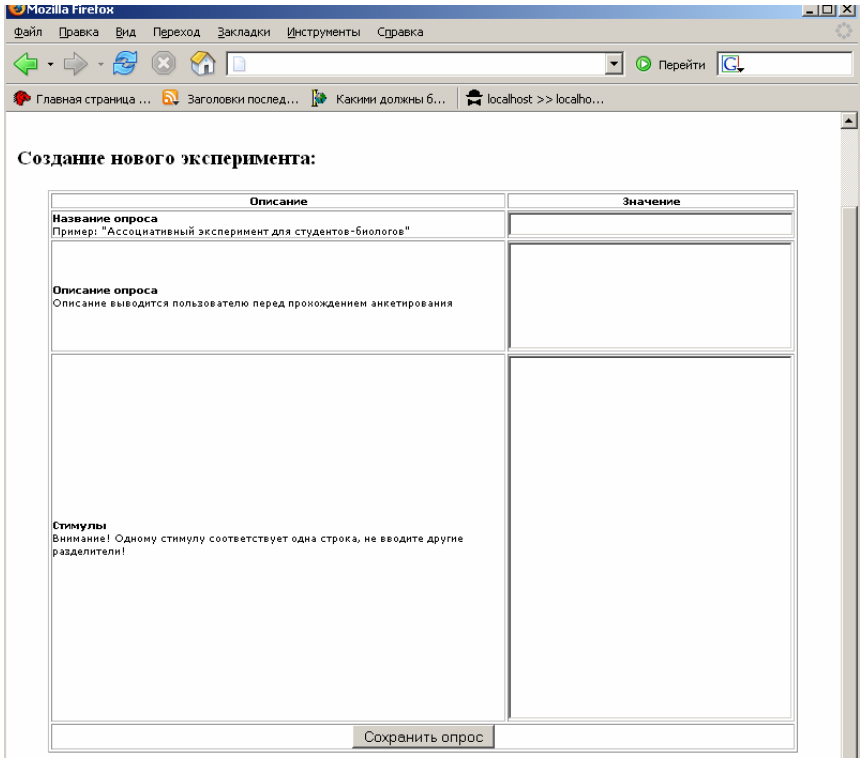
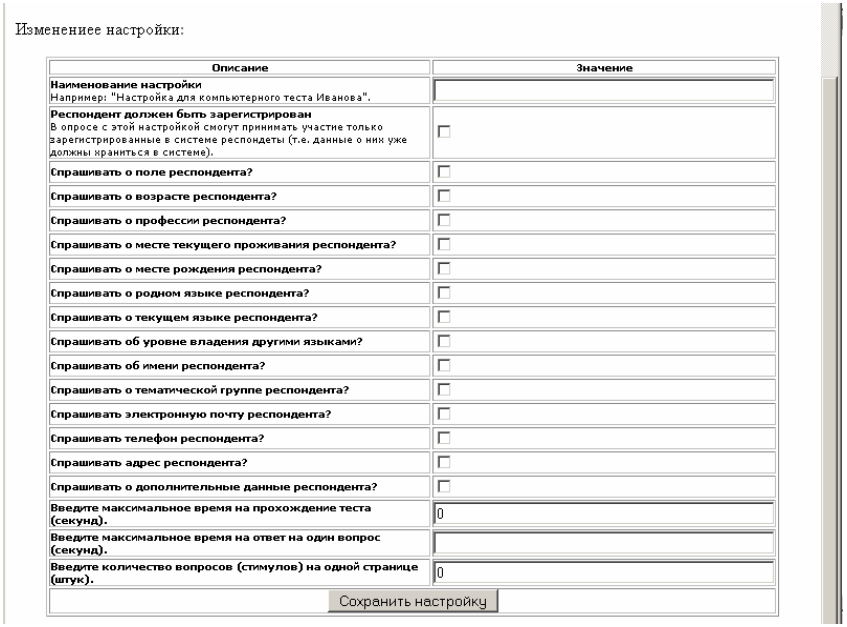


Рис. 6. Создание эксперимента

Администратор эксперимента должен ввести его название, описание и указать список стимулов. Далее для эксперимента автоматически создается настройка, которую автор эксперимента может изменить:


Измененнее настройки:

Описание	Значение
Наименование настройки Например: "Настройка для компьютерного теста Иванова".	
Респондент должен быть зарегистрирован В опросе с этой настройкой смогут принимать участие только зарегистрированные в системе респонденты (т.е. данные о них уже должны храниться в системе).	☐
Спрашивать о поле респондента?	☐
Спрашивать о возрасте респондента?	☐
Спрашивать о профессии респондента?	☐
Спрашивать о месте текущего проживания респондента?	☐
Спрашивать о месте рождения респондента?	☐
Спрашивать о родном языке респондента?	☐
Спрашивать о текущем языке респондента?	☐
Спрашивать об уровне владения другими языками?	☐
Спрашивать об имени респондента?	☐
Спрашивать о тематической группе респондента?	☐
Спрашивать электронную почту респондента?	☐
Спрашивать телефон респондента?	☐
Спрашивать адрес респондента?	☐
Спрашивать о дополнительные данные респондента?	☐
Введите максимальное время на прохождение теста (секунд).	0
Введите максимальное время на ответ на один вопрос (секунд).	
Введите количество вопросов (стимулов) на одной странице (штук).	0
Сохранить настройку	

Рис. 7. Настройка ассоциативного эксперимента

Здесь указываются те данные, которые система будет требовать ввести у каждого респондента. В самом общем случае, для проведения экспериментов могут потребоваться следующие данные:

- Общие данные
- Пол
- Возраст
- Профессия
- Территориальные данные
- Место текущего проживания (Страна, город, район)
- Место рождения (проживания в детском возрасте)
- Языковые данные
- Родной язык
- Владение другими языками (количество)
- Основной язык общения (в текущий период времени)
- Идентификационные данные
- Псевдоним (пароль)
- ФИО

- Тематическая группа
- Контактные данные
- e-mail
- телефон
- адрес
- Дополнительные данные

Также в настройках возможно указание временных характеристик эксперимента (максимальное время эксперимента и максимальное время ответа на стимул) и количества стимулов на экране. В первой версии программы эти функции не реализованы.

Перспективы развития

Первая версия программы является скорее «тестовой», поэтому она содержит лишь необходимый минимум для проведения эксперимента. В то же время в первой версии программы отсутствуют очень важные компоненты, а именно:

1) Возможность генерации разнообразных анкет;

2) Визуальные настройки отображения эксперимента;

3) Учет пользователей и групп пользователей;

4) Развитый инструмент вывода результатов и статистических данных об экспериментах.

Рассмотрим эти функциональные возможности подробнее:

1) Возможность генерации разнообразных анкет. В настоящей версии программы доступен лишь минимальный функционал – каждый эксперимент содержит список стимулов и в каждой анкете любой стимул встречается ровно один раз. В будущих версиях необходимо реализовать возможность генерации анкет – то есть из общего списка стимулов в анкету должно попадать заданное количество, при этом для каждого стимула должна быть предусмотрена возможность указать вероятность его появления. Более того, необходимо ввести механизм ограничений – чтобы стимулы не встречались в экспериментах больше определенного количества раз.

2) Визуальные настройки отображения эксперимента. В текущей версии отсутствует возможность визуальной настройки эксперимента. Визуальная настройка предполагает для каждого лингвистического эксперимента возможность настройки:

- параметров расположения стимулов (количество столбцов, количество символов в столбце, количество стимулов на странице);
- параметров шрифта (размер, стиль);
- цветовых параметров (цвет шрифта, фона, полей ввода)

3) Учет пользователей и групп пользователей. В текущей версии существует только две роли пользователей – администраторы и респонденты. Администраторы могут создавать эксперимент, а респонденты – отвечать на вопросы. В дальнейшем должна быть предусмотрена расширенная система полномочий, включающая авторизацию и аутентификацию пользователей. Как минимум, должны быть созданы и настроены следующие роли:

- Администратор

Пользователи этой роли могут добавлять авторов экспериментов, а также курировать разнообразные системные настройки (резервирование данных, выгрузку и загрузку любых экспериментов, администрировать форумы и т.д.).

- Автор эксперимента

Авторы экспериментов должны иметь возможность создать эксперимент – ввести в систему список стимулов, выбрать настройки формирования анкет и ограничители, указать необходимые настройки отображения, выбрать данные, которые необходимо ввести респонденту о себе и т.д.

- Респондент

Должна быть возможность регистрации респондентов с выдачей каждому из них уникальной пары «имя пользователя – пароль», так, чтобы впоследствии респонденту не требовалось вводить заново о себе.

4) Развитый инструмент вывода результатов и статистических данных об экспериментах. В текущей версии система может только формировать статистику об относительной частоте реакций в разрезе экспериментов. В будущем планируется реализация следующих отчетов, составляющих основу прямого и обратного ассоциативного словаря:

об абсолютной и относительной частоте реакций на заданный стимул (группу стимулов);

об абсолютной и относительной частоте стимулов, породивших выбранную реакцию (группу реакций)

Все статистики могут быть представлены в разрезе одного или всех экспериментов с возможностью разбиения респондентов по раз-

личным критериям на группы (пол, возраст, сфера деятельности, родной язык и т.д.).

Литература

1. *Черкасова Г.А.* Формирование баз лингвистических знаний с использованием технологии ассоциативного эксперимента // Третья Всесоюзная конференция по созданию Машинного фонда русского языка. Тез. докл. Часть 1. М., 1989, с. 197-199.
2. *Черкасова Г.А.* Формальная модель ассоциативного исследования // Scripta linguisticae applicatae. Проблемы прикладной лингвистики. Вып. 2. Сборник статей / Отв. ред. Н.В. Васильева. М.: «Азбуковник»., 2004. С. 139–156. 2004
3. *Филиппович Ю.Н., Черкасова Г.А., Д.Дельфт.* Ассоциации информационных технологий: эксперимент на русском и французском языках. / Серия «Компьютерная лингвистика». Вступ. Статья Н.В.Уфимцевой. М.: МГУП, 2002.— книга в комплекте с CD ROM.

Е.П.Коннова

ОСНОВНЫЕ ПОДХОДЫ К ВЫДЕЛЕНИЮ И РАНЖИРОВАНИЮ БИЗНЕС-ПРОЦЕССОВ

Выделение бизнес-процессов

На сегодняшний день проблемы повышения эффективности бизнеса и усиления его конкурентоспособности особо остро стоят перед российскими предприятиями, переживающими переходный период. Многочисленные примеры успешного внедрения процессного подхода, встречающиеся в западной литературе по менеджменту и иллюстрируемые внушительными цифрами результатов внедрения, позволяют надеяться, что эти проблемы можно эффективно решить при помощи управления на основе бизнес-процессов. Действительно, какие бы новомодные управленческие технологии ни применяла компания - реинжиниринг, "Шесть сигм" или внедрение стандартов ИСО серии 9000:2000 - в основе всегда лежит управление бизнес-процессами [1].

Сутью реинжиниринга является выделение основных бизнес-процессов организации и коренное их изменение для достижения требуемых показателей результативности. Учитывая то, что реинжиниринг заключается, прежде всего, в исследовании и пересмотре бизнес-процессов организации, основными мероприятиями в рамках реинжиниринга являются выделение основных бизнес процессов, описание их на общедоступном языке и анализ с целью дальнейшего преобразования. Главной целью мероприятий по выделению бизнес-процессов является получение цельной картины функционирования организации. Такая картина должна отражать все задействованные в функционировании организации ресурсы, выполняющиеся последовательности процедур, результаты выполнения этих процедур и т.д.

Первым шагом любой из методологий является выделение бизнес-процессов, определение их четких границ и назначение владельцев. Менеджеры, получившие образование в разных местах и имеющие различный опыт, понимают термин "бизнес-процесс", выделяют и описывают процессы по-своему.

Наиболее распространены три основных подхода к определению границ бизнес-процессов, которые часто являются причиной основных

разногласий среди менеджеров, хотя и это деление достаточно условно. Итак, процесс - это последовательность действий, сгруппированных [2]:

1. по виду деятельности (схожие функции);
2. по результату деятельности (продукту);
3. по добавленной ценности для клиента.

Первый подход ориентирован на описание последовательности действий, производимых работниками для достижения результата в рамках своего функционального подразделения; второй позволяет сгруппировать работы по принципу выделения заказчика и продукта для него; третий выделяет и рассматривает процессы как совокупность действий, добавляющих ценность для клиента. Рассмотрим области применения, достоинства и недостатки каждого подхода

Вид деятельности. Первый подход часто применяется при работе над различными проектами автоматизации. При этом делаются "фотографии" существующих и будущих операций на предприятии, зачастую даже без построения моделей верхнего уровня, а если они и строятся, то скорее напоминают функциональную иерархию [2]. Такой подход вполне приемлем для привязки IT-решений к реально действующему предприятию, позволяет на этапе проектирования продемонстрировать заказчику предполагаемые результаты деятельности и вполне адекватно проводить работы по постановке IT-решений и внедрению программного обеспечения.

На схеме 1 показана модель OBM (Oracle Business Model), отображающая такой подход. Модели этого типа отличаются существенным недостатком - предприятие описывается в терминах функциональной деятельности. Поэтому при декомпозиции модели бизнес-процессы и операции описываются как деятельность, распределенная по различным функциональным подразделениям и специалистам, что нарушает главный принцип реинжиниринга - "один процесс - одно подразделение - один бюджет - один владелец процесса". Именно этот принцип (принцип процессного управления) ставили во главу угла М. Хаммер, Дж. Чампи и другие гуру в области реинжиниринга и процессного подхода.

Результат деятельности (продукт). Второй подход основан на выделении процессов по результатам деятельности (а не по предмету, как в предыдущей модели). Наиболее известными моделями, использующими данный подход, являются тринадцати- и восьмипроцессные универсальные модели, а также модель Шеера, их особенность заключается в четком агрегировании работ "по результату".

Схема 1. Модель OBM (Oracle Business Model)

Если при внедрении процессного управления владельцу процесса административно подчиняются все участники процесса, такие модели позволяют разрабатывать и внедрять "плоские" структуры. Такие структуры предъявляют крайне жесткие требования к квалификации исполнителей, плохо понимаются линейными управленцами (заказчиками) и крайне сложны в разработке - в силу высокой абстрагированности принципов и понятий, применяемых при моделировании. В то же время эти структуры, в случае их внедрения, позволяют существенно сокращать численность персонала, в действительности оптимизировать деятельность предприятия, придавать "прозрачность" и управляемость бизнесу. При применении данного подхода следует иметь в виду, что поскольку понятие "результат" само по себе не является однозначным, данный подход предполагает множество вариаций на тему. Наверное, самая большая опасность при применении данного подхода кроется в определении результата, поскольку, умело жонглируя этим понятием, не очень сложно представить каждую функцию как отдельный процесс, в результате которого что-либо производится, а затем объединить полученные "процессы" в уже известную модель "по предмету", тем самым сведя на нет все преимущества данного подхода.

Тринадцатипроцессная модель.

Технический отчет ИСО и МЭК (ИСО/МЭК/ТО 15504), классифицирует не только процессы, но и их результаты, выделяя следующие категории:

- "потребитель - поставщик" – процессы, непосредственно затрагивающие потребителя, поддерживающие разработку и передачу продукта потребителю и обеспечивающие правильные эксплуатацию и использование продукта;
- инженерные - процессы, непосредственно специфицирующие, реализующие и сопровождающие продукт;
- вспомогательные - процессы, результаты которых могут быть использованы в любых других процессах (включая и другие вспомогательные процессы) на различных этапах жизненного цикла продукта;
- управленческие - процессы, содержащие общие действия, которые могут быть использованы теми, кто управляет проектом любого типа или процессом в рамках жизненного цикла продукта;
- организационные - процессы определения бизнес-целей организации и разработки процессов, продуктов или развития активов.

Данные категории могут быть сгруппированы по трем основным типам:

- основные процессы - "потребитель - поставщик", инженерные;
- вспомогательные процессы - вспомогательные;
- организационные процессы - управленческие, организационные.

Тринадцатипроцессная модель (схема 2), часто используемая отечественными консультантами, например, ЗАО "Логика бизнеса", как при проведении семинаров, так и в реальных проектах, составлена по данным Международной бенчмаркинговой палаты (International Benchmarking Clearinghouse). Блок "Производство и поставка продуктов и услуг" обычно разбивается на составляющие: "производство продукции" и "оказание услуг".

Существует также и восьмипроцессная модель, которая была разработана и успешно применяется сотрудниками консалтинговой компании BKG Profit Technology. В своей модели. А.В. Шеер [1] выделяет две ключевые категории основных процессов, вокруг которых группируются информационные и координационные процессы - логистика (материально-техническое обеспечение) заказов и разработка нового изделия.

Схема 2. Тринадцатипроцессная модель.

Добавленная ценность для клиента. Третий подход основывается на описанной М. Портером цепочке создания ценности. В цепочке выделяются основные бизнес-процессы, обеспечивающие операционный цикл производства, выполняющиеся последовательно и поддерживающие бизнес-процессы, обеспечивающие функционирование бизнес-системы и сопровождающие создание продукта на всем протяжении его жизненного цикла [4].

М. Портер указал, что покупатели приобретают не продукт как таковой, а его ценность лично для себя, и поэтому чтобы предприятие могло точно определить свои конкурентные преимущества, необходимо рас-

смотреть всю последовательность процесса создания именно этой ценности[4]. Иными словами, цепочка создания ценности представляет собой инфраструктуру, показывающую значимость бизнес-процессов. Первичными являются бизнес-процессы, предназначенные непосредственно для создания результатов деятельности предприятия - ценности для клиента. Вторичные бизнес-процессы играют вспомогательную роль, обеспечивая необходимую инфраструктуру и средства управления при выполнении первичных бизнес-процессов. При решении вопроса о границах процессов М. Портер предположил, что границы звеньев цепочки, а, следовательно, и бизнес-процессов, находятся там, где каждый внутренний подпроцесс что-то добавляет к ценности продукта. Из этого предположения М. Портера вытекает интересный вывод: не существует стандартного списка бизнес-процессов, каждое предприятие должно разработать собственный перечень основных бизнес-процессов, так как продукт, как ценность для клиента, для каждого предприятия уникален.

Рекомендации по выбору подхода

Вопрос, какой из подходов лучше, остается открытым. Именно в силу применимости всех трех подходов, возникают разногласия и путаница в головах менеджеров, но бесспорным остается только одно - выделение бизнес-процессов, их анализ и последующее совершенствование - колоссальный потенциал для повышения конкурентоспособности компании и эффективности ее работы. В статье приводятся следующие рекомендации для выбора подхода к выделению бизнес-процессов. Выбор и подхода целиком зависит от поставленных целей, преследуемых при постановке задачи описания и оптимизации бизнес-процессов. Следует придерживаться функционального подхода при выделении бизнес-процессов при постановке нижеследующих целей:

Функциональный подход

- Прозрачность, контролируемость и управляемость бизнеса, наведение порядка, реализация стратегии, поддержание роста.
- Построение эффективной организационной структуры реструктуризация.
- Тиражирование бизнеса.
- Правильный подбор персонала, мотивация.
- Уменьшение зависимости от персонала.
- Регламентация и высвобождение времени руководителей.
- Автоматизация.
- Повышение эффективности работы персонала.

Продуктовый подход

- Финансы (себестоимость объектов учета, управленческий учет, бюджетирование).
- Автоматизация.

Ценностный подход

- Реинжениринг бизнес-процессов.
- Система сбалансированных показателей (BSC).
- Проектирование новых бизнес-направлений и бизнес-процессов.

Любой из трех подходов обеспечивает:

- Повышение рыночной стоимости, инвестиционной привлекательности, имиджа, выход на новые рынки, ISO 9000.
- Оптимизация бизнес-процессов.

Методика выделения бизнес-процессов

Первым шагом при работе с процессами следует провести выделение бизнес-процессов верхнего уровня, провести их четкие границы и назначить им владельцев. Среди всего многообразия бизнес-процессов банка следует выделить процессы различного класса: основные; обеспечивающие или вспомогательные; управления; развития.

При этом следует придерживаться следующих критериев классификации и определения. Если бизнес-процесс создают добавленную стоимость предлагаемому компанией продукту, или создает продукт, представляющий ценность для внешнего клиента, представляет процесс, за который клиент готов платить деньги, то данный процесс следует отнести к основному.

В качестве основных бизнес-процессов для банков, например, могут являться следующие процессы [3] (схема 3): кредитования. обслуживания депозитов, кассового обслуживания юридических лиц, кассового обслуживания физических лиц, расчетного обслуживания юридических лиц, осуществления валютно-обменных операций, осуществления услуг по доверительному управлению имуществом, оказания информационно-консультационных услуг, оказания услуг по хранению ценностей в индивидуальных сейфах, оказания услуг по проектному (реальному) инвестированию, осуществления операций с драгоценными металлами.

Если клиентами бизнес-процесса являются основные бизнес-процессы, или бизнес-процессы создают инфраструктуру организации, то такие процессы следует относить к обеспечивающим или вспомогательным. Так, например, вспомогательный бизнес-процесс обеспечения бухгалтерского учета в кредитной организации создает такую ценность

Схема 3. Банковские бизнес-процессы.

для клиента, как лицевой счет, которая используется в большинстве основных бизнес-процессов. Вспомогательными процессами могут являться: обеспечение кассовых операций, ведение бухгалтерского учета, проведение электронных платежей, осуществление внутрихозяйственной деятельности, противодействие «отмыванию» средств, полученных преступным путем.

Бизнес–процессы управления обеспечивают выживание и развитие организации, регулируют ее текущую деятельность. Прямой целью таких бизнес-процессов является управление деятельностью организации. Например, в банковской структуре можно выделить следующие бизнес-процессы управления: управление продуктами, управление стратегическим развитием, планирование и бюджетирование, региональное развитие.

Также следует выделить бизнес-процессы развития (проекты и программы и развития), если таковые существуют в банке. Прямой целью таких процессов является получение прибыли в долгосрочной перспективе. К ним относятся бизнес-процессы совершенствования и развития деятельности организации.

Ранжирование бизнес-процессов

Для осуществления выбора, по каким именно бизнес-процессам в банке необходимо провести моделирование следует провести ранжирование бизнес-процессов. Существует несколько способов определения приоритетности процессов с целью проведения их описания и последующего реинжиниринга. Два из них являются основными: ранжирование по параметрам (методом определения важности и проблемности или с использованием критериев приоритезации по Хаммеру) и ранжирование по стратегической значимости (сопоставление критических факторов успеха и процессов или соответствие стратегии компании).

Ранжирование по параметрам. При выборе ранжирования по параметрам следует преследовать следующую методику[5]:

- Определить параметры важности.
- Разработать методику определения (измерения) параметров важности.
- Определить параметры проблемности.
- Разработать методику определения (измерения) параметров проблемности.
- Разработать систему взвешивания параметров важности и проблемности.

- Построить матрицу ранжирования.

В качестве *параметров важности* банковского бизнес-процесса могут выступать следующие параметры: доля в выручке компании, доля прибыли, потенциал роста. В кредитной организации могут быть разработаны уникальные методики для определения и измерения этих параметров в соответствии с имеющимися средствами автоматизации, подсчета статистик и ведения аналитических данных.

Итоговым оценивающим параметром важности процессов может выступать - *интегральная важность*, которая оценивается как:

$Iv = (P1v * g1v + P2V * g1v + \ldots + Pkv * g1v) / * k$, где:

$P1v, P2V, Pkv$ – параметры важности процесса;

k- количество параметров важности процесса;

$g1v, g2v .. gkv$ – приведенные весовые коэффициенты параметров важности процесса, значения задаются в диапазоне от 0 до 1.

В качестве *параметров проблемности* процесса могут выступать следующие: доля в расходах компании, удовлетворенность клиентов, % сбоев при выполнении процессов, степень фрагментарности.

Итоговым оценивающим параметром проблемности процессов может выступать - *интегральная проблемность*, которая оценивается как:

$Ip = (P1p * g1p + P2V * g2p + \ldots + Pkp * gmp,) / * m$, где:

$P1v, P2V, Pkv$ – параметры проблемности процесса;

m- количество параметров проблемности процесса;

$g1p, g2p .. gkp$ – приведенные весовые коэффициенты параметров проблемности процесса, значения задаются в диапазоне от 0 до 1.

Результаты оценки важности и проблемности бизнес-процессов следует свести в итоговую таблицу оценок. Пример оценки важности и проблемности нескольких бизнес-процессов одного банка сведены в таблицы 1 и 2.

Таблица 1. Важность процесса

Процесс	Доля в выручке компании, %	Доля прибыли, %	Потенциал роста (3 года), %	Интегральная важность, Iv
Кредитные продукты	12	15	100	3,4
Пластиковые карты	55	45	30	4,5
Векселя собственные	30	35	20	2,2

Таблица 2 Проблемность процесса

Процесс	Доля в расходах компании, %	Удовлетворенность клиентов	% сбоев при выполнении процессов	Степень фрагментарности	Интегральная проблемность , Ip
Кредитные продукты	45	5.00	7	40	4.8
Пластиковые карты	37	2	0.4	20	2.7
Векселя собственные	15	75	0.1	24	1.2

Расчет степени фрагментарности и построение матрицы ранжирования бизнес-процесса

В качестве одного из параметров проблемности была выбрана степень фрагментарности бизнес- процесса. В одном бизнес-процессе могут осуществляться несколько несвязанных с собой работ, при этом каждая из работ может быть выполнена как в одном структурном подразделении, так и в различных подразделениях.

Переход выполнения процесса от одной работы к другой называется *функциональным переходом.* Выполнение отдельно взятой работы в отличном от других работ в структурном подразделении несет в себе *организационный разрыв.*

Таким образом, расчет степени фрагментарности бизнес-процессов осуществляется следующим образом:

$Fr = (Norg / Nfunc) * 100\%$, где:

Fr – степень фрагментарности бизнес-процесса;

Norg - количество организационных разрывов

Nfunc - количество функциональных переходов

Степень фрагментарности варьируется от 0 до 100:

$0\% <= Fr <= 100\%$

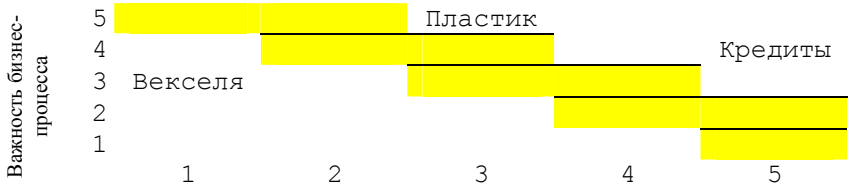
Нулевое значение степени фрагментарности свидетельствует о ее полном отсутствии, 100 % - о полной фрагментарности бизнес-процесса.

При построении матрицы ранжирования бизнес-процессов следует руководствоваться следующими правилами:

- в качестве одного измерения следует выбрать шкалу оценки важности бизнес-процесса;
- в качестве второго измерения выбирается проблемность бизнес-процесса;
- бизнес-процесс согласно вычисленным для него значениям интегральных параметров важности и проблемности размещается в матрице;
- итерация повторяется со всеми оцениваемыми бизнес-процессами;
- в матрице выделяются зоны приоритетности: высокий, средний и низкий приоритеты.

По бизнес-процессам, попавшим в зону с высоким приоритетом, моделирование следует производить в первую очередь. Процессы со средним приоритетом попадают во вторую очередь, с низким – в последнюю. Согласно тестовым данным, приведенным в таблицах 1 и 2, была построена матрица ранжирования.

Таблица 3. Матрица ранжирования бизнес-процессов



высокий приоритет — ячейки в правом верхнем углу таблицы;
средний приоритет — ячейки в середине таблицы;
низкий приоритет — ячейки в левом нижнем углу таблицы.

Критерии установки приоритетов процесса по М. Хаммеру

Согласно теории М. Хаммера следует применять другие критерии для установки приоритетов бизнес-процессов. Бизнес-процесс попадает под установленный критерий, если он отвечает на определенные вопросы:

Таблица 4. Критерии приоритетов процессов по Хаммеру

Беспомощность, дисфункция	Какие процессы выглядят хуже всех остальных, имеют самые большие проблемы в работе банка?
Важность,	Какие процессы имеют наибольшее воздействие на

ценность	внешнего потребителя?
Выполнимость, скорость	Какой процесс дает наиболее важные характеристики и ингредиенты, нужные для легкого и быстрого улучшения ситуации.
Изучение, маркетинг, инфраструктура	Какой процесс дает возможность для улучшения техники и хода дальнейших улучшений?
Зависимость, связи	Какие процессы должны производиться совместно, а какие строиться один на другом?

Процессно-ориентированное управление — подход сложный, но в то же время интересный и перспективный. Некоторые руководители называют его идеальным управленческим инструментом, поскольку он позволяет не только снижать затраты, устраняя не приносящие добавочной стоимости работы, повышать качество обслуживания клиентов и соответственно прибыльность бизнеса, но и принимать стратегически верные решения, ориентируясь на потребности клиента и игнорируя иерархические противостояния.

Литература

1. *Шеер А.-В.* Моделирование бизнес-процессов. — М.: Весть-МетаТехнология, 2000.
2. *Кремлева И.В., Риб С.И.* Различные подходы к выделению и описанию бизнес-процессов, www.betec.ru, 03.2007.
3. *Будков С.Б.* “Базовые принципы реинжиниринга бизнес-процессов в коммерческом банке “, МИТС-НАУКА, № 2, 2007.
4. *Портер М.* Конкуренция: Пер. с англ. / Под ред. Я.В. Заблоцкого. - М.: Издательский дом "Вильямс", 2001.
5. *Кулик В.В., Кремлева И.В.* Материалы семинара “Совершенствование системы управления коммерческим банком на основе описания и оптимизации бизнес-процессов” – Москва, 2007, www.betec.ru.
6. *Шеер А.-В.* Бизнес-процессы. Основные понятия. Теория. Методы. — М.: Весть-МетаТехнология, 1999.

А.А.Кононков

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА ПРИ ПРОВЕДЕНИИ ИССЛЕДОВАНИЙ

Применение кластерного анализа в современных исследованиях в настоящее время получило широкое распространение. Области его использования разнятся от формирования инвестиционных портфелей ценных бумаг при работе банков и паевых инвестиционных фондов до использования алгоритмов кластеризации в биологии. Все большее количество методов автоматической классификации (кластерного анализа) поставляются в качестве функционала известных статистических пакетов, таких как Statistica, SPSS и др., а для того чтобы более эффективно использовать предлагаемые методики необходимо знать суть методов кластерного анализа и понимать какими ограничениями они обладают. В настоящем обзоре представлены, как и монографии, в которых излагается в достаточном объеме теоретический материал по кластерному анализу, так и примеры его применения.

Остановимся чуть более подробно на самом термине кластерный анализ, по сложившейся практике, его следует понимать как обобщенное название множества методов и процедур, используемых при создании классификации. Классификация дает возможность выявить набор схожих объектов исследования, что позволяет значительно упростить их анализ. Проблема выделения однородных групп рассматривается в статистике как задача группировки исходных данных. Принципиально выделяют два типа группировок:

1) типологическая группировка – это разбиение совокупности на качественно однородные группы, характеризующие некоторые типы (классы) явлений, например группировка людей по используемому в их речи диалекту;

2) структурная группировка – расчленение качественно однородной совокупности на группы, характеризующие строение совокупности, ее структуру. Фактически под структурой следует понимать распределение объектов по интервалам.

При формировании групп объектов встает проблема схожести или близости между объектами, она решается введением понятия расстояния

между объектами, выбор меры расстояния существенно влияет на результаты кластеризации, поэтому данной проблеме в литературе уделено достойное место.

Ниже приведен обзор литературы по кластерному анализу, а также его комбинированному использованию с другими методами многомерного статистического анализа.

Представляется целесообразным разбить обзор на две части – в начале обзора рассмотрены три публикации, в которых излагается теория кластерного анализа, а также его методы без конкретизации области исследований, во второй части представлены несколько публикаций опубликованных во всемирной сети Интернета, что подтверждает актуальность применения методов кластеризации в различных исследованиях. Далее в обзоре рассмотрены статьи, рассматривающие применения кластерного анализа в конкретной области, например социальных исследованиях, это сделано для того, чтобы показать, что выбор методов все же зависит от области исследования и ожидаемого результата.

Дюран Б. и Оделл П. Кластерный анализ. Пер. с Демиденко. Под ред. Боярского.- М.: «Статистика», 1974.- 128 с.

Обзор состояния теории и практики применения «кластерного анализа» приведен в книге Дюрана и Одель. В ней авторы подчеркивают, что этот метод имеет все преимущества метода комбинаторной группировки, но свободен от его главного недостатка – распыления материала, что открывает большие перспективы применения рассматриваемого метода в статистическом анализе, в классификации объектов, в исследовании связей и типизации выборки. Эта книга отличается полнотой и доступностью и вместе с тем краткостью изложения материала. Она разбита на 6 глав – в начале книги даются основные идеи кластеризации, дальше затрагиваются проблемы, с которыми сталкивается исследователь, если использует кластеризацию полным перебором. В книге также освещаются методы графического представления матриц сходства и собственно результатов кластеризации – наиболее авторы выделяют метод дендограмм.

Мендель И.Д. Кластерный анализ.- М.: Финансы и статистика, 1988.- 176с.: ил.

В данной книге предпринята попытка систематизировать и дать сравнительный анализ многочисленных алгоритмов кластеризации. В ней также приводятся примеры использования кластерного анализа в социально-экономических исследованиях. В книге наилучшим образом изложена проблема измерения близости объектов, с которой сталкиваются все исследователи, пытающиеся применять кластерный анализ в своих работах. Это обусловлено тем, что объекты обычно имеют параметры, изме-

ряемые в различных величинах, иногда даже качественные, в этом случае встает проблема нормировки значений и формализации отношения близости или удаления объектов друг от друга.

Ким Дж. – О. Ким Факторный, дискриминантный, кластерный анализ: пер с англ./ Дж. – О. Ким, Ч.У. Мюллер, У.Р. Клекка и др.; Под ред. И.С. Енюкова.- М.: Финансы и статистика,1989.- 215 с.: ил.

Книга представляет сборник работ американских ученых, в которых рассмотрен аппарат факторного, дискриминантного и кластерного анализа, широко применяемый в социально-экономических классификациях и анализе неявных закономерностей в экономических и социальных процессах. В настоящую книгу вошли три выпуска серии «Quantitative Application in Social Sciences» издаваемых американской фирмой «Sage Publication», представляющие некоторые важнейшие методы многомерного статистического анализа.

Данное издание является хорошим справочником для научных работников, преподавателей, студентов, аспирантов экономических вузов.

Айвазян С.А. Прикладная статистика : Классификация и снижение размерности: Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Е.Д. Мешалкин; под ред. С.А. Айвазяна.- М: Финансы и статистика, 1989.- 607 с.

Книга является третьей в трехтомном справочном издании этих же авторов посвященным прикладной статистике. Данный том посвящен актуальным аспектам общей проблемы статистического анализа данных – задачам классификации объектов, снижению размерности исследуемого признакового пространства и статистическим методам их решения. В настоящее время, когда вычислительные мощности уже достигли определенного уровня, главной проблемой теории и практики классификации и снижения размерности стало развитие достаточно изощренного и эффективного в приложениях математического аппарата, при написании данной книги авторы ставили целью достаточно полно систематизировать все достижения. В результате им удалось выстроить методологическую схему применения методов статистического анализа данных, снабдив ее необходимыми практическими рекомендациями (включая вопросы преодоления вычислительных трудностей и использования подходящего программного обеспечения).

Настоящая книга будет полезна для специалистов, применяющих методы анализа данных.

Абдулганеев М.Т., Владимиров В.Н. Типология поселений Алтая 6-2 вв. до н.э.: Монография. Барнаул: Изд-во Алт. ун-та, 1997. 148 с.

В монографии представлена типология поселений лесостепного, степного и предгорного Алтая скифского времени, созданная на основе применения методов многомерного статистического анализа (факторный анализ, кластер-анализ, fuzzy classification). В основе типологии лежит изучение орнаментальных традиций основных археологических культур, распространенных на территории Алтая в то время. Весь анализ проведен на компьютере с использованием стандартного и оригинального программного обеспечения. Это позволило внести существенные коррективы в культурно-хронологическую схему развития населения Алтая в скифское время, четко разграничить выделенные на этой территории археологические культуры.

Кораблёва Г.В., Широков Л.А. Алгоритмы классификации и кластеризации компонентов сложных изделий для адаптивной системы планирования Машиностроительного производства.// Материалы ежегодной конференции.-Свободный доступ.- http://www.msiu.ru/conference/section_2/2_6_conference_1.doc

В статье представлены принципы эффективности планирования производства, на примере машиностроительного производства. Показан один из возможных путей достижения оптимальности и эффективности процесса планирования. Он основан на использовании экономико-математических моделей, которые позволяют сократить количество уравнений для расчета модели планирования производства. С целью уменьшения времени нахождения оптимального производственного плана применяются методы классификации и кластеризации. Также приведен пример использования методов классификации для уменьшения анализируемых данных, в частности: с целью сокращения числа строк в матрице запасов ресурсов производится их классификация по известным признакам. При классификации применяются многомерные методы обобщенных расстояний Махаланобиса и алгоритм Беккера.

Статья будет полезна для исследователей, желающих ознакомиться с практикой применения кластерного анализа в конкретной области.

Коновалов А.В. Об особенностях применения многомерного статистического анализа в исторических исследованиях.- Свободный доступ.- http://sodmu.narod.ru/archive/konovalov.doc

В статье приведен пример использования многомерного анализа в исторических исследованиях. На основе совместного использования кластерного, и факторного анализов. Проводились исследование социального и материального деления хуторян и общинников Симбирской губернии, данные для исследований были взяты из массового статистического

источника – «Кратких бюджетных сведений по хуторским и общинным крестьянским хозяйствам Симбирской губернии».

Методы многомерного анализа применяются последовательно, сначала с помощью агломеративно-иерархического метода кластерного анализа (основная идея этого метода заключается в последовательном объединении группируемых объектов – сначала самых близких, затем все более удаленных друг от друга) получают группировку объектов в исходном многомерном пространстве признаков, а затем с помощью факторного анализа выявляют небольшое количество основных факторов. В результате исследователи получают наглядную интерпретацию результатов на плоскости. Что дает исследователю возможность более четко интерпретировать полученное разбиение на классы.

Ценность статьи заключается в примере комбинирования методов многомерного статистического анализа, а также примере использования автоматических процедур кластеризации пакета Statistica.

С.А.Айвазян, В.С.Степанов Программное обеспечение по статистическому анализу данных: методология сравнительного анализа и выборочный обзор рынка.- Свободный доступ.- http://stphs.narod.ru/SAS/STatProg.htm

В статье описывается методология сравнительного анализа множества однотипных статистических программных продуктов (СПП), а также подходы к ценообразованию на рынке СПП. Предлагаемая методология является развитием и определенной коррекцией методики, разработанной и активно используемой Национальной лабораторией по тестированию программных продуктов (США). Предложения авторов статьи сводятся, в основном, к необходимости включения в принятую ранее блок-схему нового важного базового свойства «Степень интеллектуализации СПП», к разработке существенно иного подхода к определению «весов» детализированных характеристик и различных базовых свойств, а также к увязке этих вопросов с вопросами ценообразования на рынке СПП.

Статья содержит также выборочный аналитический обзор мирового рынка СПП, особенно подробный применительно к пакетам, решающим задачи статистической классификации и снижения размерности (независимо от предметной области их применения).

Содержащаяся в статье информация может оказаться полезной пользователям (в частности, в более квалифицированном подходе к решению вопроса о приобретении того или иного пакета), специалистам-разработчикам наукоемких программ и исследователям, а также руководителям различных аналитических служб (банков, бирж, маркетинговых

и консалтинговых агентств, экономических, медицинских, геофизических, промышленных исследовательских центров и т.п.).

Многомерный статистический анализ в социально-экономических исследованиях// Под ред. Айвазян С.А., Френкель А.А - М.:«Наука», 1974.-416с.

В настоящем издании освещены теоретические, методологический и экономико-прикладные аспекты многомерного статистического анализа, рассматриваются вопросы построения экономико-статистических моделей с помощью факторного анализа, метода главных компонент, кластерного анализа и других методов многомерной статистики. Большое внимание уделено описанию основных алгоритмов и их приложений в экономике, приведены результаты различных практических исследований с применением многомерного статистического анализа. Эта книга интересна тем, что в ней можно проследить, как разные ученые решают схожие проблемы кластеризации данных и сравнить различные методы кластерного анализа.

Ле Ба Хонг Минь

АЛГОРИТМ ПЕРЕВОДА СЛОВ С РУССКОГО ЯЗЫКА НА ВЬЕТНАМСКИЙ

Одной из главных функций системы машинного перевода является функция «Трансфер». Основной ее задачей является перевод русской семантической сети во вьетнамскую семантическую сеть. Из результатов курсовой работы «Формальные признаки структур для перевода» нам известно, что машинный перевод будет правилен, если две сети полиморфны, то есть их графы схожи по дугам и вершинам. В этой статье мы рассматриваем алгоритм нахождения вьетнамского слова, которое больше всего соответствует русскому не только с грамматической точки зрения, но и с точки зрения лексики, то есть значения этого слова.

Элементы модуля «Трансфер»

1. Входные данные

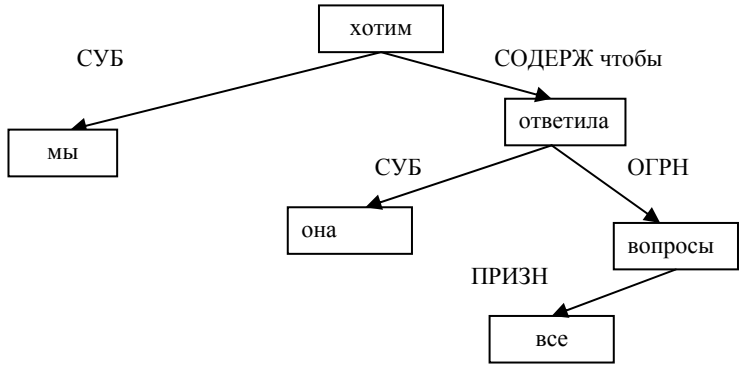
Система машинного перевода – это последовательность различных элементов системы. Результаты предыдущего элемента процесса являются входными данными для следующего элемента. Входными данными для модуля «Трансфер» являются выходные данные модуля «Семантического анализа», то есть семантическая сеть предложения, требующего перевода. В этом графе вершинами являются слова или словосочетания, у которых определены лексические значения и грамматические свойства, а дуги отображают отношения между словами.

Пример структуры слова в словаре:

статья Валить:
КАТ = 1 ЭТК.СИТ
ГХ = 1 ГЛ:ГГ
ВАЛ = АГЕНТ , А1 , С
ОБ , А2 , С
ГХ1 = 1 подл : И
СХ1 = 1 ОДУШ
ГХ2 = 1 п_доп :

Семантическая сеть.

Мы хотим, чтобы она ответила на все вопросы



2. Словарь

Бинарный словарь — определяет все вьетнамские слова, которые подходят по значения к русскому слову.

Пример:

Гулять: Đi chơi
Đi nghỉ

ВОСС вьетнамский общесемантический словарь — определяет все грамматические и лексические значения вьетнамских слов. Структура этого словаря похожа на структуру словаря РОСС.

Пример:

статья " địa chỉ" :
ГХ = 1 VERB
СХ = 1 КОММУНИК
ВАЛ = СУБ , А1 , С
АДР , А2 , С
СОДЕРЖ , А3 , С
СПОСОБ , А4 , С
ГХ1 = 1 chủ thể
ГХ2 = 1 đối tượng
СХ2 = 1 ОДУШ
ГХ3 = 1 đối tượng
2 cùng với + danh từ
СХ3 = 1 ИНФ
2 СИТУАТ

ГХ4 = 1 như + danh từ
НЕСОВМ = 2.1 , 3.1

3. Алгоритм выбора слова

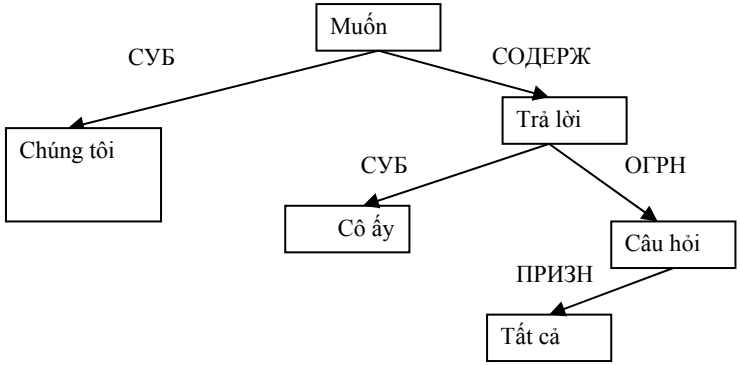
Задача этого алгоритма выбор вьетнамских слов, которые наиболее соответствуют русскому слову в этой семантической сети, используя данные словаря.

4. Выходные данные

На выходе функции «Трансфер» будет получена вьетнамская семантическая сеть, которая идентична входящей. Каждая вершина этой сети – вьетнамское слово, грамматические значения которых определены.

Пример:

Мы хотим, чтобы она ответила на все вопросы – Chúng tôi muốn cô ấy trả lời tất cả câu hỏi.



Алгоритм функции «Трансфер»

1. Получить семантическую сеть, соответствующую русскому предложению.

2. Получить слово из этой сети.

3. Найти перевод этого слова на вьетнамский язык.

3.1. Если было найдено только одно похожее вьетнамское слово, то это слово является результатом функции.

3.2. Если русскому слову соответствует несколько вариантов вьетнамских слов, то

3.2.1. Определить грамматические и лексические характеристики слов, используя словарь ВОСС.

3.2.2. Для каждого слова определить вероятность связи этого слова с другими словами семантической сети, которые связаны с этим словом

3.2.3. Выбрать слово имеющее самую большую вероятность

4. Построить семантическую сеть вьетнамского предложения с помощью слов определенных на предыдущем этапе, соответствующую семантической сети русского предложения.

Литература

1. *Н.Н.Леонтьева* “Автоматическое понимание текстов: системы, модели, ресурсы” — М.: ACADEMA, 2006 г.
2. “Lý thuyết đồ thị và ứng dụng ” — Ханой: Научная редакция, 1980 г.
3. “Трансфер” http://www.aot.ru/docs/transfer.html

И.М.Нейский

КЛАССИФИКАЦИЯ И СРАВНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Методы кластерного анализа являются наиболее распространенными в различных технологиях обработки данных. Приведем их классификацию и описание.

Методы по способу обработки данных [1]:

Иерархические методы:

Агломеративные методы AGNES (Agglomerative Nesting):

- CURE;
- ROCK;
- CHAMELEON и т.д.

Дивизимные методы DIANA (Divisive Analysis):

- BIRCH;
- MST и т.д.

Неиерархические методы.

Итеративные

- К-средних (k-means);
- PAM (k-means + k-medoids);
- CLOPE;
- LargeItem и т.д.

Методы по способу анализа данных:

- Четкие;
- Нечеткие.

Методы по количеству применений алгоритмов кластеризации:

- С одноэтапной кластеризацией;
- С многоэтапной кластеризацией.

Методы по возможности расширения объема обрабатываемых данных:

- Масштабируемые;
- Немасштабируемые.

Методы по времени выполнения кластеризации [1]:

- Потоковые (on-line);
- Не потоковые (off-line).

Иерархическая кластеризация

При иерархической кластеризации выполняется последовательное объединение меньших кластеров в большие или разделение больших кластеров на меньшие [1].

Агломеративные методы AGNES (Agglomerative Nesting).

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

АЛГОРИТМ CURE
(CLUSTERING USING REPRESENTATIVES).

Выполняет иерархическую кластеризацию с использованием набора определяющих точек для определения объекта в кластер.

Назначение: кластеризация очень больших наборов числовых данных.

Ограничения: эффективен для данных низкой размерности, работает только на числовых данных.

Достоинства: выполняет кластеризацию на высоком уровне даже при наличии выбросов, выделяет кластеры сложной формы и различных размеров, обладает линейно зависимыми требованиями к месту хранения данных и временную сложность для данных высокой размерности.

Недостатки: есть необходимость в задании пороговых значений и количества кластеров.

Описание алгоритма [3]:

Шаг 1: Построение дерева кластеров, состоящего из каждой строки входного набора данных.

Шаг 2: Формирование «кучи» в оперативной памяти, расчет расстояния до ближайшего кластера (строки данных) для каждого кластера. При формировании кучи кластеры сортируются по возрастанию дистанции от кластера до ближайшего кластера. Расстояние между кластерами определяется по двум ближайшим элементам из соседних кластеров. Для определения расстояния между кластерами используются «манхеттенская», «евклидова» метрики или похожие на них функции.

Шаг 3: Слияние ближних кластеров в один кластер. Новый кластер получает все точки входящих в него входных данных. Расчет расстояния до остальных кластеров для новообразованного кластера. Для рас-

чета расстояния кластеры делятся на две группы: первая группа – кластеры, у которых ближайшими кластерами считаются кластеры, входящие в новообразованный кластер, остальные кластеры – вторая группа. И при этом для кластеров из первой группы, если расстояние до новообразованного кластера меньше чем до предыдущего ближайшего кластера, то ближайший кластер меняется на новообразованный кластер. В противном случае ищется новый ближайший кластер, но при этом не берутся кластеры, расстояния до которых больше, чем до новообразованного кластера. Для кластеров второй группы выполняется следующее: если расстояние до новообразованного кластера ближе, чем предыдущий ближайший кластер, то ближайший кластер меняется. В противном случае ничего не происходит.

Шаг 4: Переход на шаг 3, если не получено требуемое количество кластеров.

Дивизимные методы DIANA (Divisive Analysis).

Эта группа методов характеризуется последовательным разделением исходного кластера, состоящего из всех объектов, и соответствующим увеличением числа кластеров.

В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

АЛГОРИТМ BIRCH
(BALANCED ITERATIVE REDUCING AND CLUSTERING USING HIERARCHIES).

В этом алгоритме предусмотрен двухэтапный процесс кластеризации.

Назначение: кластеризация очень больших наборов числовых данных.

Ограничения: работа с только числовыми данными.

Достоинства: двухступенчатая кластеризация, кластеризация больших объемов данных, работает на ограниченном объеме памяти, является локальным алгоритмом, может работать при одном сканировании входного набора данных, использует тот факт, что данные неодинаково распределены по пространству, и обрабатывает области с большой плотностью как единый кластер.

Недостатки: работа с только числовыми данными, хорошо выделяет только кластеры сферической формы, есть необходимость в задании пороговых значений.

Описание алгоритма [4]:

Фаза 1: Загрузка данных в память.

Построение начального кластерного дерева (CF Tree) по данным (первое сканирование набора данных) в памяти;

Подфазы основной фазы происходят быстро, точно, практически нечувствительны к порядку.

Алгоритм построения кластерного дерева (CF Tree):

Кластерный элемент представляет собой тройку чисел (N, LS, SS), где N – количество элементов входных данных, входящих в кластер, LS – сумма элементов входных данных, SS – сумма квадратов элементов входных данных.

Кластерное дерево – это взвешенно сбалансированное дерево с двумя параметрами: B – коэффициент разветвления, T – пороговая величина. Каждый нелистьевой узел дерева имеет не более чем B вхождений узлов следующей формы: $[CF_i, Child_i]$, где i = 1, 2, …, B; $Child_i$ – указатель на i-й дочерний узел.

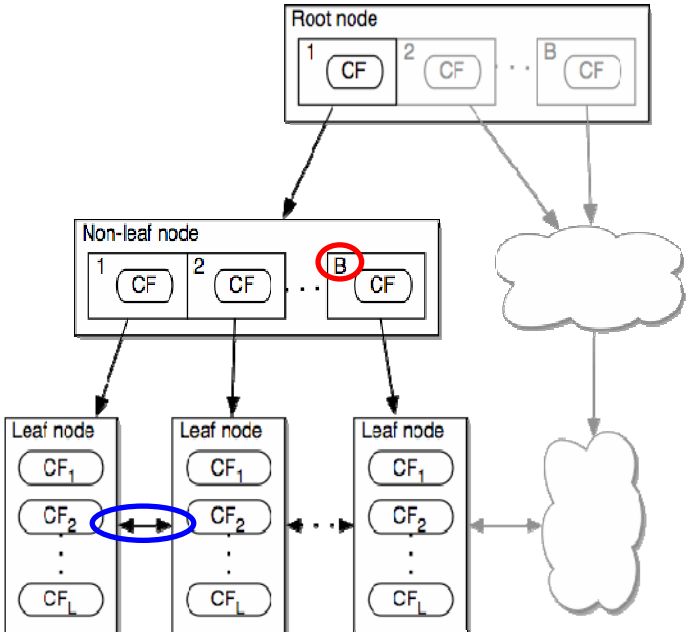


Рис. 1. Построение кластерного дерева.

Каждый листьевой узел имеет ссылку на два соседних узла. Кластер, состоящий из элементов листьевого узла, должен удовлетворять следующему условию: диаметр или радиус полученного кластера должен быть не более пороговой величины T.

Фаза 2 (необязательная): Сжатие (уплотнение) данных.

Сжатие данных до приемлемых размеров с помощью перестроения и уменьшения кластерного дерева с увеличением пороговой величины T.

Фаза 3: Глобальная кластеризация.

Применяется выбранный алгоритм кластеризации на листьевых компонентах кластерного дерева.

Фаза 4 (необязательная): Улучшение кластеров.

Использует центры тяжести кластеров, полученные в фазе 3, как основы.

Перераспределяет данные между «близкими» кластерами. Данная фаза гарантирует попадание одинаковых данных в один кластер.

АЛГОРИТМ MST
(ALGORITHM BASED ON MINIMUM SPANNING TREES).

Назначение: кластеризация больших наборов произвольных данных.

Достоинства: выделяет кластеры произвольной формы, в т.ч. кластеры выпуклой и вогнутой формы, выбирает из нескольких лучших решений оптимальное.

Описание алгоритма [5]:

Шаг 1: Построение минимального остовного дерева:

Связный, неориентированный граф с весами на ребрах G(V, E), в котором V — множество вершин (контактов), а E — множество их возможных попарных соединений (ребер), для каждого ребра (u,v) однозначно определено некоторое вещественное число w(u,v) — его вес (длина или стоимость соединения).

Алгоритм Борувки:

1: Для каждой вершины графа находим ребро с минимальным весом.

2: Добавляем найденные ребра к остовному дереву, при условии их безопасности.

3: Находим и добавляем безопасные ребра для несвязанных вершин к остовному дереву.

Общее время работы алгоритма: O(ELogV).

Алгоритм Крускала:

Обход ребер по возрастанию весов. При условии безопасности ребра добавляем его к остовному дереву.

Общее время работы алгоритма: O(ELogE).

Алгоритм Прима:

1: Выбор корневой вершины.

2: Начиная с корня, добавляем безопасные ребра к остовному дереву.

Общее время работы алгоритма: O(ELogV).

Шаг 2: Разделение на кластеры. Дуги с наибольшими весами разделяют кластеры.

Представление алгоритмов построения остовных деревьев на примерах:

Неиерархическая кластеризация.

АЛГОРИТМ K-СРЕДНИХ (K-MEANS).

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Ограничения: небольшой объем данных.

Достоинства: простота использования; быстрота использования; понятность и прозрачность алгоритма.

Недостатки: алгоритм слишком чувствителен к выбросам, которые могут искажать среднее; медленная работа на больших базах данных; необходимо задавать количество кластеров.

Описание алгоритма [5]:

Этап 1. Первоначальное распределение объектов по кластерам.

Выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр.

Выбор начальных центроидов может осуществляться следующим образом:

выбор k-наблюдений для максимизации начального расстояния;

случайный выбор k-наблюдений;

выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

Этап 2. Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;

число итераций равно максимальному числу итераций.

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.

АЛГОРИТМ PAM (PARTITIONING AROUND MEDOIDS).

Ограничения: небольшой объем данных.

Достоинства: простота использования; быстрота использования; понятность и прозрачность алгоритма, алгоритм менее чувствителен к выбросам в сравнении с k-means.

Недостатки: необходимо задавать количество кластеров; медленная работа на больших базах данных.

Описание алгоритма [5]:

Данный алгоритм берет на вход множество S и число кластеров k, на выходе алгоритм выдает разбиение множества S на k-кластеров: S_1, S_2, …, S_k. Этот алгоритм аналогичен алгоритму k-средних, только при работе алгоритма перераспределяются объекты относительно медианы кластера, а не его центра.

АЛГОРИТМ CLOPE.

Назначение: кластеризация огромных наборов категорийных данных.

Достоинства: высокие масштабируемость и скорость работы, а так же качество кластеризации, что достигается использованием глобального критерия оптимизации на основе максимизации градиента высоты гистограммы кластера. Он легко рассчитывается и интерпретируется. Во время работы алгоритм хранит в RAM небольшое количество информации по каждому кластеру и требует минимальное число сканирований набора данных. CLOPE автоматически подбирает количество кластеров, причем это регулируется одним единственным параметром – коэффициентом отталкивания.

Описание алгоритма [2]:

Пусть имеется база транзакций D, состоящая из множества транзакций $\{t_1,t_2,\ldots,t_n\}$. Каждая транзакция есть набор объектов $\{i_1,\ldots,i_m\}$. Множество кластеров $\{C_1,\ldots,C_k\}$ есть разбиение множества $\{t_1,\ldots,t_n\}$, такое, что $C_1 \cup \ldots \cup C_k=\{t_1,\ldots,t_n\}$ и $C_i \neq \emptyset$ и $C_i \cap C_j = \emptyset \ \forall\ i \geq 1,\ k \geq j$. Каждый элемент C_i называется кластером, а n, m, k – количество транзакций, количество объектов в базе транзакций и число кластеров соответственно.

Каждый кластер C имеет следующие характеристики:

D(C) – множество уникальных объектов;

Occ(i,C) – количество вхождений (частота) объекта i в кластер C;

$$S(C) = \sum_{i \in D(C)} Occ(i,C) = \sum_{t_i \in C} |t_i|,$$

$$W(C) = |D(C)|, H(C) = S(C)/W(C)$$

Функция стоимости:

$$Profit(C) = \frac{\sum_{i=1}^{k} G(C_i) * |C_i|}{\sum_{i=1}^{k} |C_i|} = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} * |C_i|}{\sum_{i=1}^{k} |C_i|},$$ где

$|C_i|$ – количество объектов в i-ом кластере, k – количество кластеров, r – коэффициент отталкивания ($0 < r \leq 1$).

С помощью параметра r регулируется уровень сходства транзакций внутри кластера, и, как следствие, финальное количество кластеров. Этот коэффициент подбирается пользователем. Чем больше r, тем ниже уровень сходства и тем больше кластеров будет сгенерировано.

Формальная постановка задачи кластеризации алгоритмом CLOPE выглядит следующим образом: для заданных D и r найти разбиение C: $Profit(C,r) \rightarrow \max$.

САМООРГАНИЗУЮЩИЕСЯ КАРТЫ КОХОНЕНА.

Назначение: кластеризация многомерных векторов, разведочный анализ данных, обнаружение новых явлений.

Достоинства: используется универсальный аппроксиматор – нейронная сеть, обучение сети без учителя, самоорганизация сети, простота реализации, гарантированное получение ответа после прохождения данных по слоям.

Недостатки: работа только с числовыми данными, минимизация размеров сети, необходимо задавать количество кластеров.

Описание алгоритма [6]:

Самоорганизующая карта Кохонена – нейронная однослойная сеть прямого распространения.

Этап 1. Подготовка данных для обучения.

Обучающая выборка должна быть представительной, не должна быть противоречивой, преобразование и кодирование данных, нормализация данных.

Этап 2. Начальная инициализация карты.

На этом этапе выбирается количество кластеров и, соответственно, количество нейронов в выходном слое.

Перед обучением карты необходимо проинициализировать весовые коэффициенты нейронов сети. Существуют два способа инициализации весовых коэффициентов:

Инициализация случайными значениями, когда всем весам даются малые случайные величины.

Инициализация примерами, когда в качестве начальных значений задаются значения случайно выбранных примеров из обучающей выборки.

Этап 3. Обучение сети.

Карта Кохонена представляет собой нейронную сеть, состоящую из двух слоев: входного и выходного.

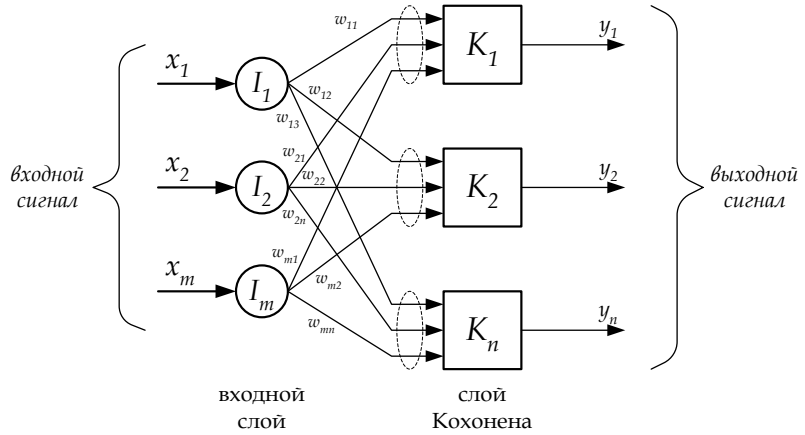


Рис. 2. Карта Кохонена.

Обучение состоит из последовательности коррекций векторов, представляющих собой нейроны. На каждом шаге обучения из исходного набора данным случайно выбирается один из векторов, а затем производится поиск наиболее похожего на него вектора коэффициентов нейронов. При этом выбирается нейрон – победитель, который наиболее похож на вектор входов. Под похожестью в данной задаче понимается расстояние между векторами, обычно вычисляемое в евклидовом пространстве. Таким образом, если обозначить нейрон – победитель как c, то получим: $\|x - w_c\| = \min_i \{\|x - w_i\|\}$.

После того, как найден нейрон – победитель производится корректировка весов нейросети. При этом вектор, описывающий нейрон – победитель, и вектора, описывающие его соседей в сетке, перемещаются в направлении входного вектора.

Для модификации весовых коэффициентов используется формула:

$$w_i(t+1) = w_i(t) + h_{ci}(t) * [x(t) - w(t)],$$

где t – номер эпохи, вектор x(t) – выбирается случайно из обучающей выборки на итерации t, функция h(t) – функция соседства нейронов.

Функция соседства нейронов представляет собой невозрастающую функцию от времени и расстояния между нейроном – победителем и соседними нейронами в сетке. Эта функция разбивается на две части: функцию расстояния и функцию скорости обучения от времени.

$$h(t) = h(\|r_c - r_i\|, t) * a(t),$$

где r – определяет положение нейрона в сетке, a(t) – функция скорости обучения сети.

Функции от расстояния применяются двух видов:

Простая константа: $$h(d,t) = \begin{cases} const, d \le \delta(t) \\ 0, d > \delta(t) \end{cases}$$

Гауссова функция: $$h(d,t) = e^{-\frac{d^2}{2*\delta^2(t)}}$$

Функция скорости обучения сети: $$a(t) = \frac{A}{t+B},$$

где A и B – константы.

Обучение состоит из двух основных фаз: на первоначальном этапе выбирается достаточно большое значение скорости обучения и радиуса обучение, что позволяет расположить вектора нейронов в соответствии с распределением примеров в выборке, а затем производится точная подстройка весов, когда значения параметров скорости обучения много меньше начальных. В случае использования линейной инициализации первоначальный этап грубой подстройки может быть пропущен.

Этап 4. Использование карты.

При использовании карты входной вектор предъявляется на вход, после чего на выходе активизируется нейрон или группа нейронов, которые соответствуют тому или кластеру, полученному в процессе обучения сети.

АЛГОРИТМ HCM (HARD C – MEANS).

Назначение: кластеризация больших наборов числовых данных.

Достоинства: легкость реализации, вычислительная простота.

Недостатки: задание количества кластеров, отсутствие гарантии в нахождении оптимального решения.

Описание алгоритма [7]:

Шаг 1. Инициализация кластерных центров c_i (i = 1, 2, …, c). Это можно сделать выбрав случайным образом c – векторов из входного набора.

Шаг 2. Вычисление рядовой матрицы M.

Матрица M состоит из элементов m_{ik}:

$$m_{ik} = \begin{cases} 1, \left\|u_k - c_i\right\|^2 \le \left\|u_k - c_j\right\|^2 \\ 0, остальное \end{cases}$$, для всех i ≠ j, i = 1, 2, …, c, k = 1, 2, …, K, где K – количество элементов во входном наборе данных.

Матрица M обладает следующими свойствами:

$$\sum_{i=1}^{c} m_{ij} = 1, \sum_{j=1}^{K} m_{ij} = K$$

Шаг 3. Расчет объектной функции:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, u_k \in C_i} \left\|u_k - c_i\right\|^2 \right)$$

На этом шаге происходит остановка и выход из цикла, если полученное значение ниже пороговой величины или полученное значение не сильно отличается от значений, полученных на предыдущих циклах.

Шаг 4. Пересчет кластерных центров.

Пересчет кластерных центров выполняется в соответствии со следующим уравнением:

$$c_i = \frac{1}{|Ci|} * \sum_{k, u_k \in C_i} u_k ,$$

где $|C_i|$ – количество элементов в i-ом кластере.

Шаг 5. Переход на шаг 2.

Нечеткая кластеризация

АЛГОРИТМ FUZZY C-MEANS.

Назначение: кластеризация больших наборов числовых данных.

Достоинства: нечеткость при определении объекта в кластер позволяет определять объекты, которые находятся на границе, в кластеры.

Недостатки: вычислительная сложность, задание количества кластеров, возникает неопределенность с объектами, которые удалены от центров всех кластеров.

Описание алгоритма [7]:

Пусть нечеткие кластеры задаются матрицей разбиения:

$F = [\mu_{ki}], \mu_{ki} \in [0,1], k = \overline{1, M}, i = \overline{1, c}$, где μ_{ki} - степень принадлежности объекта k к кластеру i, c – количество кластеров, M – количество элементов.

При этом: $\sum_{i=1}^{c} \mu_{ki} = 1, k = \overline{1, M}$; $0 < \sum_{k=1}^{M} \mu_{ki} < M, i = \overline{1, c}$. (1)

Этап 1. Установить параметры алгоритма: c - количество кластеров; m - экспоненциальный вес, определяющий нечеткость, размазаность кластеров ($m \in [1, \infty)$); ε - параметр останова алгоритма.

Этап 2. Генерация случайным образом матрицы нечеткого разбиения с учетом условий (1).

Этап 3. Расчет центров кластеров: $V_i = \frac{\sum_{k=1}^{M} {\mu_{ki}}^m * |X_k|}{\sum_{k=1}^{M} {\mu_{ki}}^m}, i = \overline{1, c}$

Этап 4. Расчет расстояния между объектами X и центрами кластеров:

$$D_{ki} = \sqrt{\left\| X_k - V_i \right\|^2}, k = \overline{1, M}, i = \overline{1, c}$$

Этап 5. Пересчет элементов матрицы разбиения с учетом следующих условий:

если $D_{ki} > 0 : \mu_{kj} = \dfrac{1}{\left({D_{jk}}^2 * \sum_{j=1}^{c} \dfrac{1}{{D_{jk}}^2} \right)^{1/(m-1)}}, j = \overline{1, c}$

если $D_{ki} = 0 : \mu_{kj} = \begin{cases} 1, j = i \\ 0, j \neq i \end{cases}, j = \overline{1, c}$

Этап 6. Проверить условие $\left\| F - F^* \right\| < \varepsilon$, где F* - матрица нечеткого разбиения на предыдущей итерации алгоритма. Если «Да», то переход к этапу 7, иначе к этапу 3.

Этап 7. Конец.

Литература

1. *И. А. Чубукова* Data Mining. Учебное пособие. – М.: Интернет-Университет Информационных технологий; БИНОМ. Лаборатория знаний, 2006. – 382 с.: ил., табл. – (Серия «Основы информационных технологий»).

2. *Н. Паклин.* «Кластеризация категорийных данных: масштабируемый алгоритм CLOPE». Электронное издание.

Ссылка: ttp://www.basegroup.ru/clusterization/clope.htm

3. *Sudipto Guha, Rajeev Rastogi, Kyuseok Shim* «CURE: An Efficient Clustering Algorithm for Large Databases». Электронное издание.

4. *Tian Zhang, Raghu Ramakrishnan, Miron Livny* «BIRCH: An Efficient Data Clustering Method for Very Large Databases». Электронное издание.

5. *Daniel Fasulo* «An Analysis Of Recent Work on Clustering Algorithms». Электронное издание.

6. *Н. Паклин* «Алгоритмы кластеризации на службе Data Mining». Электронное издание.

Ссылка: http://www.basegroup.ru/clusterization/datamining.htm

7. *Jan Jantzen* «Neurofuzzy Modelling». Электронное издание.

Д.А.Никитченко

ИНТЕЛЛЕКТУАЛИЗАЦИЯ ДЕЯТЕЛЬНОСТИ МЕХАНИЧЕСКОГО ЗАВОДА

В последнее время в машиностроительной отрасли наблюдается существенная заинтересованность предприятия в повышении качества выпускаемой продукции, а также в сокращении сроков выполнения заказов уникальных изделий. Это связано в первую очередь с постоянно возрастающей конкуренцией производителей на рынке изделий с высокой трудоемкостью. С другой стороны, потенциальные заказчики в этой ситуации вправе сотрудничать с тем предприятием, которое наиболее полно удовлетворяет их требования, предъявляемые ими к качеству и цене приобретаемой продукции. Таким образом, для привлечения новых клиентов и установления долговременного сотрудничества с ними, необходимо постоянно повышать требования к производственному процессу с целью роста качества выпускаемой продукции.

Анализ деятельности отечественных производственных предприятий и всестороннее изучение их финансово-экономического положения позволяет сделать выводы о путях преодоления кризисов на предприятиях, а также о повышении устойчивости функционирования, как с экономической точки зрения, так и с точки зрения производственно-технической деятельности. При этом нельзя ограничиваться только мерами чисто экономического характера или же мерами по модернизации и реструктуризации материально-производственной инфраструктуры предприятия.

Информационная система промышленного предприятия

Автоматизированные системы управления предприятием (АСУП) необходимы, в первую очередь, для получения оперативной, достоверной информации о деятельности производственного предприятия с высокой степенью детализации, получения отчетов в различных разрезах, а также для снижения трудоёмкости получения информации по управленческому, налоговому и бухгалтерскому учёту. Также они позволяют существенно улучшить производственный учет и более рационально

организовать производство путем упорядочивания и дальнейшей оптимизации технологических маршрутов и производственных процессов, а также улучшения структуры изделия и расчета его состава для сокращения сроков комплектования заказов.

Таким образом, автоматизированная система управления предприятием (АСУП) является одной из важнейших составляющих для поддержания конкурентоспособности предприятия и удержания его от глубокого кризиса на этапе его внутренней реструктуризации, в том числе и при изменении производственной деятельности. Следует помнить, что снижение, даже на небольшую долю, затрат на обработку информации в административно-управленческой деятельности ведет к значительной экономии трудовых ресурсов [1].

На современном производственном предприятии без всестороннего применения компьютерных информационных систем затруднительны или практически невозможны все составляющие его деятельности, такие как исследование рынка, совершенствование и повышение качества изготавливаемой продукции, её конструкторское и производственное сопровождение, управление бизнес-процессами, мониторинг финансовой деятельности, периодическая реструктуризация предприятия и модернизация производственного процесса (для адаптации к постоянно изменяющимся условиям функционирования и повышения качества выпускаемой продукции для повышения уровня её конкурентоспособности) и эффективное управление работой предприятия.

Время первых радостей от использования разрозненных офисных программ (текстовых редакторов, электронных таблиц и др.) и бухгалтерии необратимо уходит. Современные производства поддерживают свою конкурентоспособность, как при регулярном производстве, так и при периодической реструктуризации — многофункциональными интегрированными компьютерными системами. Руководители отечественных промышленных предприятий как можно скорее должны придти к осознанию непродуктивности ресурсной и организационной дискриминации вопросов развития информационных систем их предприятий [2].

В условиях полной хозяйственной и технической самостоятельности большинства отечественных предприятий, ограниченных возможностей их финансовой поддержки, при отсутствии у них самих значительных ресурсов, — заметной поддержкой для них могло бы стать принятие на себя надпроизводственными отечественными структурами (в т.ч. и уполномоченными государством — т.е. министерствами и ведомствами) вопросов, носящих универсальный, инвариантный характер. Автомати-

зированная система управления предприятием жизненно необходима для сокращения издержек и повышения мощи производства и качества выпускаемой продукции.

Оценивая состояние дел в области успешного внедрения и дальнейшей эксплуатации автоматизированных систем на производственных предприятиях, необходимо, прежде всего, непредвзято осмыслить сложившуюся ситуацию, предстоящие задачи, после чего сформировать ясную позицию по поводу:

- зависимости экономических результатов работы промышленного предприятия от развития его информационной системы;
- существенного отставания отечественных предприятий в этой области;
- необходимости осуществления реструктуризации информационных систем предприятия;
- решения задачи финансового обеспечения этой реструктуризации.

В соответствии с доводами, приводимыми в [2], промышленным предприятиям России при выходе на открытый рынок товаров и услуг придется выдерживать жесткую конкуренцию по меркам международного рынка. И, заняв ту или иную нишу на рынке, удерживать ее придется в режиме, соответствующем общепризнанным и общепринятым темпам и стилю деятельности на этом рынке.

Это однозначно предопределяет необходимость введения конструирования продукции в режиме параллельного инжиниринга (с одновременной подготовкой производства и обратной связью на процесс конструирования), широкого использования средств быстрого прототипирования, кинематического и технологического имитационного моделирования, компьютерного дизайна, автоматизации технологической и технической подготовки производства, вплоть до разработки управляющих программ – см. рис. 1.

На рис. 1 представлены следующие сокращенния:

ИЭА – инженерно-электронный архив,

ОЭА – офисный электронный архив,

CAD (Computer Aided Design) – блок автоматизации конструирования и моделирования,

CAM (Computer Aided Manufactoring) – блок автоматизации технологии изготовления,

CAE (Computer Aided Engineering) – блок автоматизации инженерных расчётов,

ERP (Enterprise Resource Planning) – блок управления ресурсами предприятия,

БД – база данных,

BOM (Bill of material) – список материалов (состав изделия, спецификация, перечень сборочных узлов и компонент),

STEP (Standard for the Exchange of Product model data) — формат представления данных об изделии для обмена в виде информационной модели,

CALS (Continuos Asqusition and Lifecycle Support) – непрерывная информационная поддержка жизненного цикла продукта,

PDM (Product Data Management) – управление данными об изделиях,

OLAP (Online Analytical Processing) — оперативный анализ данных.

ОС – операционная система,

Сетевое ПО – сетевое программное обеспечение,

АиЭП – административная и эксплуатационная поддержка,

СУБД – система управления базами данных.

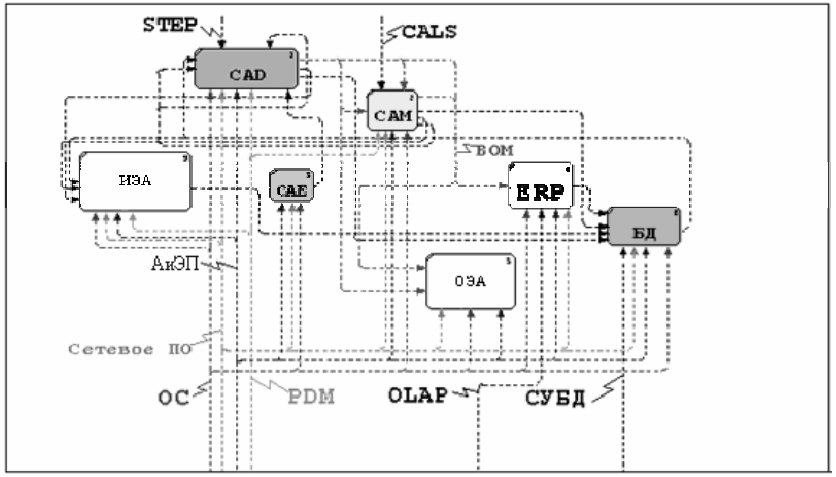


Рис. 1. Структура компьютеризированного сектора информационной системы промышленного предприятия

Автоматизация производства и система управления качеством

При внедрении на предприятии автоматизированной системы приходится сталкиваться с целым комплексом проблем взаимодействия. К

примеру, как организовать взаимодействие конструкторов, технологов, производственников и менеджеров по продажам. По сути, они выполняют совершенно разные функции в производственно-финансовом цикле деятельности предприятия. Но их общей задачей является устойчивое и эффективное развитие существующей производственной базы и нового гибкого производства с максимальной ориентацией на потребителя с целью максимального удовлетворения требований клиента путем предоставления ему высококачественной продукции и полного комплекса качественных услуг, включая гарантийное и сервисное обслуживание. В свою очередь, при автоматизации деятельности предприятия необходимо решить проблему технологической и конструкторской подготовки производства и их связи с оперативным управлением производственным процессом.

Для решения этих задач на производственных предприятиях необходимо внедрять системы управления качеством на основе ISO 9001:2000 и постоянно её развивать. Они позволят, в частности, снизить количество брака и возврата продукции. Построение и развитие на предприятиях таких систем взаимосвязано с внедрением корпоративных информационных систем (КИС) на основе информационных технологий (ИТ). АСУП, построенная на наиболее подходящей для данного предприятия платформе КИС, позволит устранить устарелые и реализовать наиболее современные и эффективные методы организации производственного цикла, а также поспособствует развитию командной формы управления как инструмента укрепления и совершенствования предприятия.

Вместе с тем необходимо разделять понятия внедрения ИТ и сертификации систем качества. Хотя и система управления предприятием, базирующаяся на стандартах ERP, и сертифицированная система качества затрагивают весь спектр вопросов деятельности предприятия, суть их различна. Первая охватывает более широкий комплекс вопросов, вторая рассматривается как часть первой. Можно сказать, что внедрение систем класса ERP – более дорогой и болезненный, но и более фундаментальный путь к кардинальному решению проблемы управления предприятием вообще и управлением качеством в частности. Несомненно, что эти два процесса должны быть тесно увязаны между собой, но наибольший эффект достигается, если они идут параллельно [3].

Для более полного представления о стандартах ERP рассмотрим историю их появления и развития.

Исходным стандартом, разработанным еще в конце 60-х годов прошлого века, был MRP, включающий только планирование материалов для производства. Расширенный до MRPII, этот стандарт обеспечивал

планирование всех производственных ресурсов предприятия (сырья, материалов, оборудования и т. д.). Дальнейшим развитием стала концепция ERP, которая позволяла объединить все ресурсы предприятия, добавив управление заказами, финансами и т. д. Сейчас практически все производственные системы, представленные на рынке, отвечают требованиям стандарта ERP [4]. Наконец, самый последний по времени стандарт CSRP охватывает также и взаимодействие с клиентами: оформление наряд-заказа, технического задания, поддержку заказчиков на местах и пр. Таким образом, CSRP вышел за рамки отдельного предприятия [5].

Фундаментальными направлениями развития современного машиностроения является создание компьютеризированных интегрированных производственных систем (Computer Integrated Manufacturing System — CIMS) и ориентированная на рынок поддержка полного жизненного цикла изделий на базе CALS-технологий.

Важнейшей составной частью CIMS является компьютеризированная система обеспечения и/или управления качеством (Computer Aided Quality — CAQ) производственных процессов на всех этапах жизненного цикла изделия. Содержанием системы CAQ являются разнообразные измерительные процессы, достоверность которых определяет качество как самой CIMS, так и создаваемого в ней изделия.

Рассмотрим построения CAQ на примере только одного этапа жизненного цикла изделия — технологического процесса изготовления машины. Как известно, любая технология реализуется в результате сложного взаимодействия трех основных продуктообразующих потоков: 1) материального, 2) энергетического, 3) информационного. Оставляя в стороне содержание первых двух потоков в машиностроительных технологиях, рассмотрим более подробно структуру третьего — информационного потока. Информационный поток, связанный с управлением, то есть организацией целесообразных воздействий, необходимо дополнить еще одним потоком, который несет информацию о текущем состоянии самого технологического процесса и создаваемого изделия, а также, в общем случае, — о состоянии материального потока (например, входной контроль), энергетического потока, системы управления, в частности, о ходе исполнения технологической и манипуляционной программ и т.д. [6]

Таким образом, в каждом технологическом процессе участвует еще одна информационная система — наблюдатель. Функции этой системы кардинально отличны от функций управления. Наблюдатель должен осуществлять сбор информации о показателях, характеризующих со-

стояние технологического процесса и создаваемого в нем изделия, а также достоверную и достаточно полную модель состояния технологической системы. На основании оценивания полученной модели состояния и сравнения ее оценок с желаемыми значениями наблюдатель формирует решение об управлении. Естественно, оно должно быть оптимальным.

Следовательно, с информационной точки зрения любой технологический процесс имеет не одну, а две стороны: управления и наблюдения. Под наблюдением (the observation) понимается процесс системной динамики объектов. Управление и наблюдение есть две стороны общего информационного потока в технологическом процессе.

Система CAQ, интегрируемая в CIMS, есть обобщенный автоматизированный наблюдатель производственной системы. Структура CAQ определяется действующим в CIMS технологическим сценарием и показателями качества процессов и производства. В общем случае, CAQ представляет собой сложную многоуровневую измерительно-вычислительную сеть, в которой осуществляется разнообразное перемещение измерительной информации.

Интеграция технологических процессов в единой производственной системе приводит к необходимости интеграции разнообразных измерительных процессов и процедур в едином CAQ-наблюдателе. Получаемые на разных уровнях наблюдения, в различном сочетании, одновременно или с разнесением во времени и/или в пространстве, вне (off-line) или внутри (in-line) технологического оборудования, измерительные процессы в CAQ-наблюдателе могут иметь различное алгоритмическое содержание. В качестве таковых могут быть: измерение, контроль, диагностика, обнаружение событий, идентификация, распознавание образов. Рассматривается соотношение между уровнями управления и наблюдения в гибком производственном модуле и ЧПУ.

В более общем случае CAQ-наблюдатель должен включать подсистему метрологического обеспечения или поддержки. В этой подсистеме аккумулируются результаты аттестации и сертификации технологического оборудования, поверки, учета и выбора средств измерений (СИ), другие функциональные возможности.

Построение оптимального CAQ-наблюдателя CIMS требует разработки, как минимум, нескольких проблем моделирования: 1) создание достоверных моделей состояния конкретных технологических процессов и производств, полнота которых достаточна для обеспечения заданного качества; 2) разработка стратегии наблюдения за параметрами технологического процесса и участвующих в нем потоков, достаточной для

требуемой полноты моделей состояния; 3) разработка стратегии оценивания наблюдаемых параметров; 4) разработка стратегии выработки решения на оптимальное управление.

Все перечисленные проблемы в той или иной степени сталкиваются с априорной неопределенностью информации о моделируемых объектах. Это обстоятельство существенно уже на этапе формализации моделей конкретных технологических процессов, видов оборудования, материальных потоков, средств измерения, контроля и т.п. Это касается объемных размерных цепей, нормирования точности деталей, узлов, сборок, процессов физико-механического взаимодействия режущего инструмента и деталей, неопределенности вероятностных характеристик, процессов и объектов, изменчивости характеристик в статике и динамике вследствие износа, старения, энергетических воздействий в процессе обработки, влияния окружающей среды и других факторов. Можно утверждать, что, чем выше требования к качеству процессов и производств (точности, надежности и др.) и чем сложнее структура объектов CIMS, охватываемых CAQ-наблюдателем, тем выше степень априорной неопределенности в оценке параметров моделируемых объектов. Кроме того, гибкость, заложенная в CIMS, приводит к изменчивости характеристик, параметров и структуры в процессе самого функционирования.

Интеллектуализация инженерно–конструкторской и технологической деятельности на примере крупного производственного предприятия – механического завода

Автоматизированная система управления предприятием на базе стандартов ERP, обеспечивая весь необходимый набор информационных услуг, от планирования до управления складами и производством, от закупок сырья, материалов, комплектующих, до сбыта готовой продукции, включая финансы и бухгалтерию, должна опираться на абсолютно точные данные.

Чтобы информацию можно было превращать в данные и наоборот, нужна система, способ представления информации или, иначе говоря, язык. Такая система далеко не однозначна. Поэтому одна и та же информация может быть представлена различными данными и, наоборот, одни и те же данные могут изображать различную информацию. Более того, данные могут и не изображать никакой информации, если они не удовлетворяют требованиям никакой системы представления данных, или же информация не может быть извлечена, если система представления нам неизвестна. На систему представления информации влияют как

содержание информации, так и цель, ради которой эта информация сохраняется, а также условия хранения информации [1]. При этом возникает проблема формализации информации [7]. Важную роль в функционировании автоматизированных систем играет системный аналитик [8].

Информационная платформа КИС в реальных условиях представляет собой сложную иерархическую структуру, в которой в первую очередь должны быть выделены источники первичной информации, впоследствии формирующие единую интегрированную базу данных: одну таблицу изделий, один файл поставщиков, одну таблицу маршрутов и ресурсов и т. д.

Единственным способом получения относительно полной информации о состоянии дел на предприятии является интеграция разрозненных информационных ресурсов для получения единой, сводной информационной базы данных. Последняя должна содержать необходимую для анализа информацию, свободную от неиспользуемых данных и противоречий [9].

Таким образом, точность данных в подобных системах наряду с технологическими процедурами контроля и обработки информации обеспечивается определением единого информационного пространства, базирующимся на едином источнике данных, для которого назначены ответственные за эти данные. Наличие только одного источника каждого вида данных значительно повышает их корректность, так как в этом случае они будут вводиться в систему только один раз, и при изменении этих данных все пользователи будут иметь доступ к актуальным на текущий момент данным, отражающим текущее состояние дел на предприятии.

С одной стороны, определить, какие подразделения станут источниками данных, нетрудно: например, конструкторские спецификации введет конструкторский отдел, занимающийся проектированием этих изделий, а технологический маршрут — технологические службы; с другой стороны, в реальных условиях это требует значительных усилий и затрат.

Рассмотрим более детально вопросы, связанные с формированием отдельных информационных слоев, на примере организации работ на конкретном предприятии.

Продукция на предприятии изготавливается по конструкторской документации, как собственной разработки, так и сторонних организаций. Характер учета позаказный. Оплата труда основных производственных рабочих — сдельно-премиальная. Основной оплатный документ — рабочий наряд на выполненную работу.

Производство носит единичный и мелкосерийный характер. Сложность продукции – порядка 2200…2500 детале-сборочных единиц (ДСЕ).

Традиционно запуск изделий в производство начинался с составления ведомости материалов (ВМ), которая до автоматизации включала следующие реквизиты: обозначение, наименование и количество ДСЕ, норма расхода необходимого материала, размер заготовки, количество деталей из заготовки, маршрут изготовления. В ВМ включалась также информация в виде отдельных технологических требований, как на процесс получения заготовки, так и на саму деталь. После этого разрабатывались (или уточнялись для уже изготавливаемых ДСЕ) технологические процессы, проектировались или заказывались недостающие средства технологического оснащения (СТО), которые должны быть изготовлены к моменту поступления заготовок на механообрабатывающие участки.

Недостатки традиционных, «ручных» методов обработки информации можно видеть хотя бы на таком примере. Если в качестве виртуального «пассажира» информационных потоков, которые являются основой процессов управления, взять конструкторскую спецификацию (КС), то станет ясно, что она, видоизменяясь, составляет основу ВМ, ведомости норм времени, ведомости СТО и т д. Учитывая текущие изменения КС, поддерживать в актуальном состоянии вышеперечисленные документы с одинаковой достоверностью практически невозможно.

На предприятии работы по внедрению автоматизированных способов обработки инженерно – конструкторской и технологической документации велись в следующем направлении: описание конструкторского состава изделий, составление материальных нормативов, маршрута изготовления (формирует отдел Главного технолога — ОГТ) и пооперационно-трудовых нормативов (формирует отдел охраны труда и занятости — ООТиЗ).

В качестве первоочередной была определена задача по автоматизации формирования ВМ. Для решения этой задачи было разработано программное обеспечение по созданию и ведению баз данных конструкторских спецификаций и картотеки деталей и сборочных единиц, которые ведутся в отделе Главного технолога. При этом была решена проблема разузлования изделия, т.е. на основании базы спецификаций и картотеки возможно построение полного состава изделия. Кроме того, на основе этих баз решаются задачи применяемости ДСЕ и задача входимости материала в изделие. Был выполнен большой объем работ по организации и ведению массива конструкторских спецификаций.

В Бюро программирования были разработаны общезаводские классификаторы на стандартные изделия, материалы и ПКИ (покупные комплектующие изделия) и программное обеспечение для ведения этих баз.

Подготовленные таким образом базы данных в основном обеспечивают формирование ВМ, сводной материальной спецификации (СМС), выписок из ВМ по цехам, сводной материальной спецификации по цехам.

Далее стала решаться задача определения трудоемкости изделия. С этой целью в ООТиЗ была создана база пооперационных нормативов трудоемкостей по ДСЕ, содержащимся в базе картотеки, и база тарифов. В результате, на основании базы составов изделий, сформированной в ОГТ, и баз, сформированных в ООТиЗ, производится расчет трудоемкости любой сборочной единицы.

Структура информационных связей задач автоматизации формирования ВМ и определения трудоемкости изделия представлена на рис. 2.

Дальнейшая интеллектуализация деятельности предприятия была связана с внедрением на предприятии корпоративной интегрированной автоматизированной системы управления предприятием. С внедрением этой системы удалось решить задачу управления производством, которая разбивается на независимые во времени, но неразрывно связанные между собой этапы:

1) Формирование портфеля заказов на месяц;

2) Формирование месячных подетальных планов (ППЗ);

3) Учет и движение детале-сборочных единиц;

4) Анализ затрат.

Очередность выполнения этих этапов устанавливалась с учетом того, в какой последовательности могут быть созданы информационные слои, образующие единую платформу.

Для успешного внедрения системы управления производством был проведен ряд организационно-технических мероприятий:

1) Подключение к общезаводской вычислительной сети подразделений инженерных служб;

2) Обеспечение подразделений инженерных служб средствами вычислительной техники в необходимом объеме;

3) Создание регламентов работы с системой для каждой инженерной службы;

4) Разработка программного обеспечения для формирования месячного портфеля заказов;

5) Формирование конструкторской базы знаний, позволяющей на основе месячного портфеля заказов создать подетальное ППЗ по всем заказам;

6) Интеграция АСУП с существующими базами данных.

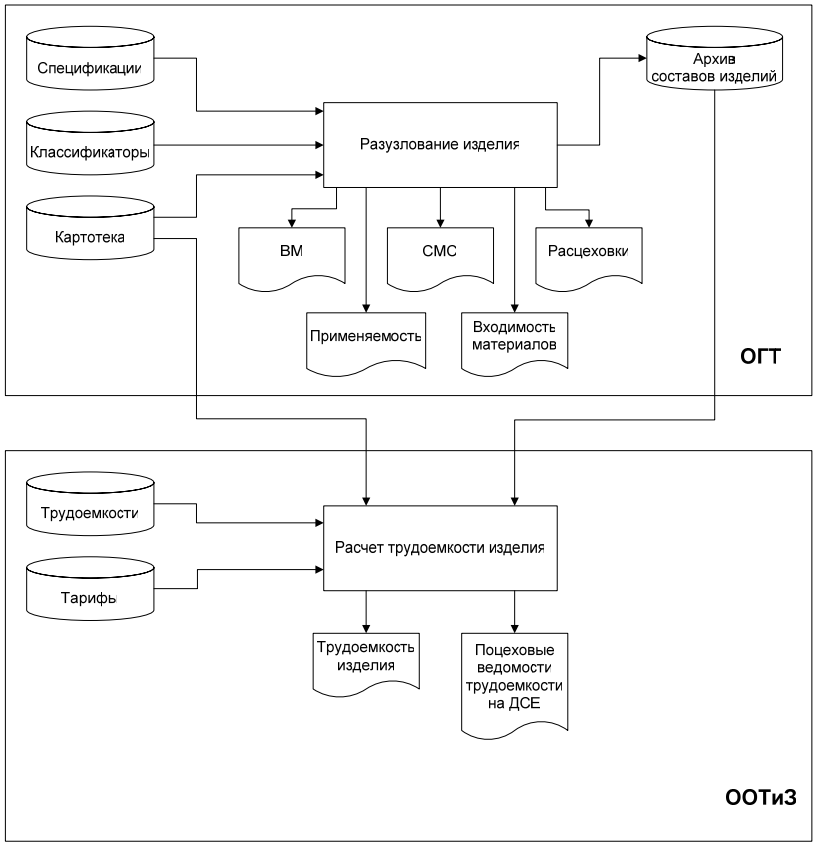


Рис. 2. Структура информационных связей задач автоматизации формирования ВМ
и определения трудоемкости изделия

Таким образом, при интеллектуализации производственного предприятия были внедрены стандарты ERP, которые стали эффективным инструментом развития производства и повышения качества выпускае-

мой продукции. Важным условием успешного внедрения корпоративной системы является наличие у предприятия корпоративных стандартов учета и управления и готовность к внедрению ролевого принципа построения системы, то есть к привязке каждой функции системы к некоторой роли пользователя. Но успешная работа АСУП на предприятии связана, в свою очередь, с квалификацией персонала, который имеет доступ к системе и непосредственно влияет на её функциональность и работу. Следовательно, требования к знаниям и навыкам пользователей АСУП должны предъявляться в соответствии с функциональностью системы и местом в ней каждой роли пользователя, и определяться регламентом работы подразделений производственного предприятия, который должен быть составлен для работников каждого подразделения и функционал которого реализован в системе.

Сегодня заметно некоторое оживление на рынке промышленной автоматики. Отчасти это объясняется крайней степенью износа доперестроечных систем, отчасти — оживлением некоторых отраслей промышленности. Основная же проблема банальна — низкий уровень платежеспособности предприятий, подлежащих автоматизации. Из-за этого они часто применяют дешевое оборудование с невысокой надежностью, в результате чего страдает качество системы автоматизации. С другой стороны, закупка комплексного решения по автоматизации процесса или объекта у западной фирмы-интегратора, значительно удорожает проект.

Несмотря ни на что отечественная промышленность развивается, и мы это ощущаем в своей работе. Проблемы есть, но они, в основном, носят финансовый характер. Многие заказчики желают "автоматизироваться", произвести обновление существующих технических средств или создать новые системы, но их сдерживает отсутствие или недостаток средств.

Зачастую, решение руководства предприятия является определяющим при выборе той или иной стратегии развития. Так как руководителями могут быть люди часто достаточно далекие от вопросов информационных технологий, то велика вероятность ошибочно принятого решения, безуспешность проекта и необоснованно нерационально потраченных на проект денежных средств. Поэтому при выборе платформы и методики внедрения АСУП необходимо всесторонне учитывать мнения не только службы информационных технологий предприятия, но и привлекать сторонних консультантов и независимых экспертов, а также учитывать опыт как успешных, так и не оправдавших себя внедрений АСУП в своей отрасли производства.

Литература

1. Электронные вычислительные машины: В 8-ми кн. Кн.6. Средства общения с ЭВМ: Практ. пособие для вузов / А.В.Савченко, Ю.В.Сальников, А.Н.Филиппов; Под ред. А.Я.Савельева. — 2-е изд., перераб. и доп. — М.: Высш. шк., 1991. — 127 с.: ил.

2. *Дубейковский В.И.* Компьютеризация информационной системы промышленного предприятия. — сайт компании Interface Ltd., http://www.interface.ru/misc/rmd/komp.htm

3. *Хачатуров А.Е., Куликов Ю.А.* Основы менеджмента качества. — М.: ДиС, 2003. С.134–170.

4. *Верников Г.И.* Основы систем класса MRP-MRPII. — сайт «Корпоративный менеджмент», http://www.cfin.ru/vernikov/mrp/mrpmine.htm

5. *Верников Г.И.* Планирование ресурсов, синхронизированное с покупателем (CSRP). — сайт «Корпоративный менеджмент», http://www.cfin.ru/vernikov/mrp/csrp.shtml

6. *Телешевский В.И.* К проблеме интеллектуализации измерительных процессов в производственных системах. — Site of Information Technologies http://inftech.webservis.ru/it/conference/scm/1999/session7/telechevskiy.htm.

7. *Белоногов Г.Г., .Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации.—М.: Наука, Главная редакция физико-математической литературы, 1979, 256 стр.

8. Перспективы развития вычислительной техники: В 11 кн.: Справ. Пособие / Под ред. Ю.М. Смирнова. Кн. 2. Интеллектуализация ЭВМ / Е.С. Кузин, А.И. Ройтман, И.Б. Фоминых, Г.К. Хахалин. — М.: Высш. шк., 1989. — 159 с.: ил.

9. *Дзегеленок И.И., Сорокин П.М.* Технология актуализации знаний на основе интеграции сетевых информационных ресурсов — ВС/NW 2003г., №1(3)/17.1

М.А. Павлова

ИНТЕЛЛЕКТУАЛЬНЫЕ МЕТОДЫ И ПОДХОДЫ К ОРГАНИЗАЦИИ ПОИСКА ИНФОРМАЦИИ

Введение

В сфере электронных информационных массивов, увеличивающихся с каждым годом, громадную роль играет документальный поиск необходимой информации. Объемы этой информации предполагают увеличение ее рассеянности, что обостряет актуальность проблемы организации эффективного поиска. Все сложнее и сложнее становиться осуществлять поиск в возрастающем своде электронных документов, предпринимались и предпринимаются попытки интеллектуализировать этот процесс. Интеллектуальный поиск, кроме выдачи найденных документов, позволяет учесть морфологию запроса, введенного на естественном языке, подправить орфографические ошибки в запросе, предусмотреть синонимичные варианты. Основным способом осуществления подбора синонимичных вариантов в процессе поиска является использование информационно-поискового тезауруса (ИПТ), что в результате способствует повышению полноты поиска.

В исследовательской работе поставлена задача разработать подобный метод поиска на базе тезауруса для специальной учебно-научной библиотеки из 25 книг, посвященных проблемам сурдопедагогики, логопедии, дактилологии, истории глухих и слабослышащих и физиологии органов речи и слуха. Основным предметом изучения данной работы является разработка математической модели и алгоритма выборки документов, отвечающих наиболее полно и релевантно заданному запросу. Результаты работы будут иметь большое значение в качестве базы знаний для преподавателей в области специальной педагогики, студентов педагогических вузов и других, в том числе глухих и слабослышащих.

Понятие тезауруса

Тезаурус представляет собой словарь лексических единиц дескрипторного *информационно-поискового языка* (ИПЯ) или нормативный

словарь *дескрипторов* и ключевых слов с зафиксированными парадигматическими отношениями между этими единицами, предназначенный для координатного *индексирования* документов и информационных запросов. Разработка любого тезауруса начинается с построения множества ключевых слов, на базе которых выстраиваются связи различного типа (синонимия, род-вид, часть-целое и др.) с другими словами. Таким образом, тезаурус представляет собой не что иное, как семантическую сеть или если говорить упрощенно список терминов, их синонимов и связей. Ключевое слово представляет собой отдельное слово или словосочетание естественного языка, выделяемое из текста документа или запроса и несущее существенную смысловую нагрузку с точки зрения информационного поиска. Оно отражает основное содержание документа при индексировании.

Термин *"тезаурус"* древнего происхождения. Впервые его применил еще в 13 веке Брунетто Латини в своем произведении "Книга о сокровище". Первые тезаурусы составлялись без всякой связи с потребностями информационной деятельности. Наибольшую известность получил тезаурус, составленный в 1852 англичанином Роджетом "для облегчения выражения мыслей и помощи при написании сочинений". В 1956 он был использован Кембриджской группой в работах по автоматизированному переводу. Тезаурус – своего рода "обращенный" толковый словарь: если в обычном толковом словаре по слову находится его значение, то в тезаурусе по значению, записанному определенным способом, находят одно или несколько слов, выражающих искомое значение.

Информационно-поисковые тезаурусы используются в *информационно-поисковых системах (ИПС)* как средство перевода с естественного языка на информационно-поисковый язык при индексировании документов и запросов, он обеспечивает формализованное представление смыслового содержания документов в АИС. ИПТ позволяют установить взаимосвязи между тремя терминологическими областям: авторской терминологией, терминологией ИПС и терминологией пользователя при формировании его запросов.

Информационно-поисковые тезаурусы, как правило, строятся для отдельной области знания. Стремление построить универсальный тезаурус ухудшает его использование в области автоматизированного поиска, и делает невозможным применение его в библиотечных системах. Построение универсального тезауруса требует длительной согласованной работы огромного коллектива специалистов в разных областях знания.

Принципы построения тезауруса

Построение тезауруса сочетает два метода: а) научную разработку классификационных схем понятий и б) выявление терминологии из представительного массива документов с последующим его дополнением терминами, взятыми из вспомогательных источников.

В качестве вспомогательных источников могут выступать: тезаурусы родственной тематики; терминологические, толковые, энциклопедические словари; научно-технические словари и справочники; таблицы ББК, УДК; тематические рубрикаторы; руководства, стандарты; задания на НИР и ОКР; патентные описания; рефераты статей и др.

Формирование тезауруса включает следующие этапы:

1) отбор терминов и составление словника (перечня терминов, выделяемых в качестве ключевых слов);

2) составление концептуальных моделей области деятельности;

3) обработка словника и дополнение его по каждому ключевому слову лексическими единицами из различных вспомогательных источников, перечисленных выше, или другими отобранными ключевых словами с последующим распределением по семантическим сферам.

Ключевые слова могут выбираться в соответствии с принятой методикой индексирования. Выделение ключевых слов, является очень важным этапом на начальных стадиях составления тезауруса, этот этап зависит от субъективных факторов и его качественное исполнение будет во многом определять процент выданных релевантных документов. Выбор каждого ключевого слова зависит от следующих факторов:

– важность данного ключевого слова для описания содержания документа с точки зрения информационного поиска;

– его связи с ключевыми словами, отобранными ранее;

– точность и приемлемость ключевого слова с точки зрения терминологии рассматриваемой области науки или техники;

– решения специалистов в соответствующей области знания.

Дополнение тезауруса новыми лексическими единицами, выявляемыми при индексировании документов и запросов, производится при эксплуатации системы постоянно. При небольшой частоте появления новых ключевых слов (порядка одного на 10 тыс. слов текста) считается, что информационно-поисковый тезаурус насыщен терминами по данной тематике.

Выделение ключевых слов из текста сопровождается их обработкой, которая может заключаться:

– в принятии решения о разделении выделенного словосочетания из

двух или более слов или сохранении его в качестве целостного ключевого слова;

– в принятии решения об использовании сложного слова в качестве ключевого или членении его на два или более ключевых слов;

– устранение омонимии и полисемии;

– приведение слов и словосочетаний к нормальной форме.

Распределение ключевых слов по семантическим сферам производится в соответствии с концептуальной схемой. Могут использоваться синонимичные, родовидовые отношения и отношение часть-целое, кроме них могут быть введены и другие. Родовидовое отношение подразумевает связь между двумя дескрипторами, одно из которых имеет объем понятия, включаемое в объем понятие другого дескриптора. В таких случаях тезаурус напоминает иерархическое дерево, по которому дескрипторы распределяются по более узким семантическим категориям.

Основные аспекты и понятия поиска информации

Существует распространенное мнение, что каждое новое поколение программных продуктов становиться более совершенным, то есть из этого убеждения выходит, что раньше все программы были «плохие», а сейчас очень интеллектуальны и более приспособлены под потребности пользователя. Что же все-таки поменялось за последние годы? Понятно, что за несколько лет развития информационных систем произошли изменения в алгоритмах, в структурах данных, в математических моделях, но самое главное – изменилась парадигма использования систем. Другими словами, потребителями электронной информации и систем стали вполне обыкновенные люди – домохозяйки, подростки и другие. Кроме тотальной востребованности поисковых систем стало ясно, что люди не только «думают словами», но и «ищут словами». В ответе системы они ожидают увидеть слово, набранное в запрос на требуемую информацию. Трудно переучить человека правильно искать информацию в поисковых системах, также как и переучить его говорить и писать. Не случайно в Интернете стали появляться статьи и советы о том, как правильно оформлять свои запросы, прокручивать возможные варианты. В 60-е – 80-е годы были предприняты попытки организовать итеративное уточнение запросов, генерацию связного ответа на вопрос, эти способы интеллектуализации поиска с трудом выдерживают сейчас испытание реальностью.

Рассмотрим основные понятия поискового механизма.

Перевод документа на информационно-поисковый язык представляет собой переход от такого документа к его сокращенному описанию

с целью использования этого описания для последующего поиска. По определению А.И. Михайлова, А.И. Черного, Р.С. Гиляревского, информационно-поисковый язык – это специализированный искусственный язык, предназначенный для выражения основного смыслового содержания документов или информационных запросов с целью отыскания в некотором множестве документов таких, которые отвечают на поставленный информационный запрос.

Описание содержания документа с помощью ИПЯ представляет собой поисковый образ документа, а описание содержания запроса – поисковый образ запроса. Правила составления поисковых образов документов и запросов являются правилами перевода текстов с естественного языка на ИПЯ. При наличии массива документов и соответствующих им поисковых образов поиск отвечающего на запрос документа сводится к сопоставлению поисковых образов документов и запросов. Для того чтобы оценить степень их соответствия, необходимо сформулировать критерий смыслового соответствия – формальное правило, по которому поисковые образы документа и запроса считаются совпадающими или несовпадающими. При формальном совпадении поисковых образов документа и поискового образа запроса документы считаются отвечающими на запрос. Однако такое совпадение не означает содержательного соответствия выданного документа запросу. Документ, смысловое содержание которого соответствует информационному запросу, называется *релевантным* этому запросу. Но если ИПЯ неточно выражает смысл документов и запросов, то может оказаться, что близкие по смыслу документы и запросы обладают разными поисковыми образами и, наоборот, у далеких по смыслу друг от друга документов поисковые образы оказываются сходными. В этом случае не все документы, формально соответствующие запросу, соответствуют ему в действительности, т.е. релевантны. Явление, при котором в ответ на запрос система выдает документы, не соответствующие запросу, называется *поисковым шумом*. По тем же причинам может оказаться, что часть документов, релевантных запросу, все же оказалась невыданной, тогда говорят о *потерях информации*.

Можно количественно оценить информационный шум и потери информации с помощью полноты и точности поиска. Данные показатели технической эффективности поисковой системы введены давно, но до сих пор они являются оценочными параметрами поиска.

Полнота поиска R определяется отношением числа выданных в результате поиска релевантных документов к общему числу релевантных документов (выданных и оставшихся невыданными):

R = a / (a + c) (1).

Точность поиска P представляет собой отношение количества выданных релевантных документов к общему числу документов в выдаче:

P = a / (a + b) (2).

Использованные обозначения в формулах (1), (2):

a – число релевантных документов в выдаче;

c – число релевантных документов, оставшихся невыданными (потери информации);

b – число выданных нерелевантных документов (поисковый шум).

Возникает вопрос: возможно ли разработать такой ИПЯ, который бы точно передавал смысл документа, и поиск бы оценивался максимальной полнотой и точностью? Попытка достичь таких показателей неизбежно подчиняется жизненному закону «приобретаем, теряя». Информация, содержащаяся в документах, объективно следует закону рассеяния. Это означает, что поисковая система в одном случае может выдать несколько публикаций, точно отвечающих на запрос, но не выдать релевантную информацию, рассеянную среди других источников, то есть, выданы не все имеющиеся релевантные документы, в другом – может выдать и релевантную информацию. Полнота поиска возрастет. Однако в этом случае будет иметь место больший поисковый шум. Исходя из этого, можно сделать вывод, что невозможно достичь одновременного достижения максимальных показателей полноты и точности поиска. Увеличивая полноту поиска, уменьшаем его точность и наоборот.

Перевод содержания документа на ИПЯ, то есть индексирование, зависит от субъективного фактора, так как этот процесс – ручная работа человека. В результате разные люди могут проиндексировать один и тот же документ по-разному. Поэтому очевидно, что неточность описания содержания документа, в качестве этого процесса может оказаться этап выделения из текста ключевых слов, не может не сказаться при его поиске. Таким образом, можно ответить на вышепоставленный вопрос: не может быть разработан такой ИПЯ, который бы гарантировал максимальные точность и полноту поиска. Данный анализ вопроса показывает, что не нужно стремиться к таким наивысшим показателям, а следует находить компромисс между двумя этими показателями. Именно от этого зависит качество работы всей информационно-поисковой системы. Также необходимо заметить, что большое внимание на ранних этапах разработки поисковых механизмов, нужно большое внимание уделять ИПЯ.

Поисковый образ документа формируется совокупностью терминов

тезауруса – дескрипторами. Точно так же на язык дескрипторов переводится и запрос. Таким образом, поиск сводится к выявлению дескрипторов, принадлежащих одновременно и поисковому образу документа и поисковому образу запроса, то есть к нахождению пересечения их множеств дескрипторов. Минимальная величина зоны пересечения оговаривается принятым критерием смыслового соответствия. Изменяя его, можно варьировать точность и полноту поиска в зависимости от нужд потребителей информации.

Требования к информационно-поисковому тезаурусу очень высоки. Должны быть зафиксированы некоторые отношения между терминами (род – вид, часть – целое и другие), служащие целям повышения точности и полноты поиска.

Лингвистические проблемы

Лексику тезаурусов составляют не только дескрипторы, но и их синонимы, которые не являются дескрипторами. Присутствие в тезаурусе синонимов имеет большое значение. Поясним это на примере.

Пусть имеется два термина из области «Специальная педагогика для глухих и слабослышащих»: «Нейросенсорная тугоухость» и «Глухота», данные понятия могут выступить как синонимы для людей, которые имеют и близкое, и отдаленное представление о данной области знания. Но одно понятие не может быть представлено в тезаурусе двумя различными терминами, потому что это привело бы при поиске документов по запросу, который содержит термин «Глухота» к следующей ситуации: поисковая система не выдала бы те документы, содержащие в поисковом образе термин «Нейросенсорная тугоухость», хотя они подлежат выдаче, так как вполне соответствуют запросу. Использование синонимичных терминов приводит к потерям информации. Чтобы предотвратить это, из двух синонимов в качестве дескриптора выбирают один – термин «Нейросенсорная тугоухость», а другой является отсылкой этого термина – «Глухота». Такая пометка означает, что вместо одного термина при составлении поисковых образов документов или запросов следует использовать другой, являющийся дескриптором. Именно так ликвидируется в тезаурусах синонимия.

Если из нескольких синонимов один выбран в качестве дескриптора, то остальные (в нашем случае это термин «Глухота») при этом получают название ключевых слов. Наличие в тезаурусе ключевых слов с отсылками к соответствующим дескрипторам облегчает индексирование документов, обеспечивает быстрый поиск нужного термина, способствует повышению качества функционирования ИПС.

Кроме того, существуют другие лингвистические проблемы:

1) исключение неинформативных слов в запросе (стоп-слов);

2) лемматизация: приведение словоизменительных форм к нормальной форме, то есть словарной;

3) решение проблем, возникающих при омонимии, в данном случае используется вероятностный подход;

4) принятие решений об использовании в качестве дескриптора словосочетания, при решении такой проблемы можно руководствоваться ГОСТом СИБИД 7-25-2000: «допускается включать словосочетания в словник, если в качестве опорного слова они содержат существительное и если выполнено одно из следующих условий:

– Значение словосочетания не выводится из значений его компонентов.

– Хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле.

– Для данного словосочетания в словнике ИПТ существуют полные синонимы.

– Данное словосочетание является устойчивым словосочетанием с именем собственным.

– Отдельные слова словосочетания имеют слишком широкое значение.

– Для данного словосочетания в словнике ИПТ существует общепринятая аббревиатура.

– Разбиение словосочетаний на отдельные компоненты приводит к потере важных для поиска семантических связей.

Словосочетания, которые не удовлетворяют перечисленным условиям, разбивают на компоненты.»

Доказано, что для некоторых языков лингвистические алгоритмы не вносят существенного прироста точности и полноты, например, в английском языке, но все же основная масса языков требует хотя бы минимального уровня лингвистической обработки. Почти все лингвистические проблемы могут решаться методами, которые опираются на словари (морфологические, синтаксические, семантические), созданные ранее.

Другие лингвистические задачи, решаемые современными системами, могут быть следующими: автоматическое определение языка документа; графематический анализ: выделение слов, границ предложений. Еще реже можно встретить алгоритмы статистического характера (LSI, нейронные сети), а толково-комбинаторные или семантические словари, также в крайне узких предметных областях.

Математические модели в поиске

Примерно половина поисковых систем и модулей функционируют без всяких математических моделей, то есть их разработчики не сопровождают разработку реализацией абстрактной, математически формализуемой моделью. Иногда даже не подозревая о ее существовании, просто стремятся сделать программу, в которой хоть как-то был бы организован поиск и выдавался какой-нибудь результат, а все остальные проблемы возлагались на пользователя.

Но в сегодняшних информационно-поисковых системах очень остро встает вопрос о повышении качества поиска, о потоке пользовательских запросов, при которых кроме эмпирически подобранных коэффициентов может оказаться полезным воспользоваться каким-нибудь пусть и несложным теоретическим аппаратом. Модель поиска – это некоторое упрощение реальности, на основании которого получается формула, позволяющая программе принять решение: какой документ считать найденным и как его ранжировать. После разработки такой модели те эмпирически подобранные коэффициенты приобретают физический смысл, их становится легче подбирать.

Различают следующие семейства моделей, используемые в информационном поиске:

– теоретико-множественные (булевская, нечеткие множества, расширенная булевская);

– алгебраические (векторная, обобщенная векторная, нейросетевая, латентно-семантическая);

– вероятностные.

Рассмотрим кратко простейшие из этих моделей.

Булевская модель. Модель низкого уровня, предполагается возможным два результата поиска: если слово из запроса в документе есть, то документ считается найденным, если нет – не найденным. Данную модель используют некоторые программисты, реализующие полнотекстовый поиск. Таким образом, эта модель предоставляет небольшой инструмент манипулирования данными. Эта модель подвергается жесткой критике, поскольку она грубовата и непригодна для ранжирования документов.

Векторная модель. Разработана в 1957 году Джойсом и Нидхэмом, которые для развития булевской модели предложили учитывать частотные характеристики слов, чтобы «... операция сравнения была бы отношением расстояния между векторами...». Данная модель была успешно реализована на практике в 1968 году Джерардом Солтоном в поисковой системе SMART (Salton's Magical Automatic Retriever of Text). Ранжиро-

вание в этой модели основано на естественном статистическом наблюдении — чем больше локальная частота термина в документе и больше «редкость» (т.е. обратная встречаемость в документах) термина в коллекции, тем выше вес данного документа по отношению к термину. Другое обозначение векторной модели TF*IDF.

Вероятностная модель. Данную модель разработали и обосновали в 1977 году Робертсон и Спарк-Джоунз. Релевантность в этой модели рассматривается как вероятность того, что данный документ может оказаться интересным пользователю. При этом подразумевается наличие уже существующего первоначального набора релевантных документов, выбранных пользователем или полученных автоматически при какомнибудь упрощенном предположении. Вероятность оказаться релевантным для каждого следующего документа рассчитывается на основании соотношения встречаемости терминов в релевантном наборе и в остальной, «нерелевантной» части коллекции. Хотя вероятностные модели обладают некоторым теоретическим преимуществом, ведь они располагают документы в порядке убывания "вероятности оказаться релевантным", на практике они так и не получили большого распространения.

Заметим, что в каждом из семейств простейшая модель исходит из предположения о взаимонезависимости слов и обладает простым условием фильтрации: документы, не содержащие слова запроса, никогда не бывают найденными. Продвинутые модели каждого из семейств не считают слова запроса взаимонезависимыми, а, кроме того, позволяют находить документы, не содержащие ни одного слова из запроса.

Применение тезаурусного метода поиска в Интернет

Тезаурусы находят сегодня ограниченное применение в универсальных полнотекстовых поисковых машинах Интернета. Одна из причин ограниченности состоит в том, что чрезвычайно трудно построить тезаурус, который соответствовал бы тематическому разнообразию информации, индексируемой универсальной машиной поиска. С другой стороны, полнота не является критическим параметром универсальных поисковых систем Интернета.

Короткие запросы – характерная черта интернет-поиска. Средняя длина запроса к поисковой системе Яндекс за последнюю неделю февраля 2004 года составила 2,81 слова. Если сравнить этот показатель с данными 1997 и 1999 годов (1,2 и 2,7 слова соответственно), то можно предположить, что рост средней длины запроса замедлился. Короткие запросы, увеличение объема и разнообразия информации в Сети, а также тот факт, что большинство пользователей анализируют только одну

две страницы результатов поиска, заставляют разработчиков уделять большое внимание механизмам ранжирования результатов.

Приведем реальные примеры Интернет-механизмов ранжирования:

– DirectHit. Один из первых механизмов, который отслеживал выбор пользователей и учитывал «мнение большинства» при формировании отклика на одинаковые запросы;

– поисковый сервис Eurekster, запущен в начале 2004 года. Способен ранжировать результаты поисковой машин в AlltheWeb в соответствии с предпочтениями участников определенного сообщества;

– механизм PageRank и аналогичные подходы, основанные на анализе ссылочной структуры Web, позволили существенно повысить качество откликов поисковой системы в ответ на короткие запросы;

– поисковая машина Rambler, которая при ранжировании результатов учитывает популярность ресурсов в рейтинге Rambler Top100.

Тезаурусы, используемые в Интернет поисковых системах, могут быть построены автоматически на основе анализа совместной встречаемости слов, а также вручную. Построенные вручную тезаурусы обычно дают хорошие результаты в различных приложениях. Можно привести примеры: в конце 90-х годов прошлого века поисковая машина AltaVista предоставляла сервис AltaVista Refine, который позволял устранять неоднозначность терминов запроса с помощью словаря совместной встречаемости слов. В настоящее время Google предлагает поиск по синонимам и грамматическим вариантам для ограниченного набора английских слов. Так, в ответ на запрос *'~cats'* будут найдены документы, содержащие слова *'cat'*, *'dogs'*, *'pets'* и *'kitten'*.

Применение тезауруса как рубрикатора

Существует ряд тезаурусов, основная задача которых не индексация ресурсов, а их классификация. В этом случае основными объектами таких тезаурусов выступают не термины, а понятия (рубрики), и, часто, идентифицирующие их уникальные идентификаторы (коды классификации). Отношения в таком тезаурусе – не семантические связи между терминами, а характеризующие логику описываемой предметной области отношения между понятиями (рубриками). Примерами таких тезаурусов могут служить тематические классификаторы в разных отраслях науки, например, MSC, PACS, DDC.

Структура классификатора соответствует структуре обычного тезауруса, поскольку связи между его рубриками по смыслу те же, что и между терминами тезауруса, и классификатор является его частным

случаем. Однако при классификации в соответствие ресурсам ставятся не термины, а обозначаемые ими понятия. Потому в схеме данных информационной системы понятия тезауруса должны быть выделены в самостоятельные объекты. Это означает, что такая схема должна иметь структуру, отличную от вышеописанных стандартов, в которых понятия не выступают отдельными объектами, а есть лишь термины и связи между ними. В то же время, схема должна позволять работать с тезаурусами, описанными в соответствии с этими стандартами, т.е. быть совместима с ними.

Применение в электронных библиотеках

Если библиотека собрана по одной тематике, например как в моей исследовательской работе используется библиотека, представляющая собой набор из 25 книг, посвященных проблемам сурдопедагогики, логопедии, дактилологии, истории глухих и слабослышащих и физиологии органов речи и слуха, то поиск на основе тезауруса будет очень эффективным в таком случае. При разработке подобных систем ИПТ строится вручную на основе анализа текстов по выделению ключевых слов. Но, к сожалению, это достаточно длительный процесс, поэтому широкого применения в реальной жизни не наблюдается, по крайней мере, в России кроме близких к этому аналогов, кроме УИС Россия, не найдены. Для более широкого распространения таких разработок следует привлекать группу специалистов для качественного составления тезауруса. Если говорить о зарубежных электронных библиотеках с использованием смыслового поиска, то нужно отметить Национальную медицинскую библиотеку США.

В 60-е – 80-е годы предпринимались попытки создать мощное лингвистическое обеспечение в электронных библиотеках. Приведем примеры.

ГАСНТИ. В его реализации участвовали наиболее квалифицированные специалисты бывшего СССР. Проектирование было начато с 1965 г. и в 1969 г. появился первый системный проект под названием “Комплекс средств индексирования научно-технической информации”. В течение 20 лет шли разработки по широкому классу проблем, связанных с лингвистическим обеспечением. В результате к концу 1980-х гг. в ГАСНТИ была сформирована достаточно стройная система языковых средств по всему необходимому спектру функциональных задач АСНТИ, поддержанная развитой системой государственных стандартов и специализированной организационной структурой. Можно утверждать, что созданное лингвистическое обеспечение ГАСНТИ соответствовало

наиболее высокому уровню информационной науки того времени, в его состав входило до 200 тезаурусов и рубрикаторов по всем отраслям народного хозяйства. Кризис 1990-х гг. в системе НТИ России совпал со сменой поколений АС НТИ (сначала распространение ПЭВМ, затем Интернет), что в совокупности привело к почти полной утрате достижений того времени. В настоящее время поддерживается ГРНТИ и частично УДК, по дескрипторным языкам и языкам метаданных системная работа не ведется.

ЕСКК ТЭИ (Единая система классификации и кодирования технико-экономической информации). Чисто научный уровень этих разработок был несколько ниже, чем в ГАСНТИ, зато масштабы работ гораздо больше. В результате была создана система общероссийских классификаторов, число которых к концу 1980–х гг. достигло 35, а их общий объем превысил 3 млн. позиций. Кризис 1990-х гг. также почти полностью разрушил эту систему.

Обзор тезаурусных систем

В России к настоящему времени создано несколько универсальных тезаурусов для задач информационного поиска.

Тезаурус Медиалингва. Компания Медиалингва предлагает разработчикам информационно-поисковых систем тезаурус для расширения поискового запроса (150 000 входов). Тезаурус содержит словари синонимов, антонимов, родственных слов и родовидовых связей.

Тезаурус Гарант-Парк-Интернет. Тезаурус присутствует и в семействе продуктов компании Гарант-Парк-Интернет. По информации разработчиков, в состав тезауруса вошло около 75 тыс. слов и словосочетаний, объединенных в 22 тыс. гипонимических рядов, в том числе 17 тыс. синонимических рядов, охватывающих 45 тыс. слов.

Тезаурус УИС Россия. В университетской информационной системе Россия (УИС Россия) используется общественно-политический тезаурус, который представляет собой иерархическую сеть более 42 тыс. понятий (более 95 тыс. русских слов и выражений). Тезаурус используется как для расширения запросов, так и для тематического индексирования документов.

Среди зарубежных систем существуют также популярные тезаурусы.

Roget's. Наиболее популярный тезаурус, который организован вниз вплоть до набора синонимов. Поэтому он часто используется для того, чтобы подыскать более подходящий синоним к слову, и дополнен грамматическими сведениями в каждой своей статье.

DUDEN. Представляет собой книгу с картинками на правой стороне (по разным ПО) с пронумерованными графически мельчайшими их деталями. На правой стороне этот нумерованный список сопровождается названиями (даже на двух языках). Например, на целой странице нарисованы железнодорожная техника, станции, пути и т.п. Справа можно найти названия стрелок, семафоров, костылей.

SNOMED. Это огромный компьютеризированный тезаурус медицинской терминологии. Данный тезаурус используется в мощной поисковой службе медицинской информации Medical World Search, в первую очередь для профессионалов – научных работников и врачей-практиков. Основные характеристики, отличающие MWS от других поисковых служб – это использование поисковой машины, которая обрабатывает (индексирует) полные тексты с наиболее важных Web узлов с качественной медицинской информацией и обеспечивает высокую точность поиска, благодаря использованию специального тезауруса медицинских терминов. Когда пользователь отправляет запрос MWS, заданные им концепты и слова сравниваются с заранее созданным индексом. Результатом является список Web-страниц, содержащих эти слова и концепты. Затем происходит ранжирование этих страниц в порядке значимости концептов и слов в запросе. Для ранжирования по степени релевантности запросу используются знания о медицинских концептах и их отношениях, а также частота встречаемости их на странице и размер этой страницы. База знаний MWS содержит 500 000 медицинских терминов с их отношениями – синонимами, вышестоящими и нижестоящими в иерархии терминами, а также их определения.

Тезаурус NASA. Как и SNOMED — этот тезаурус одно из больших свершений в области лингвистики. Систематизированный свод терминов по ракетной технике и смежным областям.

Из российских систем наиболее известной и большой разработкой является УИС Россия для доступа к разнородным архивам документов, с предоставлением интеллектуального поиска на основе общественно-политического тезауруса. Другие программные продукты, поставляемые на наш российских рынок, являются модулями, на основе которых требуется самим разрабатывать поисковые машины.

Аннотации статей

Чурсин Н.Н. Популярная информатика. Выход в автоматизации? // Электронная библиотека «Наука и техника» – Москва, 2000. – http://www.n-t.org/ri/ch/pi06.htm

В статье описываются первые попытки человека механизировать и

затем автоматизировать поиск информации и его последующее развитие. Представлены характеристики поиска информации такие, как релевантность, полнота и поисковый шум. Автор пытается раскрыть лингвистические проблемы любого поиска, обращает внимание на важность создания информационно-поискового языка и индексирования. В статье рассматриваются принципы использования тезауруса, подчеркивается важность его правильного составления с учетом синонимии. Приведено множество примеров реального использования информационно-поискового тезауруса в различных системах поиска разных стран.

В третьей части статьи рассматривается развитие и использование вычислительных средств в системах поиска в 50-70-е годы нашей страны. Анализируются и описываются задачи и функции интегральных информационных систем. Читателю представляются этапы развития информатики, способы ввода информации с помощью вычислительной техники в 60-80-е годы нашей страны, предлагаются способы структурирования и экономичного хранения информации с помощью микроносителей с описанием реального применения этой практики. Рассматриваются структура и принципы работы документального фонда АСНТИ «Реферат» и иностранных информационных систем.

Анализируются недостатки и проблемы, которые порождал автоматизированный информационный поиск того времени, и способы их решения. Описывается деятельность, развитие, структура и функции ГСНТИ, существовавшее во времена социалистического государства.

Текст статьи написан автором в 1982 году и включен в список раритетных изданий. Информация во многом устарела, но основные принципы методов поиска, описанных в статье, актуальны и в настоящее время. Статья является познавательной и расширяет кругозор читателя, интересующегося принципами создания информационно-поисковых систем, в том числе и на основе информационно-поискового тезауруса.

Лавренова О.А. Информационно-поисковый тезаурус (ИПТ): библиотечная энциклопедия // Российская государственная библиотека – Москва, 2000. – http://www.rsl.ru/pub.asp?bib=1&ch=8&n=4

Автором широко раскрывается термин информационно-поискового тезауруса, его важные составляющие элементы. Термин «тезаурус» был рожден в 13 веке Брунетто Латини в произведении «Книга о сокровище», но известную популярность получил тезаурус, составленный англичанином Роджером в 1852 году. Рассматриваются области использования и назначение информационно-поискового тезауруса. Освещается метод построения тезауруса, который включает в себя создание класси-

фикационной схемы отношений и выявление терминологии. Описывается выделение и обработка ключевых слов тезауруса, отмечены процесс составления тезауруса, как длительной процедуры, и реальность его составления для отдельной отрасли библиотечной системы.

Рыков В.В. Теория тезауруса (Лекция 4) // Курс «Обработка нечисловой информации» – МФТИ: Москва, 1999. – http://ryk-kypc1.narod.ru/tsd.html

В данной лекции приводится обоснование в необходимости применения тезауруса, в частности Интернет. Отмечается, что тезаурус представляет собой, прежде всего, иерархическую классификацию для нахождения нужного денотата. Приведены популярные тезаурусы в мире и их особенности, такие как Roget's, DUDEN, SNOMED, тезаурус НАСА, Wordnet.

Чурсин Н.Н. Популярная информатика (часть 5) // Сайт поддержки научных изобретений. – Москва, 2000. – http://www.inventors.ru/index.asp?mode=5090

Рассматривается модель передачи семантической информации, предложенная Ю.А. Шрейдером, она позволяет разобраться в механизме изменения знания человека. Предпринимается попытка расширить понятие тезауруса, как структуризированное знание в виде иерархического дерева. Описанную выше модель можно применить к процессу передачи информации, в результате которого тезаурусы, или знания, могут обмениваться информацией или точнее пополняться, таким образом, возникают частные случаи пересечения тезаурусов или включения нового фрагмента в тезаурус. В процессе восприятия знания может образоваться новый смысловыражающий элемент, который становится составной частью тезауруса, это возможно в том случае если в тезаурусах, обменивающихся знаниями, существуют для этого элемента какие-то общие связи с другими терминами. Если таких связей нет в получаемом знании, то есть новый термин будет проигнорирован. Таким образом, на основании этих положений модели может предугадать результат коммуникации, например, между людьми различных специальностей. Анализируется на примере реальных случаев важность определенного соотношения тезаурусов при передаче информации.

В данной статье подчеркивается такой недостаток тезаурусного метода поиска, как его ограниченность в пределах определенной предметной области.

Маркарова Т.С. Отраслевой тезаурус в информационно-поисковой системе // Научная техническая библиотека – http://www.gpntb.ru/win/ntb/ntb2002/5/f5_13.htm

Освещается необходимость лингвистического обеспечения в автоматизированных библиотечных системах, в частности в Государственной научной педагогической библиотеке им. К.Д. Ушинского. В качестве наиболее репрезентативного тезауруса библиотека использует тезаурус ЮНЕСКО. Обоснован выбор подобного тезауруса и рассматривается отраслевой тезаурус в области знания – педагогическая наука.

Базовыми единицами данного тезауруса являются дескрипторы. Анализируется применение синонимии в отраслевом тезаурусе и предпринимается попытка использовать родовидовые отношения. Разработанный тезаурус описан восемью предметно-понятийными полями. Сформулированы лингвистические принципы пополнения тезауруса новыми терминами и обосновывается его применение в информационно-поисковой системе на основании лексической наполненности и семантического потенциала.

В результате делается вывод, что отраслевой тезаурус есть некая итоговая форма описания и представление систематизированного материала и может служить лингвистической моделью целостного знания научных текстов. Сформулировано понятие индексирования.

Список цитируемой в статье литературы:

1. UNESCO: IBE education thesaurus, 5-е изд., 1990.
2. Краткий словарь лингвистических терминов. М.: Рус. яз., 1995. С. 31.
3. *Гендина Н.И.* Лингвистическое обеспечение автоматизированных библиотечных систем. Алма-Ата: Гылым, 1991. С. 136.

Журавлев С.В., Добров Б.В. УИС Россия автоматическое тематическое индексирование полнотекстовых документов: основные тезисы доклада // Научно-практическая конференция «Проблемы обработки больших массивов неструктурированных текстовых документов» – http://www.fep.ru/text/dataarrays03.html

В докладе представляются различные технологические решения, осуществленные АНО Центром информационных технологий, наиболее крупным и известным из которых является полнотекстовая информационная ИС Россия для доступа к разнородным архивам документов. Приведены основные функциональные и технические характеристики этой системы. Описаны возможности конверторов для работы с разнородными документами, этапы автоматизированной лингвистической обработ-

ки документов, которая включает в себя морфологический и терминологический анализ, рубрицирование и аннотирование. Кроме того, разработан общественно-политический тезаурус русского языка, перечислены его особенности и структура.

Для более широкого представления приведены конкретные числовые данные разработанного тезауруса и примеры поиска информации в текстовой коллекции УИС Россия.

Браславский П.И. Расширение поискового запроса с помощью тезауруса с сильно дифференцированными связями // – 2000 – http://www.krc.karelia.ru/structure/math/conf/nit/papers/paper.html?file=braslavsky.html&pdate=2000-042&ptitle=Query+expansion+by+using+thesaurus+with+%C1+highly+diversified+ relation+set.

Подчеркиваются основные проблемы пользователя, набирающего запрос в Интернет и значение эвристики и ранжирования найденных документов. Тезаурус можно использовать для расширения полноты поиска, но в Интернет его применение наталкивается на трудности, в частности, его ограниченности в определенной тематике. Создание универсального тезауруса не реально, но для конкретной предметной области он приобретает актуальность. Наиболее адекватное представление тезаурусе можно достичь с помощью набора сильно дифференцированных связей. Такие тезаурусы находятся на стороне пользователя и позволяют тонко настраивать запрос к универсальной машине поиска. Круг пользователей расширенных запросов могут составлять как обычные пользователи, так и специалисты.

Подобный подход построения запроса предполагает итеративность, поиск будет эффективным, но займет время у пользователя, и поэтому подход в принципе не сможет найти широкое распространение.

Список использованной в статье П.И.Браславского литературы:

1. *Браславский П.И.* Пути повышения эффективности поиска научной информации в Internet // Четвертое международное совещание по электронным публикациям (El-Pub-99), Новосибирск 1999.

2. *Браславский П.И., Гольдштейн С.Л., Ткаченко Т.Я.* Тезаурус как средство описания систем знаний // Журнал "Научно-техническая информация", Сер. 2, – 1997. – С. 16-22.

3. *Никитина С.Е.* Семантический анализ языка науки. (На материале лингвистики.) – М.: Наука, 1987.

4. *Никитина С.Е.* Тезаурус по теоретической и прикладной лингвистике. – М.: Наука, 1978.

5. *Aitchison J. et al.* Thesaurus Construction And Use: A Practical Manual. 3rd Edn. London: Aslib, 1997.

Абрагимова И. Medical World Search - механизм поиска медицинской информации в WWW // http://wwwold.aiha.com/english/programs/lrc/topics/mosc5-1.cfm.

Представлено назначение поисковой системы *Medical World Search*. В данной системе реализован полнотекстовый поиск в большом массиве Web-страниц медицинской информации, для индексирования и поиска используется словарно-статистический подход.

В данной разработке используется мета-тезаурус UMLS, включающий в себя различные медицинские словники и тезаурусы. Концепт из UMLS может иметь синонимы из различных словарей и тезаурусов, а сама система UMLS используется для сравнения термина из запроса с терминами из документов, представленных в базе данных и позволяет абстрагироваться от многообразия медицинской терминологии, способствуя ее нормализации.

Приведен список критериев отбора web-страниц для индексирования. Система состоит из трех основных компонент: Web crawler ("ползунок"), индексатор и процессор обработки запроса. Описаны обработка и ранжирование документов. Кроме того, в MWS после получения первоначального результата запроса можно вывести таблицу, в которой детально отражен весь процесс его обработки. Далее описываются дополнительные сервисы системы, и подчеркивается преимущества ее использования.

Заключение

В заключение скажем, что проведенный анализ тезаурусного метода поиска, показал, что он представляет собой мощное лингвистическое средство в системах электронной информации. Если ранее тезаурус существовал на бумаге, и поиск по нему производился вручную, так сегодня имея на руках новейшие информационные технологии, можно превратить его в интеллектуальный способ поиска информации в электронных библиотеках и найти его ограниченное, но незаменимое использование в Интернет поисковых машинах.

Литература

1. *Чурсин Н.Н.* Популярная информатика. Выход в автоматизации? // Электронная библиотека «Наука и техника». – Москва, 2000. – *http://www.n-t.org/ri/ch/pi06.ht;*

2. *Рыков В.В.* Теория тезауруса (Лекция 4) // Курс «Обработка нечисловой информации». – МФТИ: Москва, 1999. – *http://ryk-kypc1.narod.ru/tsd.html;*

3. *Лавренова О.А.* Информационно-поисковый тезаурус (ИПТ): библиотечная энциклопедия // Российская государственная библиотека. – Москва, 2000. – *http://www.rsl.ru/pub.asp?bib=1&ch=8&n=4*;

4. *Журавлев С.В., Добров Б.В.* УИС Россия автоматическое тематическое индексирование полнотекстовых документов: основные тезисы доклада // Научно-практическая конференция «Проблемы обработки больших массивов неструктурированных текстовых документов». – *http://www.fep.ru/text/dataarrays03.html*;

5. *Сегалович И.* Как работают поисковые системы // *http://www.dialog-21.ru/direction_fulltext.asp?dir_id=15539;*

6. *Г.В. Альшанский, Браславский П.И., Титов П.В.* Формирование информационных запросов к машинам поиска интернета на основе тезауруса // VIII Международная конференция по электронным публикациям "EL-Pub2003". –. Новосибирск: Академгородок, 2003. – *http://psb.sbras.ru/ws/show_abstract.dhtml?ru+76+5964*;

7. *Абрагимова И.* Medical World Search - механизм поиска медицинской информации в WWW // *http://wwwold.aiha.com/english/programs/lrc/topics/mosc5-1.cfm*;

8. *Нгуен М.Х., Аджиев А.С.* Описание и использование тезаурусов в информационных системах, подходы и реализация // Российский научный электронный журнал «Электронные библиотеки». – Москва, 2004. – *http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part1/NA*;

9. *Антопольский А.Б.* Лингвистическое обеспечение электронных библиотек // Российский научный электронный журнал «Электронные библиотеки». – Москва, 2002. – *http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2002/part2/antopolskii;*

10. ГОСТ СИБИД 7-25-2000.

А.И.Панченко

ПРОГРАММА ДЛЯ ИЗУЧЕНИЯ НОВОЙ ЛЕКСИКИ «LES MOTS»

Введение

Сегодня не надо никого убеждать в том, что необходимо изучать иностранные языки. Мировой процесс глобализации и интеграции между странами настолько усилился, что специалисту из любой области необходимы уверенные знание как минимум английского языка. В связи с этим существует огромное множество методик и техник обучения иностранному языку. Одни из них более эффективны, другие менее. Некоторые методики предполагают наличие преподавателя, в то время как другие ориентированы на самостоятельное изучения языка.

Проблема изучения новой лексики является одной из главных на любом этапе ознакомления с языком. Несмотря на это, когда человек изучает язык, как правило, он прибегает к небольшому набору самых простых техник. Обычно это или «классическое» просматривание заранее выписанного или же напечатанного в учебнике списка новых слов (возможно, с закрыванием перевода значения слов листком бумаги) или же использование *карточек со словами*. Первая методика, само собой является действенной, несмотря на всю свою простоту, потому что ей пользуется большинство людей изучающих язык. Тем не менее, у методики визуального просмотра слов написанных в учебнике или тетради существует множество ограничений и недостатков:

1. Необходимость закрывать перевод слов. Для того чтобы быть уверенным, что слово было выучено, следует закрыть листком бумаги или рукой перевод. Как правило, это сделать трудно, если слова были выписаны в тетрадь т.к. они не выровнены относительно друг друга.
2. Неудобство или невозможность изучения слов «на ходу». Известно, что для выучивания некоторого множества слов необходимо не один раз их прочитать или встретиться в тексте с ними. Если вы заучиваете слова по тетради или учебнику, то вам может просто не хватить времени чтобы повторить слова несколько раз. Гораздо более продуктивнее учить слова, заполняя этим занятием бесполезные промежутки времени, коих возникает время от времени большое множество. Так не

будет тратиться время специально на заучивание, что мало эффективно, а само запоминание будет более простым и качественным.

3. Необходимость читать все слова подряд, даже те которые уже изучены. При изучении какого-либо блока лексики все слова запоминаются по-разному, в зависимости от каждого конкретного человека. Таким образом, некоторые слова при обычном методе заучивания приходится читать снова и снова, несмотря на то, что они уже запомнились, в то время как действительно не выученным словам уделяется меньше внимания, чем необходимо.

4. Неудобство хранения и повторного доступа к блоку лексики. Если слова были записаны в тетради то, как правило, поиск нужного слова слишком долгий и его проще посмотреть в словаре.

Что касается второй популярной техники выучивания слов – *карточек со словами*, это давно известная, высокоэффективная и популярная методика изучения. Она часто рекомендуется в качестве основной для студентов лингвистических ВУЗов, которым приходится учить огромные объемы новой лексики каждый день. Суть данной методики состоит в следующем: новые слова записываются на небольшие бумажные карточки. С одной стороны такой карточки пишется слово на русском языке, с другой – на иностранном. При заучивании человек перебирает карточки и, если он забыл перевод слова, то тогда карточка переворачивается и читается перевод. Если обучающийся помнит перевод, он может отложить данную карточку как выученную. Таким образом, нивелируются такие недостатки как необходимость закрывать перевод (пункт 1) и читать уже выученные слова (пункт 3). Из этого следует, что методика изучения слов с помощью карточек лишена нескольких серьезных недостатков классической методики, однако возникают и дополнительные недостатки, главный из которых это необходимость физического создания карточек (выписывание слов, нарезка бумаги и т.п.). Так же представляется, что удобство работы с карточками вряд ли превосходит классическую методику (в транспорте можно запросто выронить стопку карточек, изученные карточки необходимо куда-либо откладывать и т.п.). С задачей удобного доступа к блоку лексики карточки справляются даже хуже чем слова, записанные в тетради. Хранение бумажных карточек тоже видится проблемой.

В последнее время, получили большое распространение *карманные персональные компьютеры* (КПК), многие из них обладают GSM-модулем и выполняют, так же функцию телефона (в этом случае их часто называют *коммуникаторами*). Такие устройства постоянно под рукой у владельца, но при этом они являются достаточно функциональ-

ными компьютерами. Современные операционные системы подобных носимых устройств, к примеру, Windows Mobile 2003/2005, позволяют легко устанавливать на них дополнительное прикладное программное обеспечение.

Рис.1. Современные КПК и коммуникаторы

Целью данной разработки является создание средства для изучения слов, в основе которого лежала бы методика изучения слов по карточкам, но который был бы лишен неудобств, связанных с изготовлением и хранением карточек. Для этого предполагается создание *электронных*

карточек, доступ к которым был бы возможен посредством ПК, КПК или мобильного телефона. Таким образом, карточки стали бы доступными в любом месте и в любое время: для того чтобы повторить слова, необходимо было бы всего лишь взглянуть на экран коммуникатора или КПК. Это позволило бы заполнять изучением лексики даже незначительные перерывы во время рабочего дня, коих всегда присутствует большое количество: ожидание транспорта, поездка в переполненном автобусе, ожидание заказа в кассе.

Установка программы и системные требования

Установка программы Les Mots состоит из двух частей: инсталлирование windows-версии и установка PDA-версии. Опишем оба этих процесса чуть более подробно:

Windows-версия

Программа работает практически со всеми современными операционными системами семейства Microsoft Windows: Windows XP, Windows 2000, Windows 98, Windows ME, Windows Server 2003, Windows Vista. Для нормальной работы программы достаточно минимальных системных требований, которые нужны для установки соответствующей ОС.

Для установки windows-версии нужно запустить инсталлятор LesMotsInstall.msi и следовать указаниям на экране монитора.

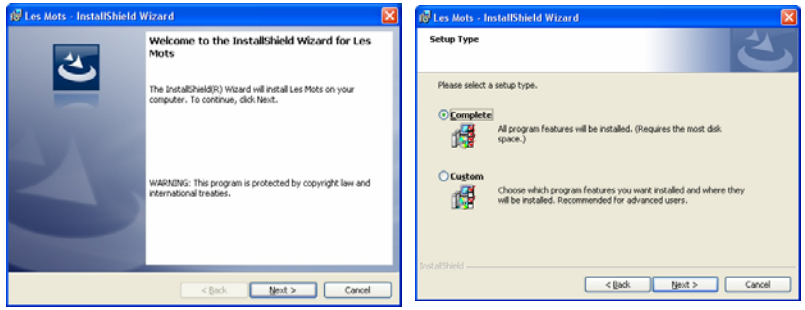


Рис. 2. Установка Windows-версии программы Les Mots

КПК-версия

КПК-версия программы может работать на устройствах под управлением ОС семейства Windows Mobile, таких как Windows Mobile 2003/2005 и Windows Mobile 6.

Для того чтобы установить КПК-версию нужно скопировать на флеш карту или внутреннюю память устройства файл LesMotsIntall.cab и запустить его. После совершения указанных выше действий, программа будет полностью установлена, и можно будет приступать к работе с ней.

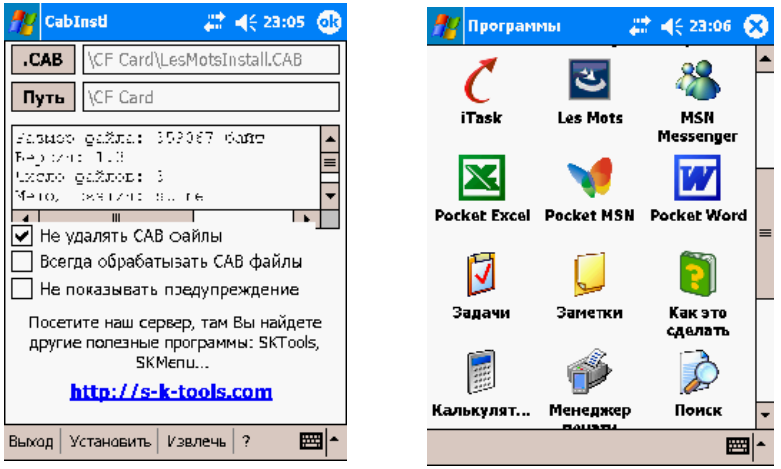


Рис.3. Установка программы Les Mots на КПК

Описание программы

Программа Les Mots состоит из Windows и КПК версий. Каждая из версий обладает практически идентичной функциональностью: возможность просмотра списка слов в наборе карточек, редактирование и создание новых электронных карточек и обучение новым словам. Обе версии, само собой, работают с одним форматом файлов электронных карточек, то есть карточки созданные в windows-версии, легко открываются на карманном компьютере и наоборот.

Интерфейс программ интуитивно понятен. Главное окно версии для Win32 состоит из таблицы с новыми словами и переводом. В данном случае выбран французский язык в качестве иностранного и русский в качестве родного.

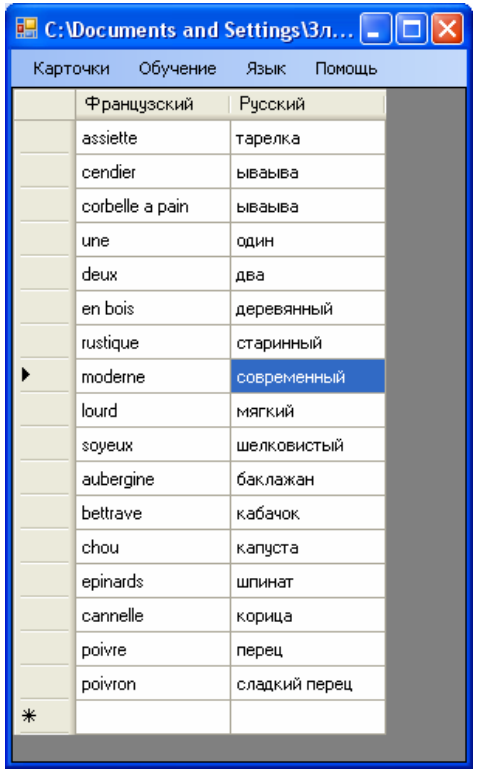


Рис.4. Интерфейс Windows-версии

Создание электронных карточек

Несмотря на то, что и Windows и КПК версии программ позволяют создавать файлы формата электронных карточек со словами, куда более эффективно подготавливать учебные материалы в настольной версии из-за гораздо более высокой скорости ввода слов. Само собой пользователь быстрее набирает слова на полноценной клавиатуре, нежели посредством последовательного ввода слов с помощью сенсорной клавиатуры коммуникатора или другого носимого устройства. Стоит создавать карточки на ПК еще и потому что при вводе слов нет необходимости переключаться между раскладками клавиатуры – в программе реализована функция автоматического переключения языков ввода.

Для того чтобы сформировать новую карточку нужно выполнить команду *Карточки → Новая*. После этого создастся новая электронная карточка, и можно будет приступать к её заполнению словами. При этом каждая строка в таблице, фактически соответствует одной бумажной карточке.

Стоит отметить то, что во время заполнения словами не нужно постоянно переключать кодировки клавиатуры, потому что программа, в зависимости от выбранных языков, сама меняет раскладки после получения фокуса ввода соответствующей ячейкой. Это крайне удобно, вследствие того, что не нужно менять раскладку после каждого введенного слова. Эта особенность windows-версии программы позволяет крайне быстро формировать наборы карточек с требуемой лексикой.

Изменить язык карточки можно используя пункты меню *Язык → Иностранный* и *Язык → Родной*.

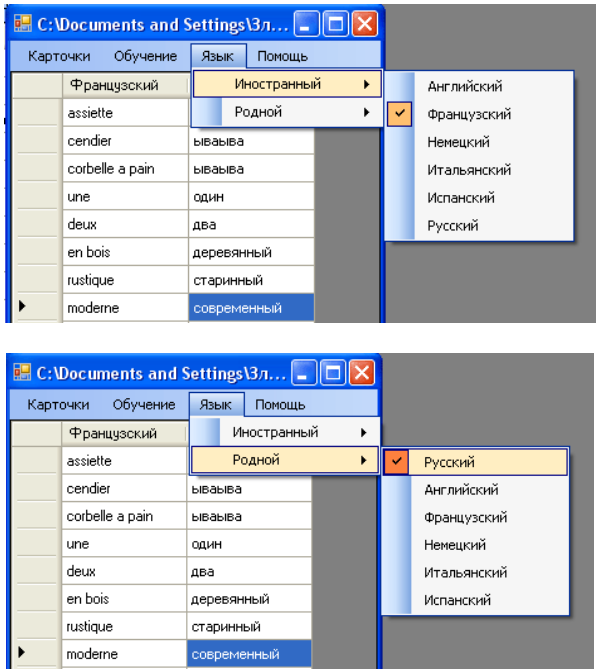


Рис.5. Изменение языка электронной карточки

Указание правильного языка для карточки важно т.к. в зависимости от этого выбора будет соответствующим образом автоматически переключаться раскладка клавиатуры.

После того как все необходимые слова будут введены, нужно сохранить набор карточек в файл посредством пункта меню *Карточки → Сохранить* или *Карточки → Сохранить как*.

В первом случае файл с созданным набором электронных карточек будет сохранен с именем заданным по умолчанию (les mots <дата создания>.xml) на Рабочий стол. В случае выбора второго пункта можно указать произвольное место на жестком диске, куда будет записан файл.

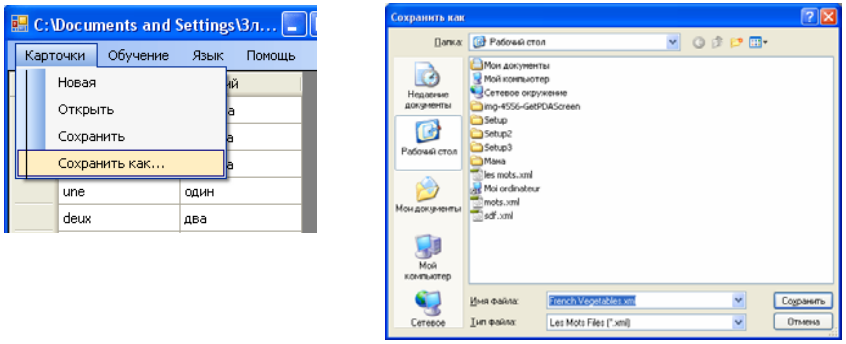


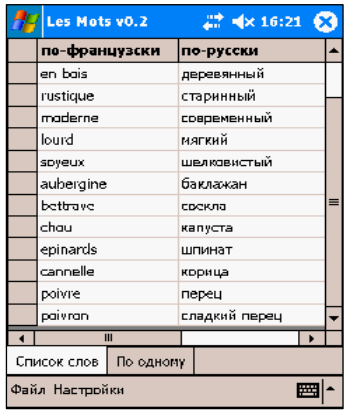
Рис.6. Сохранение файла электронных карточек

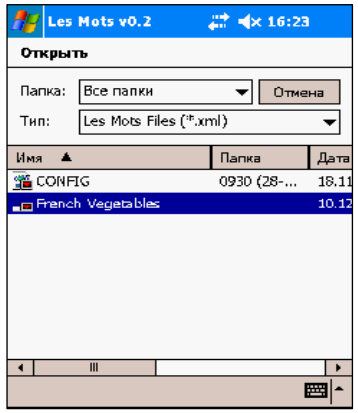
Электронные карточки хранятся в формате XML файлов и являются легко доступными для обработки. Для того чтобы редактировать уже имеющийся файл с электронными карточками нужно воспользоваться командой *Карточки →*

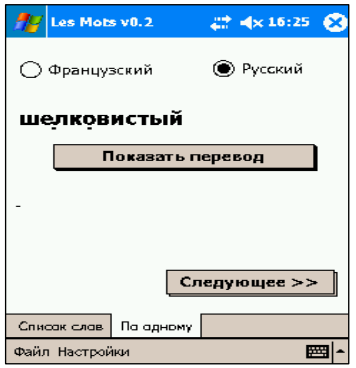
```
   <?xml version="1.0" standalone="yes"?>
   <motsDataset
xmlns="http://tempuri.org/motsDataset.xsd">
     <Words>
       <Id>0</Id>
       <Foreign>assiette</Foreign>
       <Native>тарелка</Native>
       <ForeignLearnedCount>0</ForeignLearnedCount>
       <NativeLearnedCount>0</NativeLearnedCount>
     </Words>
     <Words>
```

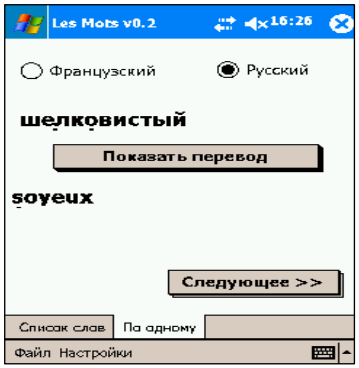
```
      <Id>1</Id>
      <Foreign>cendier</Foreign>
      <Native>ываыва</Native>
      <ForeignLearnedCount>0</ForeignLearnedCount>
      <NativeLearnedCount>0</NativeLearnedCount>
    …
    <Settings>
      <NativeLanguage>Русский</NativeLanguage>
      <ForeignLanguage>Французский</ForeignLanguage>
    </Settings>
</motsDataset>
```

Использование КПК-версии программы









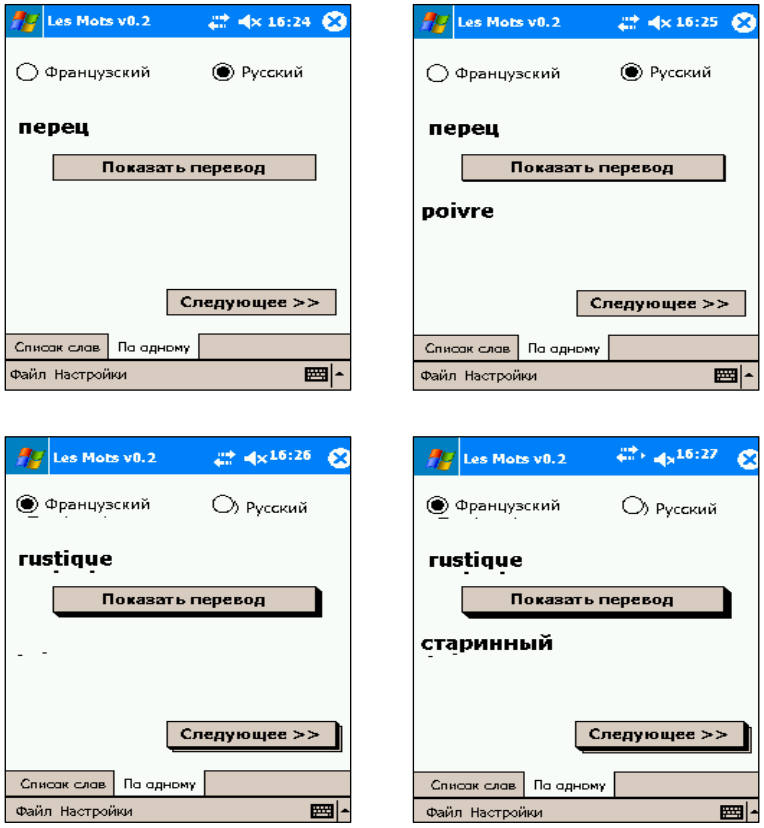


Рис. 7–14. Интерфейс КПК версии.

Заключение

В результате разработки программы Les Mots было создано средство для изучения новой лексики, которое обладает всеми преимуществами традиционной методики заучивания слов с помощью карточек: возможностью отбрасывания изученных слов, возможностью просмотра слов в произвольном порядке, контролем над степенью изученности блока лексики, а также высокой скоростью изучения слов. При этом решена проблема неудобства изготовления хранения и работы с карточками. Используя программу Les Mots, становится возможным работать с карточ-

ками практически в любой удобный момент: нет необходимости доставать стопку с бумажками, перемешивать их и куда-либо откладывать изученные карточки. Для начала работы достаточно всего лишь запустить приложение на КПК или коммуникаторе.

Более того, в данном программном продукте решена проблема доступности блоков лексики. Раньше чтобы повторить какой-либо раздел нужно было носить с собой несколько наборов карточек, что редко было возможным. Теперь же электронные карточки хранятся в виде XML-файлов и могут всегда быть под рукой – достаточно просто загрузить из программы необходимый файл.

Помимо всего этого концепция электронных карточек позволяет учителю раздавать ученикам в качестве учебного материала XML файлы с уже подготовленными карточками. Таким образом, для начала изучения слов с помощью программы остается только открыть выданный файл с набором электронных карточек и сразу же приступить к изучению новых слов.

Литература

1. *Андреасян И.М., Леонтьева Т.П., Бабинская П.К.* Практический курс методики преподавания иностранных языков: английский, французский, немецкий. Тетра-Системс, 2006.
2. *Утробина А.А.* Методика преподавания и изучения иностранного языка. Приор, 2006.
3. Статьи об изучении английского языка (http://www.native-english.ru/articles/)
4. *Китайгородская Г.А.* Методические основы интенсивного обучения иностранным языкам. МГУ, 1986.
5. Английский за 3 недели. Карточки для изучения языка. Продвинутый уровень. Дельта Паблишинг, 2004.

М.Е. Постников

МЕТОДЫ ВЫБОРА ОБОРУДОВАНИЯ ДЛЯ ОРГАНИЗАЦИИ ЭФФЕКТИВНОЙ РАБОТЫ СЕТИ*

Введение

Часто приходится принимать такие решения, последствия которых бывают очень весомы, например, выбор оборудования для организации эффективной работы системы обработки информации (СОИ) на базе ЛВС или выбор варианта архитектуры самой СОИ для офиса или подразделения банка и т.д. При этом, при выборе наилучшего варианта из набора альтернативных вариантов, следует учитывать не один, а несколько показателей, которые оценивают сравниваемые варианты с разных точек зрения. Поэтому, если такие решения принимаются без использования математического аппарата, то в большинстве случаев они не являются оптимальными.

При выборе наилучшего варианта, необходимо учитывать условия (ситуации), в которых принимаются решения: полная определенность, неопределенность или риск.

В дальнейшем будем рассматривать такие методы принятия решения, которые касаются выбора только одного, наилучшего варианта, из множества рассматриваемых и выбор варианта осуществляет лицо принимающее решение (ЛПР) в условиях полной определенности.[1, 2]

Рассмотрим более детально основные этапы процесса принятия решения, а также математические методы, используемые ЛПР при принятии решения.

Основные этапы процесса принятия решения.

При выделении основных этапов процесса принятия решений будем использовать ситуационный подход, который наиболее полно отражает реальную последовательность действий ЛПР при принятии технических, управленческих и организационных решений.

В этом случае основные этапы процесса принятия решения такие:

---

* Руководитель работы к.т.н., доцент Постников В.М.

1. Формирование цели исследования.
2. Формирование набора альтернативных вариантов, подлежащих сравнению. При этом все варианты, подлежащие сравнению, должны быть практически реализуемы.
3. Формирование набора показателей, с помощью которых осуществляется сравнение альтернативных вариантов. Суммарное количество таких показателей должно быть в пределах 3-10. При этом эти показатели должны отражать основные особенности сравниваемых вариантов.
4. Формирование набора недоминируемых вариантов (вариантов, входящих во множество Парето), среди которых и выбирается наилучший вариант. При этом доминируемые варианты, это те варианты, которые по всем показателям не лучше какого то варианта, относящегося к классу недоминируемых вариантов, отбрасываются.
5. Определение коэффициента важности показателей качества, используемых для сравнения недоминируемых вариантов. Этот коэффициент показывает степень важности (вес) одних показателей перед другими. ЛПР оценивает в баллах. степень важности каждого показателя, например, 1 балл, 2 балла и т.д. Далее определяется суммарное количество баллов. После этого количество баллов каждого показателя делится на суммарное количество баллов и получается весовой коэффициент соответствующего показателя.
6. Формирование набора критериев (правил свертки показателей качества в единый критерий) для сравнения недоминируемых вариантов и выбора наилучшего из них. Этот набор критериев формирует ЛПР из набора типовых критериев сравнения. Типовой набор критериев для сравнения альтернативных вариантов в условиях полной определенности подробно рассмотрен в следующем разделе.
7. Формирование порядка использования (приоритета использования) критериев, входящих в сформированный ЛПР набор критериев.
8. Последовательное определение количественной оценки сравниваемых вариантов по каждому из выбранных критериев
9. Определение степени устойчивости выбранного наилучшего решения. ЛПР определяет насколько прочно наилучший вариант занимает ведущие позиции при использовании разных критериев.
10. Выбор окончательного варианта наилучшего решения на основе результатов, полученных с использованием разных критериев

Критерии выбора наилучшего решения в условиях полной определенности

Метод главного локального критерия (показателя):

Y = max Бj j, где Б – наиболее важный показатель, по которому сравниваются все варианты. Выбирается вариант, у которого максимальное значение этого показателя.

Метод двух основных локальных критериев (показателей)

$$Y = \max_{j} \frac{Бj}{Cj}.$$

Все варианты сравниваются по двум показателям, например, производительность отнесенная к стоимости. Выбирается вариант, у которого максимальное значение этого показателя.

Метод взвешенной суммы равнозначных показателей сравнения.

$$Y = \max \sum_{i=1}^{n} Kij$$

n — количество показателей сравнения; m — количество вариантов сравнения, j∈m.

Будем производить сравнение j вариантов (j=1,m) по i параметрам.(i=1,n)

Kij – нормированный коэффициент соответствия i-ого параметра j-ого варианта эталонному значению, т.е. для j-ого варианта

$$Kj = \frac{Xij}{\boldsymbol{max}_{j} Xi}; \ 0 < \text{Kij} < 1$$

Xij – значение i-ого параметра j-ого варианта в реальных единицах.

варианты / показатели	1	2	M
1	K11	K12	K1m
. . . . .			
N	Kn1	Kn2	Knm
	∑k1	∑k2	∑km
наилучший вариант	Max		

Метод взвешенной суммы неравнозначных показателей сравнения.

Учитывается еще и важность (в виде весового коэффициента) каждого показателя сравнения.

$$Y = \max_{j \in m} \sum_{i}^{n} \alpha_i Kij$$

αi- весовой коэффициент i-ого показателя сравнения.
Весовые коэффициенты выбираются из условия:

$$\sum_{i=1}^{n} \alpha i = 1; \quad 0 < \alpha i < 1$$

варианты / показатели	αi	1	2	…….	M
1	α1	k11	k12	……	k1m
2	α2	k21	k22	……	k2m
……..	…..	…..	…..	……	……
N	αn	kn1	kn2	……	Knm
		∑k1	∑k2	……	∑km
наилучший вариант		max			

Метод произведения показателей сравнения.

$$Y = \max_{j \in m} \prod_{i=1}^{n} Kij$$

Метод взвешенного произведения показателей сравнения.

$$Y = \max_{j \in m} \prod_{i=1}^{n} (Kij)^{\alpha_i}$$

Метод близости сравниваемых вариантов к идеальному.

$$Y = \min_{j} \sqrt{\sum_{i=1}^{n} \alpha_i \left(\frac{Ki\max - Kij}{K_{i\max}} \right)^2}$$

Минимаксный критерий Вальда.

Минимизировать максимальное отклонение нормированного показателя сравниваемого варианта от эталонного значения $Y = \min_{j} \max_{i} W_{ij} \quad W_{ij} = (\max_{j} K_{ij} - K_{ij})$ для i = от 1 до n

Показатели от 1 до n	Варианты от 1 до m		
	1	2	3
1	W11	W12	W13
2	W21	W22	W23
n	Wn1	Wn2	Wn3
$\max_i W_{ij}$	W21	W22	W13
Наилучший вариант $Y = \min_j \max_i W_{ij}$	W21		

Пример

Принято решение о построении системы обработки информации офиса на базе беспроводной ЛВС стандарта 802,11g с использованием точек доступа со встроенными в них маршрутизаторами. Требуется выбрать вариант такой точки доступа из серийно выпускаемых, которая в наибольшей степени отвечает предъявляемым к ней требованиям. Для этого сначала необходимо выявить точки доступа, подлежащие сравнению, затем показатели сравнения оборудования и критерии, по которым следует проводить сравнение точек доступа.

Сравниваются следующие перспективные точки доступа, которые выпускаются ведущими компаниями, лидерами рынка беспроводных ЛВС [3]

Вариант 1 (B1) - Dell TrueMobile 2300
Вариант 2 (B2) - Gateway WGR-250
Вариант 3 (B3) - Linksys Wireless-G WRT54GS
Вариант 4 (B4) - U.S. Robotics USR8054
Вариант 5 (B5) - D-Link AirPremier AG DI-784
Вариант 6 (B6) - ZyXEL Prestige 334W
Вариант 7 (B7) - Linksys Wireless A+G WRT55AG

Точки доступа сравниваются по трем количественным параметрам (средняя реальная производительность, радиус действия, цена) и двум качественным (функциональные возможности и средства обеспечения безопасности). Все сравниваемые параметры имеют одинаковую важность.

Исходные данные сравниваемых вариантов точек доступа приведены в табл.1

Табл. 1

Первоначальные исходные данные
сравниваемых вариантов точек доступа

Показатели сравнения	Сравниваемые варианты точек доступа						
	B1	B2	B3	B4	B5	B6	B7
Средняя реальная производительность до L < 18 м ,(Мбит/сек)	20,2	20,5	30,2	21,7	39,3	20,1	19,3
Максимальный радиус действия (м)	50	50	45	40	45	35	45
Функциональные возможности - мастер установки - руковод по установке -ф-ии брандмауэра, DHCP сервера, UPnP -ф-ции диагностики - стандарт (802.11a/g)	Оч. хор	Отл	Отл	Хор	Отл	Хор	Отл
Средства обеспечения безопасности - алгоритмы WEP и WPA шифрования - фильтр разрешения доступа по расписанию - аутентификация RADIUS.	Оч. хор	Оч. хор	Отл	Хор	Оч. хор	Оч. хор	Хор
Цена (усл. ед)	90	99	129	99	170	90	289

Решение

Анализ вариантов показывает, что 6-ой вариант ни по одному показателю не превосходит 1-ий вариант, а 7-ой вариант ни по одному показателю не превосходит как 3-ий, так и 5-ый варианты. Поэтому 6-ой и 7-ой варианты можно отнести к классу вариантов, над которыми есть доминирующие, т.е. более лучшие варианты, и исключить их из дальнейшего рассмотрения. Таким образом, задача упрощается и необходи-

мо сравнить только пять вариантов по пяти параметрам и среди них выбрать наилучший вариант.

Для перевода качественных параметров в количественные используем следующую нормированную шкалу перевода

Отл	Оч. хор	Хор	Удовл	Плохо
1	0,9	0,8	0,7	0,6

После преобразований первоначальных значений исходных данных получаем нормированные значения исходных данных вариантов сравнения, которые приведены в табл.2.

Табл. 2

Нормированные исходные данные вариантов сравнения.

Показатели Сравнения	Варианты сравнения				
	B1	B2	B3	B4	B5
Средняя реальная производительность	0,51	0,52	0,77	0,55	1
Максимальный радиус действия	1	1	0,9	0,8	0,9
Функциональные возможности	0,9	1	1	0,8	1
Средства обеспечения безопасности	0,9	0,9	1	0,8	0,9
Цена	1	0,9	0,7	0,9	0,53

В табл.3-7 приведены результаты сравнительного анализа вариантов точек доступа по предложенным критериям.

Табл. 3

Сравнительный анализ вариантов методом взвешенной суммы равнозначных показателей сравнения.

Показатели Сравнения	Варианты сравнения				
	B1	B2	B3	B4	B5
Средняя реальная производительность	0,51	0,52	0,77	0,55	1
Максимальный радиус действия	1	1	0,9	0,8	0,9
Функциональные воз-	0,9	1	1	0,8	1

можности					
Средства обеспечения безопасности	0,9	0,9	1	0,8	0,9
Цена	1	0,9	0,7	0,9	0,53
$Y=\sum_{i=1}^{n} Kij$	4,31	4,32	4,37	3,85	4,33
$Y=\max\sum_{i=1}^{n} Kij$			4,37		
Место варианта	4	3	1	5	2

Табл. 4

Сравнительный анализ вариантов
методом произведения показателей сравнения.

Показатели Сравнения	Варианты сравнения				
	B1	B2	B3	B4	B5
Средняя реальная производительность	0,51	0,52	0,77	0,55	1
Максимальный радиус действия	1	1	0,9	0,8	0,9
Функциональные воз-можности	0,9	1	1	0,8	1
Средства обеспечения безопасности	0,9	0,9	1	0,8	0,9
Цена	1	0,9	0,7	0,9	0,53
$Y=\prod_{i=1}^{n} Kij$	0,413	0,421	0,485	0,253	0,429
$Y=\max_{j\in m}\prod_{i=1}^{n} Kij$			0,485		
Место варианта	4	3	1	5	2

Табл. 5

Сравнительный анализ вариантов методом
близости сравниваемых вариантов к идеальному варианту.

Показатели сравнения	Варианты сравнения					
	B1	B2	B3	B4	B5	Идеал
Средняя реальная производительность	0,51	0,52	0,77	0,55	1	1
Максимальный радиус действия	1	1	0,9	0,8	0,9	1
Функциональные возможности	0,9	1	1	0,8	1	1
Средства обеспечения безопасности	0,9	0,9	1	0,8	0,9	1
Цена	1	0,9	0,7	0,9	0,53	1
$Y = \sqrt{\sum_{i=1}^{n} \left(\frac{Ki\,\boldsymbol{max} - Kij}{K_{i\,\boldsymbol{max}}} \right)^2}$	0,510	0,5	0,391	0,576	0,49	
$Y = \min_{j} \sqrt{\sum_{i=1}^{n} \left(\frac{Ki\,\boldsymbol{max} - Kij}{K_{i\,\boldsymbol{max}}} \right)^2}$			0,391			
Место варианта	4	3	1	5	2	

Табл. 6

Сравнительный анализ вариантов методом
минимаксного критерия Вальда.

Показатели сравнения	Варианты сравнения				
	B1	B2	B3	B4	B5
Средняя реальная производительность	0,51	0,52	0,77	0,55	1
Максимальный радиус действия	1	1	0,9	0,8	0,9
Функциональные воз-	0,9	1	1	0,8	1

можности					
Средства обеспечения безопасности	0,9	0,9	1	0,9	0,9
Цена	1	0,9	0,7	0,9	0,53
$\max\limits_{i} W_{ij}$ $W_{ij} = (\max\limits_{j} K_{ij} - K_{ij})$	0,49	0,48	0,3	0,45	0,47
Наилучший вариант $Y = \min\limits_{j} \max\limits_{i} W_{ij}$			0,3		
Место варианта	5	4	1	2	3

Предлагается еще один метод сравнения альтернативных вариантов *Метод близости сравниваемых вариантов к рациональному.*

Рациональный вариант по своему усмотрению на основе анализа исходной информации выбирает ЛПР. Допустим, что в нашем примере ЛПР в качестве рационального варианта выбирает вариант с такими показателями:

- Средняя реальная производительность (30 Мбит/сек)
- Максимальный радиус действия (40 м)
- Функциональные возможности (Отл)
- Средства обеспечения безопасности (Отл)
- Цена (120 усл.ед)

Тогда нормированные значения показателей сравнения Кірац для (i= 1-5) рационального варианта соответственно будут:

(30/39 =0,77): (40/50=0,8); (Отл =1); (Отл =1); (90/120=0,75)

При этом, если вариант имеет лучшее значение показателя по сравнению с рациональным вариантом, то значение (Кірац - Кij) будет меньше нуля. В этом случае, если разница значений (Кірац - Кij) < 0, то мы будем считать ее равной нулю, так как нас устраивает рациональный вариант.

Табл. 7

Сравнительный анализ вариантов методом
близости сравниваемых вариантов к рациональному варианту.

Показатели Сравнения	Варианты сравнения Приведены значения (Кірац - Кij)

	B1	B2	B3	B4	B5	Кірац
Средняя реальная производительность	0,26	0,25	0	0,22	0	0,77
Максимальный радиус действия	0	0	0	0	0	0,8
Функциональные возможности	0,1	0	0	0,2	0	1
Средства обеспечения безопасности	0,1	0,1	0	0,1	0,1	1
Цена	0	0	0,05	0	0,22	0,75
$Y = \sqrt{\sum_{i=1}^{n} \left(\frac{Кірац - Кij}{K_{ірац}} \right)^2}$	0,366	0,339	0,066	0,362	0,31	
$Y = \min_{j} \sqrt{\sum_{i=1}^{n} \left(\frac{Кірац - Кij}{K_{ірац}} \right)^2}$			0,066			
Место варианта	5	3	1	4	2	

Представим результаты сравнительного анализа пяти вариантов по разным критериям в виде табл 8, в которой за первое место варианту по данному критерию даем 1 балл, за второе 2 и т.д Наилучшим считается вариант, у которого наименьшая итоговая сумма баллов.

Табл. 8

Сравнительный анализ пяти вариантов точек доступа по разным критериям

Критерии сравнения	Варианты сравнения (оценка - номер места варианта по критерию сравнения)				
	B1	B2	B3	B4	B5
Метод взвешенной суммы показателей сравнения.	4	3	1	5	2
Метод произведения показателей сравнения.	4	3	1	5	2
Метод близости сравниваемых вариантов к иде-	4	3	1	5	2

альному варианту.					
Минимаксный критерий Вальда	5	4	1	2	3
Метод близости сравниваемых вариантов к рациональному варианту.	5	3	1	4	2
Суммарная оценка	22	16	5	21	11
Место варианта	5	3	1	4	2

Наилучшим решением является вариант 3 - точка доступа Linksys Wireless-G WRT54GS. Второе место занимает вариант 5, точка доступа - D-Link AirPremier AG DI-784, а третье место вариант 2, точка доступа - Gateway WGR-250

Выводы

1. Проведен обзор типовых критериев, которые используются для выбора наилучшего варианта из набора альтернативных вариантов в условиях полной определенности.

2. Предложен еще один критерий для выбора наилучшего варианта из набора альтернативных вариантов в условиях полной определенности, критерий близости сравниваемых вариантов к рациональному варианту,.

3. Предложен подход к выбору наилучшего варианта из набора альтернативных вариантов, который учитывает результаты, полученные при использовании множества критериев выбора.

Литература

1. *Михайлов В.И.* Как принимать решения. – СПб.: Химера,, 1998,- 200 с.

2. *Волков К.Е., Загоруйко Е.А.* Исследование операций. - М.: МГТУ. 2004.- 440 с.

3. *Эллисон Крейг.* Стандарт “G” для всех // PC Magazine / Russian Edition. – 2004. - №9. - С. 124 – 142.

А.А. Проскурнин

АВТОМАТИЗИРОВАННАЯ СИСТЕМА КОНТРОЛЯ ЗНАНИЙ

Проблемы освоения компьютеров в образовании, усовершенствования педагогических технологий на основе человеко-компьютерного диалога в учебном процессе исследуются образовательным сообществом практически с момента промышленного производства компьютеров, т.е. с начала 50-х годов.

По мере совершенствования технических характеристик самого компьютера и его программного обеспечения, расширения его дидактических возможностей утвердилась идея о принципиально новых свойствах компьютера как средства обучения. Компьютер позволяет строить обучение в режиме диалога, реализовать индивидуализированное обучение, опирающееся на модель учащегося, его "историю обучения". Изменилась оценка роли и места компьютера в учебном процессе. К началу 90-х годов были созданы десятки тысяч различных обучающих систем, их общепринятой классификации не существует. Многие авторы выделяют следующие типы систем: тренировочные; контролирующие, наставнические, проблемного обучения; имитационные и моделирующие; игровые программы.

На сегодняшний день в России, на фоне информатизации системы образования, распространения дистанционного обучения, развития технологий управления персоналом актуальной является задача разработки автоматизированных информационных систем, позволяющих объективно и быстро оценивать знания, умения и навыки обучаемых.

В данной работе рассматривается автоматизированная система контроля знаний, которая разработана автором в рамках дипломного проектирования. Целью разработки системы является повышение эффективности процесса обучения, реализация методики контроля теоретических знаний, применимой в различных областях приложения – как в системе среднего и высшего образования, так и в кадровом менеджменте, для аудиторной и дистанционной форм обучения. Назначение системы – создание, применение и оценка качества учебных тестов, которые представляют собой набор тестовых заданий открытого типа специфической формы.

Данная разработка является реализацией определенной методики контроля знаний. Необходимо уточнить это утверждение. Исходя из обзора методов и моделей контроля знаний, проведенного в дипломном проекте, можно сделать вывод, что существует достаточно большое число самых разнообразных форм, методов, методик, моделей, идей, применяемых для контроля знаний. При этом невозможно сказать о том, что какая-то из моделей контроля знаний является более предпочтительной, обязательно даст наилучшие результаты. Здесь все определяется целями, которые ставит себе разработчик тестов, преподаватель, организационными возможностями, спецификой содержания различных учебных дисциплин. В то же время нельзя не отметить, что среди различных моделей контроля знаний можно выделить более субъективные, авторские модели, которые опираются, как правило, во многом, на эмпирические соображения, и можно выделить более устоявшиеся, классические модели, которые имеют достаточную научную проработку и являются признанными в теории педагогических измерений, тестологии, квалиметрии.

При разработке данного программного продукта была поставлена задача использования оригинальной информационной модели тестового задания и базы тестовых заданий, основанной на модели представления вербальных языковых знаний, предложенной Ю.Н. Карауловым [1]. В этом состоит основное отличие данного программного продукта от остальных. Задача создания наиболее адекватных моделей базы тестовых заданий является актуальной в настоящее время. Поскольку такая модель реализована в данном программном продукте, а она фактически позволяет описывать вербальные языковые знания в некоторой области, постольку можно использовать данный программный продукт для создания баз вербальных языковых знаний в различных предметных областях, независимо от его основного назначения.

Целью данной статьи не является подробное рассмотрение применяемой модели представления вербальных языковых знаний. Поэтому здесь опускаются рассуждения, которые определили переход от данной модели к информационной модели базы тестовых заданий. Информационная модель тестового задания состоит из следующих компонентов: формула смысла (формулировка задания), множество знаков (множество правильных ответов), способ задания смысла, множество дидактических единиц, входящих в разные тематические разбиения, функция знания.

В свою очередь информационная модель базы тестовых заданий состоит из следующих компонентов (множеств): знаков, формул смысла,

способов задания смысла, тематических разбиений, дидактических единиц, значений функции знания, тестовых заданий.

Выбранная модель тестового задания определила и его форму. Каждое задание представляет собой предложение, которое определяет или характеризует некоторый термин (понятие) учебного предмета. Испытуемый должен прочитать предложение, осмыслить его, и определить термин, о котором идет речь в предложении. Наиболее близкая классическая форма тестовых заданий – это так называемое тестовое задание на дополнение с ограничением на ответ, когда ученик дописывает пропущенное слово, формулу, символ или число на месте прочерка. Задания на дополнения кажутся ученикам более трудными, так как в них исключается догадка. Действительно, легче выбрать правильный ответ из предложенных, основываясь не столько на знаниях, сколько на интуиции, чем самому его сформулировать или найти в процессе решения поставленных проблем. Но именно это свойство делает задания на дополнение исключительно привлекательными для педагогов, особенно для тех, кто привык в своей работе опираться на традиционные средства контроля и не доверяет тестам. Есть и недостатки, связанные с трудностями, возникающими при оценке ответа учеников. Дописывая ответ на месте прочерка, ученик может выбрать синонимы пропущенного запланированного разработчиком слова или изменить порядок следования элементов в пропущенной формуле, что значительно затрудняет проверку и оценку результатов учеников. Рекомендуется, чтобы ответ на задание был достаточно кратким, как правило, одно слово, по крайней мере, не превышающим двух-трех слов.

Ниже приведены некоторые примеры тестовых заданий с указанием для них способа задания смысла как вида тестового задания (предмет «Информатика»):

1) Он может быть персональным, карманным и даже квантовым (КОМПЬЮТЕР). Способ: указание типов, видов для понятия.

2) «Мозг» компьютера (ПРОЦЕССОР). Способ: метафора.

3) В языке Pascal он бывает трех типов: со счетчиком, с предусловием, с постусловием (ЦИКЛ). Способ: указание типов, видов для понятия.

4) Как называется процесс, в результате которого из массива {5, 2, 1, 3, 8} получается массив {1, 2, 3, 5, 8}? (СОРТИРОВКА, УПОРЯДОЧИВАНИЕ). Способ: указание примера, поясняющего суть понятия.

5) Если он корневой, то обозначается так: «<имя диска>:\». (КАТАЛОГ, ДИРЕКТОРИЯ). Способ: указание обозначения, относящегося к понятию.

6) Пара множеств (V, E), где V – конечное множество вершин, а E – конечное множество упорядоченных пар e = (u, v), называемых дугами, где u, v – вершины. (ГРАФ). Способ: дефиниция.

7) По мнению Николауса Вирта, их структура плюс алгоритм равняется программе. (ДАННЫЕ). Способ: использование цитаты.

Многие системы тестирования предлагают использование различных форм тестовых заданий при разработке тестов. Так как в данной работе модель базы тестовых заданий основана на лингвистической модели, то, естественно, что здесь нет смысла использовать задания в закрытой форме. Поэтому программный продукт использует единственную форму тестовых заданий. Конечно, это является определенным ограничением. Но здесь есть и свои преимущества – это, во-первых, использование преимуществ данной формы заданий и данной модели базы тестовых заданий, и, во-вторых, благодаря интерактивным аналитическим отчетам, реализованным в программном продукте, возможность их всестороннего исследования.

Определившись с формой тестовых заданий и моделью представления базы тестовых заданий, необходимо было также выбрать модель процесса тестирования и модель обработки результатов (модель оценивания).

В качестве *модели процесса тестирования* выбрана самая простая модель – это последовательное предъявление вопросов в том порядке, который задан разработчиком теста. Случайная выборка не используется потому, что с точки зрения теории педагогических измерений, случайно выбранные тестовые задания не могут обладать системообразующими свойствами, и, следовательно, не могут составлять тест. Модели, которые используют специфические параметры тестовых заданий, определяемые экспертными методами, не использовались, так как, во-первых, экспертное определение параметров задания во многих случаях отрицательно сказывается на объективности результатов тестирования (например, экспертное определение сложности задания), а во-вторых, для обеспечения «чистоты» используемой модели тестового задания, основанной на понятии фигуры знания. Адаптивное тестирование является в настоящее время одним из самых разработанных методов интеллектуализации тестирования, но для реализации адаптивных моделей необходимо предварительно накопить достаточно большую базу тестовых заданий с эмпирически определенными оценками трудности заданий (причем желательно, чтобы эти оценки были достаточно устойчивыми и эффективными).

В качестве модели обработки результатов как основная была выбрана однопараметрическая модель Раша [2, 3]. Это классическая модель теории IRT, которая обладает рядом важных достоинств:

- модель Раша превращает измерения, сделанные в дихотомических и порядковых шкалах в линейные измерения, в результате качественные данные анализируются с помощью количественных методов;
- поскольку мера измерения параметров модели Раша является линейной, то это позволяет использовать широкий спектр статистических процедур для анализа результатов измерений;
- оценка трудности тестовых заданий не зависит от выборки испытуемых, на которых была получена;
- оценка уровня знаний испытуемых не зависит от используемого набора тестовых заданий;
- пропуск данных для некоторых комбинаций (испытуемый – тестовое задание) не является критическим.

Кроме использования модели Раша, для подсчета тестовых баллов можно использовать простейшую модель, основанную на проценте правильно выполненных заданий.

Для анализа качества теста и тестовых заданий реализованы как основные процедуры классической теории тестирования, так и анализ в рамках модели Раша.

Таким образом, в данной разработке по описанным выше соображениям была зафиксирована определенная модель тестового задания и базы тестовых заданий, определенная форма тестового задания, определенная модель процесса тестирования, модели обработки результатов и оценки качества тестов. Все эти модели в совокупности и сформировали разработанную методику контроля, которая определяет основные виды и способы деятельности разработчика тестов, основные этапы разработки, применения теста и анализа его качества.

Программный продукт позволяет пользователю решать следующие задачи:

1. Создание и поддержка базы тестовых заданий, основанной на модели представления вербальных языковых знаний, с возможностью эффективной навигации и поиска в этой базе.

2. Разработка на основе базы тестовых заданий учебных тестов, как самых простых, ориентированных на задачи текущего, промежуточного контроля, так и профессиональных, обладающих высоким уровнем качества и обеспечивающих представление об истинных баллах учащихся.

3. Проведение тестирования, как индивидуального, так и массового, с высоким уровнем масштабируемости и защиты от фальсификации

результатов тестирования.

4. Аналитическая обработка результатов тестирования, с использованием представления, аналогичного многомерной модели данных, позволяющая устанавливать характер зависимостей между результатами тестирования и параметрами модели тестового задания, например, способом задания смысла, функцией знания, знаком.

5. Оценка качества тестов как в соответствии с классической теорией тестирования, так и в соответствии с теорией IRT.

6. Проведение стандартизации теста и установки норм на основании интерпретации данных обработки результатов тестирования.

7. Получение объективированных оценок уровня подготовки испытуемых и уровня трудности заданий в единой интервальной шкале логитов, в соответствии с моделью Раша, как по приближенным формулам, так и численным методом с заданной точностью.

8. Линейное преобразование шкалы логитов для представления оценок уровня подготовки испытуемых и уровня трудности заданий на множестве положительных чисел.

9. Вычисление оценок испытуемых на основании процента верно выполненных заданий и норм теста, в случае, когда не используется модель Раша.

10. Получение разнообразных форм отчетности, в аналитическом и графическом виде, с возможностью экспорта данных результатов любого отчета в Microsoft Word или Microsoft Excel.

Графически область применения разработки можно изобразить следующим образом:

Система реализована в виде трех автоматизированных рабочих мест – АРМ «Администратор», АРМ «Преподаватель» и АРМ «Студент». Первые два АРМ используют двухзвенную архитектуру «клиент – сервер базы данных» («толстый» клиент), а последнее реализовано в трехзвенной архитектуре «клиент – Web-сервер – сервер базы данных» («тонкий» клиент).

Такое разделение было сделано исходя из того, что массовое тестирование студентов требует именно «тонкого» клиента, когда на рабочей станции не нужно устанавливать никакого специального программного обеспечения. Например, если тестирование должны пройти одновременно 100 пользователей, то при использовании обычной архитектуры «клиент – сервер базы данных» необходимо было бы предварительно установить на 100 компьютеров клиентскую часть СУБД, необходимые библиотеки и настроить подключение к удаленной базе данных. При

использовании «тонкого» клиента те же самые действия нужно сделать только на одной машине – там, где установлен Web-сервер.

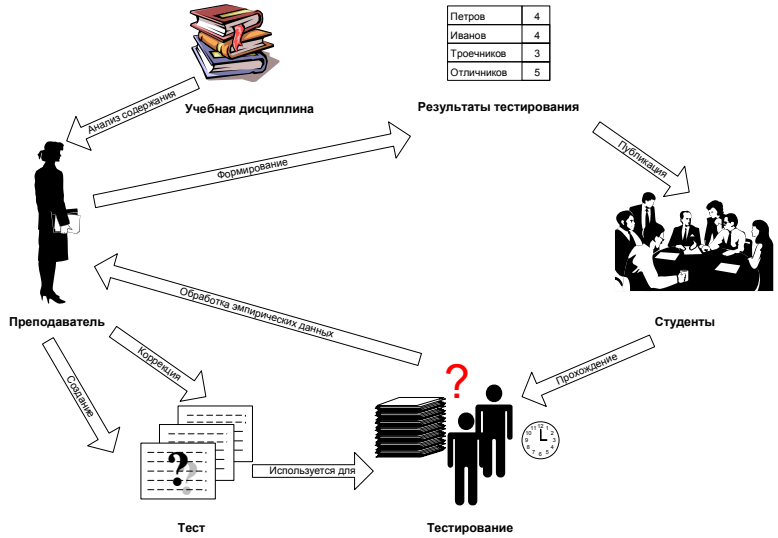


Рисунок 1. Графическое изображение предметной области.

Для АРМ администратора и преподавателя использование тонкого клиента нецелесообразно, так как оба эти рабочих места осуществляют сложное взаимодействие с базой данных, включающее как требующие достаточно большого времени выполнения запросы, так и активное DML-взаимодействие. Кроме того, интерфейс прикладного приложения, то есть в данном случае интерфейс, основанный на стандартах Windows, для постоянной работы все-таки привычней и удобней большинству пользователей, чем интерфейс Web-браузера. Потенциально возможное количество пользователей, использующих эти два АРМ, гораздо меньше, чем для АРМ студента. Вероятно, это несколько преподавателей и один администратор.

Схематически физическая структура системы изображена ниже (на этом же рисунке приведены основные задачи, решаемые пользователями системы):

Таким образом, автоматизированная система контроля знаний, разработанная в дипломном проекте, позволяет обеспечить все необходимые функциональные возможности для разработки тестовых заданий и

тестов, организации проведения индивидуального и массового тестирования, в том числе в системе дистанционного образования, адекватно оценить уровень подготовки учащихся, прошедших тест, оценить качество самого теста, а также отдельных тестовых заданий.

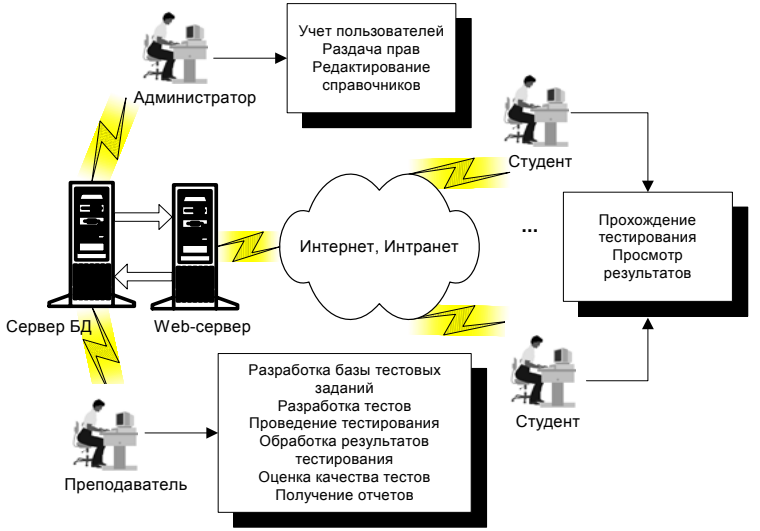


Рисунок 2. Физическая структура системы

Литература

1. *Караулов Ю.Н.* О единицах знания // Полифония образования и англистика в мультикультурном мире. Тезисы первой международной конференции Ассоциации англоведов и преподавателей английского языка 25-26 ноября 2003 г. Москва, МГЛУ, 2003.

2. *Челышкова М.Б.* Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. – 432 с.: ил.

3. *Rash G.* Probabilistick Models for Some Intelligence and Attainment Tests, 1960, Copenhagen, Denmark: Danish Institute for Educational Research.

4. *Атанов Г.А. Пустынникова И.Н.* Обучение и искусственный интеллект или Основы современной дидактики высшей школы. – Донецк: Изд-во ДОУ, 2002. – 504 с.

5. *Аванесов В.С.* Основы научной организации педагогического контроля в высшей школе. М.: Иссл. центр, 1989. – 167 с.

Б.И.Рабинович

СРЕДА ОБРАБОТКИ ПОТОКОВ ДАННЫХ

Введение

Информация — (от лат. informatio — разъяснение, изложение), первоначально — сведения, передаваемые одними людьми другим людям устным, письменным или каким-либо другим способом (например, с помощью условных сигналов, с использованием технических средств и т. д.), а также сам процесс передачи или получения этих сведений. Информация всегда играла в жизни человечества очень важную роль. Однако, в середины 20 в. в результате социального прогресса и бурного развития науки и техники роль информации неизмеримо возросла. Наблюдается лавинообразное нарастание массы разнообразной информации, получившее название "информационного взрыва". В связи с этим возникла потребность в научном подходе к информации, выявлении её наиболее характерных свойств, что привело к двум принципиальным изменениям в трактовке понятия информации. Во-первых, оно было расширено и включило обмен сведениями не только между человеком и человеком, но также между человеком и автоматом, автоматом и автоматом; обмен сигналами в животном и растительном мире [БСЭ].

В современной модели развития любого бизнеса своевременный доступ к необходимой информации является одним из определяющих факторов его развития и успеха. Именно поэтому все большую популярность завоевывают информационные системы, функциональным назначением которых является предоставление пользователям доступа к хранящейся в них информации [Integrum, Garant и т.п.]. Основной особенностью этих систем является либо узкая специализация и привязка к определенной области, либо жесткие требования к входящим потокам данных.

В данной работе рассматриваются методы, позволяющие создать систему, обрабатывающую различные потоки информации, существующей в электронном виде, методы загрузки этой информации в хранилище знаний, некоторые способы поиска и анализа накопленной информации. Система представляет собой среду обработки потоков данных (СОПД). Данные, обрабатываемые СОПД, могут быть следующих типов: текст на естественном языке (ЕЯ), структурированная информа-

ция (таблицы и биллинги), графические образы рукописных текстов, статические и динамические изображения, звуковые ряды

Для содержательной обработки информации предлагается система, основанная на соответствующих методиках обработки текстов. Особенность методик — в переносе сложных этапов лингвистического анализа на уровень обработки знаний и глубине семантического анализа [Кузнецов, 1999].

Система базируется на концептуально-лингвистической модели и методиках, развиваемых на протяжении последних десяти лет в ИПИ-РАН. Уровень полученных результатов сопоставим с передовыми научными исследованиями за рубежом [Fastus, 1996, ЭР3].

Информация, семантика, данные

Основным назначением системы, о которой пойдет речь далее, — это сбор и обработка разнородной информации. Под информацией понимаются любые сведения о каком-либо событии, сущности, процессе и т.п., являющиеся объектом некоторых операций: восприятия, передачи, преобразования, хранения или использования.

Перед тем, как определить понятие «данные», представим следующую абстрактную ситуацию: имеется некоторая система, информация о которой представляет интерес, и наблюдатель, способный воспринимать состояния системы и в определенной форме фиксировать их в своей памяти (никаких других действий наблюдатель не выполняет). В этом случае считается, что в памяти наблюдателя находятся данные, описывающие состояния системы. Итак, данные можно определить как информацию, фиксированную в определенной форме, пригодной для последующей обработки, хранения и передачи.

Данные соответствуют зарегистрированным фактам об объектах или явлениях реального мира. Чтобы в дальнейшем использовать данные, требуется их смысловое содержание – семантика данных. Поэтому в информационной системе должны быть сформулированы правила их смысловой интерпретации.

Работа с семантикой – это работа со знаниями. В системах обработки информации под знаниями понимают сложноорганизованные данные, содержащие одновременно как фактографическую (регистрацию некоторого факта), так и семантическую (смысловое описание зарегистрированного факта) информацию, которая может потребоваться пользователю при работе с данными. Основное средство представления семантики данных – естественный язык. Однако, можно использовать специальные формализованные языки, которые позволяют достаточно эффективно

организовать обработку информации для целого ряда практических задач [Григорьев, 2002].

Автоматизированная информационная система

Основной областью применения СОПД являются автоматизированные информационные системы (АИС), представляющие собой автоматизированную систему управления (АСУ). Основной метод построения такого типа систем является метод регистрации и хранения информации в виде отдельных фактов (значений, событий, операций и т.п.) с последующей их группировкой и объединением по различным признакам (в соответствии с алгоритмами управления в системе). Далее производится вывод итогов в формах и документах, необходимых и удобных для решения конкретных управленческих или пользовательских задач. На сегодняшний день АИС в большинстве случаев разрабатываются как банки данных и знаний.

Банк данных (БнД) – это АИС, включающая в свой состав комплекс специальных методов и средств (математических, информационных, программных, языковых, организационных и технических) для поддержания динамической информационной модели предметной области с целью обеспечения запросов пользователей. Под предметной областью понимается область применения конкретного БнД.

Кроме этого БнД являются составной частью любой АСУ. Соответственно к функционалу БнД, относящемуся к работе с данными, предъявляется ряд требований: хранение; добавление; удаление; модификация; извлечение и т.д.

В общем случае БнД делится на базу данных (БД) и систему управления базами данных (СУБД), где БД отвечает за хранение данных, а СУБД за все остальные функции. Крайне сложно провести четкую грань между двумя понятиями: «знания» и «данные», то же самое можно сказать и про такие понятия, как банк данных и банк знаний (БнЗ). Про БнЗ можно сказать, что это новый тип информационных систем, который нашел широкое применение в различных областях, при этом самое широкое применение находят такие разновидности БнЗ, как экспертные системы (ЭС).

Экспертные системы (ЭС) – это системы, которые создаются для решения определенного рода практических задач, таких как диагностика, прогноз, управление, проектирование, обучение и т.п. Каждая ЭС является человеко-машинной системой. В ее разработке необходимо участие экспертов, инженеров по знаниям и конструкторов. При этом эксперты – это специалисты в данной предметной области, знания кото-

рых требуется ввести в базу знаний ЭС, чтобы в дальнейшем система смогла их использовать при выполнении своих непосредственных функций. Конструкторы – группы специалистов в области АС обработки данных, поддерживающих инструментальный пакет программ, и другие технологические средства, на базе которых создается ЭС.

Основные принципы построения АИС

Перечислим основные требования, предъявляемые к разработчикам, на этапе проектирования АИС.

Масштабируемость. Управление информационным содержанием документов в традиционных и современных форматах сегодня выходит за границы корпоративной сети, давая возможность обмена данными между удалёнными подразделения предприятия, предприятиями-партнёрами и конечными клиентами. Распределённость корпоративного хранилища достигается в системе управления знаниями за счёт особой архитектуры, которая может быть основана как на едином сервере, обрабатывающем все документы территориально распределённых подразделений и подведомственных предприятий, так и на нескольких серверах, расположенных в каждом территориально-распределённом подразделении предприятия, объединённых в единую информационную сеть, с централизованным интернет-доступом к хранилищам данных предприятия.

Современная система управления знаниями должна позволять легко наращивать свои функциональные и качественные возможности. Это означает, в первую очередь, поддержку масштабируемости на системном, аппаратном и программном уровнях. Масштабирование на системном уровне должно обеспечиваться путём изменения крупных функциональных компонентов системы (системы хранения данных, серверы, переход на другие сетевые платформы, использование других систем управления базами данных и т.п.). А масштабирование на аппаратном уровне путём изменения различных аппаратных компонентов системы (процессоров, оперативной памяти, долговременных устройств хранения, телекоммутационных устройств и т.п.). Масштабирование на программном уровне должно обеспечиваться путём изменения состава используемых программных модулей.

Модульность. Система должна представлять собой совокупность аппаратных и программных модулей. Каждый модуль может быть встроен в систему, удалён из системы или модернизирован. Такое реконфигурирование системы не должно оказывать влияния на работоспособность других, входящих в систему модулей.

Интегрируемость. По мере того как крупные компании создавали всё более сложные корпоративные системы, они постепенно пришли к выводу, что безразмерный и статический подход не отвечает современным требованиям управления информационными потоками. Основой реализации системы должен быть единый стандарт обмена данных, позволяющий стыковать к ней уже написанные специализированные модули обработки информации, а также интегрировать с другими информационными системами.

Кроме того, важным аспектом является возможность интеграции с почтовыми системами, офисными приложениями и другими корпоративными информационными системами.

Защита информации на всех стадиях обработки. Вся циркулирующая в рамках системы информация должна быть надёжно защищена на всех стадиях её обработки с целью предотвращения несанкционированного доступа. Кроме того, с позиций современных требований, предъявляемых к разработке пользовательского интерфейса, наличие защиты, а также наличие недоступной для данного пользователя информации не должно быть заметно.

Для контроля доступа пользователей должен присутствовать функционал для мониторинга и получения различных отчётов, как по организации доступа, так и по участию непосредственно в рабочих процессах, чтобы служба безопасности компании могла своевременно выявить узкие места и/или неправильную организацию доступа к данным и рабочим процессам.

В случае наличия дополнительных требований по защите информации должна быть обеспечена возможность шифрования банков данных или применения механизмов электронной цифровой подписи [Губин, 2004].

Структура хранилища данных

После работ по переносу хранилища данных системы из плоских файлов в одну из современных СУБД [Рабинович, 2004] структура хранилища данных приобрела достаточно сложную структуру (Рис. 1). Сервер баз данных (СБД) теперь состоит из баз данных и базы знаний (БЗ) системы. Базы данных состоят из основной базы данных (ОБД) и внешних баз данных (ВБД). ОБД данных служит для хранения поступающих документов, изображений и звуковых файлов. База знаний системы предназначена для хранения значимой информации и связей (содержательных портретов) и обеспечивает эффективный поиск и анализ по связям. ОБД и БЗ системы объединены в один банк знаний (БнЗ), который ориентирован на работу с большими потоками информации (десятки Гб). На основе содержательных портретов строится таблица

индексов, которая тоже хранится в ОБД. Такая таблица состоит из слов (в нормальной форме), выделенных из содержательных портретов, с указанием документов, в которых они присутствуют. На основе этой таблицы, помимо средств СУДБ, обеспечивается быстрый поиск информации, что является краеугольным камнем при решении любой задачи логико-аналитической обработки.

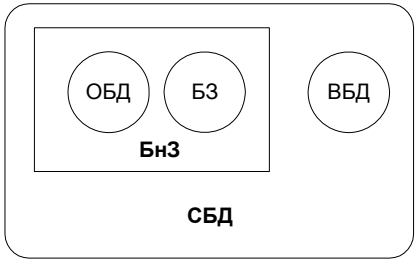


Рис. 1. Хранилище данных СОПД

СБД может быть реализован на любой СУБД, поддерживающей стандарт языка SQL-92. Возможно использование и Oracle, и MS SQL Server, и Sybase, и InterBase или других. Всё зависит от возможностей, требований и традиций конкретных пользователей системы. Обращение к такому СБД осуществляется с помощью SQL-запросов. Если СУБД поддерживает этот стандарт, то принципиальных возражений в использовании той или иной конкретной СУБД нет. Различия будут в том, как быстро и с каким объемом данных она сможет работать, а также в различиях условий эксплуатации (лицензирование, сопровождение и поддержка) различных СУБД. [Мацкевич, 2003].

Кроме этого, одним из важных фактором при выборе конкретной СУБД является сложность настройки функционала системы в рамках конкретной СУБД.

Выбор СУБД

В качестве возможных СУБД были рассмотрены следующие существующие системы: MySQL, MsSQL, Interbase, Oracle. Такой выбор обосновывается требованиями к функционалу СУБД, выполнение которых необходимо для создания СОПД [ЭР2].

Наиболее простой подход при выборе СУБД основан на оценке того, в какой мере существующие системы удовлетворяют основным требованиям создаваемого проекта информационной системы. Более слож-

ным и дорогостоящим вариантом является создание испытательного проекта на основе нескольких СУБД и последующий выбор наиболее подходящего из кандидатов. Но и в этом случае необходимо ограничивать круг возможных систем, опираясь на некие критерии отбора. При этом перечень требований к СУБД, используемых при анализе той или иной информационной системы, может изменяться в зависимости от поставленных целей. Можно выделить несколько групп критериев:

- *Моделирование данных* (Используемая модель данных, Триггеры и хранимые процедуры, Средства поиска, Предусмотренные типы данных, Реализация языка запросов)
- *Особенности архитектуры и функциональные возможности* (Мобильность, Масштабируемость, Распределенность, Сетевые возможности)
- *Контроль работы системы* (Контроль использования памяти компьютера, Автонастройка)
- *Особенности разработки приложений* (Средства проектирования, Многоязыковая поддержка, Возможности разработки Web-приложений, Поддерживаемые языки программирования)
- *Производительность* (Рейтинг Transactions per Cent, Возможности параллельной архитектуры, Возможности оптимизирования запросов)
- *Надежность* (Восстановление после сбоев, Резервное копирование, Откат изменений, Многоуровневая система защиты)
- *Требования к рабочей среде* (Поддерживаемые аппаратные платформы, Минимальные требования к оборудованию, Максимальный размер адресуемой памяти, Операционные системы, под управлением которых способна работать СУБД)
- *Смешанные критерии* (Качество и полнота документации, Локализованность, Модель формирования стоимости, Стабильность производителя, Распространенность СУБД)

Если исследовать выделенные параметры в случае каждой конкретной СУБД, то сравнение двух различных систем является достаточно трудоемкой задачей. Тем не менее, четкий и глубокий сравнительный анализ на основании вышеперечисленных критериев в любом случае позволяет рационально выбрать подходящую систему для конкретного проекта.

Следует отметить, что по существующей практике решение об использовании той или иной СУБД принимает один человек – обычно, руководитель предприятия, а он может опираться отнюдь не на техни-

ческие критерии. Здесь свою роль могут сыграть такие, с технической точки зрения, незначительные факторы как популярность компании-производителя СУБД, использование конкретных систем на других предприятиях, стоимость. При этом последний фактор может трактоваться в двух противоположных смыслах в зависимости от финансового состояния и политики предприятия. С одной стороны, это может быть принцип, – чем дороже, тем лучше. С другой стороны использование пиратской продукции. Очевидно, последний подход чреват коллизиями и не может привести к успеху в долгосрочной работе.

В результате сравнения вышеописанных СУБД по данным критериям выбор пал на СУБД ORACLE по следующим причинам:

1) Данная СУБД хорошо известна разработчику.

2) Плюсы данной СУБД, как и минусы давно известны. Возможности ORACLE полностью удовлетворяют всем требованиям, поставленными перед СУБД в данном проекте. Наиболее известными минусами является дороговизна и сложность в настройке и администрировании. Кроме этого, как уже отмечалось ранее, существует возможность переноса хранилища данных из одной СУБД в другую, например, в бесплатную и простую в администрировании СУБД MySQL.

Описание системы – ядра СОПД

Рассматриваемая система позволяет автоматизировать многие процессы, связанные с обработкой разнородной информации, такой как тексты на ЕЯ, таблицы, статические и динамические изображения, звуковую информацию. Ядром такой системы является База Знаний на Расширенных Семантических Сетях (РСС), управление данными осуществляется СУБД Oracle. В качестве языка манипуляции знаний выступает декларативный язык ДЕКЛ [Шарнин, 1988]. Такое сочетание является очень хорошей базой для построения разного рода программных продуктов.

Система предназначена для областей, где имеют место потоки текстовой информации на естественном языке: сводки происшествий, СМИ, сообщения о новом оборудовании, дефектах, катастрофах, организациях, ценах и др. Ядром системы является лингвистический процессор (ЛП), который обеспечивает:

- автоматическую формализацию текстовой информации на русском языке с выявлением лиц, организаций, промышленных изделий, событий, дат и др., их связей и создание на этой основе собственной базы знаний;

- автоматическое построение каталогов информационных объектов;
- идентификацию информационных объектов с их приведением к одному виду;
- автоматическое заполнение информационными объектами тематических полей Базы Данных (в автономном режиме) [Кузнецов, 2002].

Лингвистический процессор включает в себя лексикографический, морфологический, терминологический и синтактико-семантический анализ.

Морфологический анализ необходим, чтобы избавиться от различных форм написания слов, что облегчает поиск.

Терминологический анализ обеспечивает выделение терминов, а также синонимичные преобразования. Синтактико-семантический анализ осуществляется специальными контекстными правилами и служит для выделения из документа значимых компонент и связей.

Блок лексикографического анализа обеспечивает:

- автоматическое деление текста на самостоятельные части (например, выделение документов из сводок);
- определение начала и конца предложения, а также начала и конца абзаца.

Морфологический анализ обеспечивает приведение слов в каноническую форму. Каждому слову присваиваются признаки, которые делятся на три группы:

- лексические (слово с большой буквы, большими буквами, с точкой на конце или это отдельная буква и др.);
- морфологические (грамматическая категория слова, род, число, падеж для существительных и т.д.);
- семантические (фамилия, имя, отчество и др.).

Количество семантических признаков может увеличиваться — за счет специальных словарей — организаций, стран, городов и др. Само слово в нормальной форме тоже считается признаком.

Блок морфологического анализа основан на обобщенных окончаниях слов. В этот блок введено лишь несколько десятков тысяч слов, из которых специальной программой выделены типовые окончания слов различных грамматических категорий. Благодаря ним обеспечивается морфологический анализ неизвестных слов, что осуществляется с достаточно высокой надежностью. Результатом работы блока морфологического анализа является семантическая сеть (РСС), представляющая пространственную структуру текста. В нее включены слова в нормальной форме с их признаками и указа-

нием их последовательности. Последующая обработка сводится к преобразованию сетей на основе заданных правил.

Терминологический анализ выполняет синонимичные преобразования, расшифровку сокращений, выделение терминов [Кузнецов, 2004].

Знания (предметные и лингвистические) в БЗ представляются в виде структур, которые записываются в нотации семантических сетей, дополненных средствами представления событийных компонент и комплексных связей. В результате образуются РСС.

РСС ориентированы на отображение особенностей семантики ЕЯ — упоминаемых объектов, их связей, а также возможности интеграции множества связанных объектов в один объект, что выражается на ЕЯ в виде форм с отглагольными существительными. Понятие связи рассматривается в широком смысле. Это могут быть не только отношения, но и зависимости. Связанными считаются также объекты, участвующие в одном действии. Группа связанных объектов может быть связана с другой группой, что на ЕЯ выражается в виде глагольных форм с отглагольными существительными.

РСС состоят из однотипных N-арных фрагментов. В каждый из них введена вершина, называемая кодом фрагмента и соответствующая всей представленной в нем информации [Кузнецов, 1986]. Помимо этого, вводится множество "внутренних" вершин, которые порождает сама система по мере необходимости и которые сопоставляются неименованным объектам.

Сети (РСС), представляющие объекты и связи какого-либо документа, образуют так называемые содержательный портрет этого документа. Такие портреты необходимы для обеспечения быстрого и качественного поиска информации по значимым компонентам и связям.

Содержательные портреты (как и документы) шифруются и помещаются в СУБД, ориентированную на большие потоки информации (гигабайты) и обеспечивающую их быстрый выбор — за счет механизмов СУБД. Такие портреты подкачиваются в оперативную память по мере необходимости, образуя активную часть БЗ [Мацкевич, 2003].

РСС образуются в результате прохождения неструктурированной информации через лингвистический процессор. Граф такой РСС представлен на рисунке 1 [Рабинович, 2005].

Для обработки знаний используются специальные инструментальные средства — язык ДЕКЛ. Он состоит из правил, у которых в левой и правой части — семантические сети. Обработка идет за счет управления вызовом таких правил. Как показывает опыт, использование таких средств заметно упрощает построение прикладных интеллектуальных

систем (Спрут, Аналитик, Криминал, ДИЕС и др.).

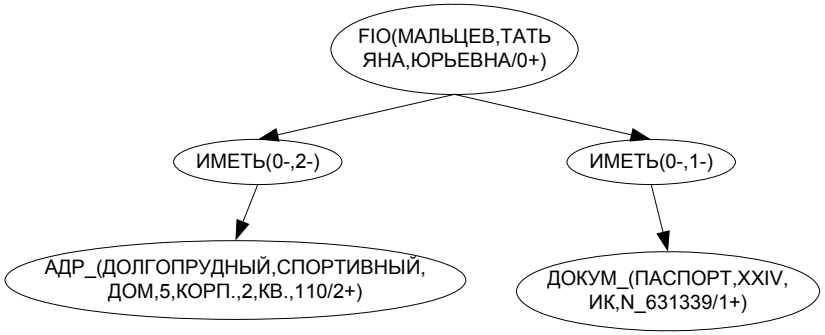


Рис.2. Граф РСС

Структурная схема СОПД
Виды обрабатываемых данных

Данные, которые обрабатывает СОПД, могут быть следующих типов: текст на естественном языке, структурированная информация (таблицы и биллинги), графические образы рукописных текстов, статические и динамические изображения, звуковые ряды. Для каждого типа данных используется свой способ обработки.

Отметим, что для таких типов данных, как графические образы рукописных текстов, таблицы, изображения, звуковые ряды далее описаны существующие методы автоматической обработки, которые не применяются в системе. В СОПД для этих типов данных человек-оператор на стадии загрузки конкретного экземпляра строит его естественно-языковое описание. Соответственно, при загрузке документов изображения, звуковые ряды, таблицы, графические образы рукописных текстов поступают в систему в исходном виде, а их смысловое содержание отражается в ЕЯО. ЕЯО, пройдя через ЛП, преобразовывается в РСС и поступает в БЗ системы. Поиск связей и анализ осуществляется по РСС этих документов.

Кроме перечисленных типов данных, внешними источниками данных могут быть ВБД и Интернет. Результатом обработки всех типов данных являются РСС, которые помещаются в БЗ системы. Исходные данные помещаются в БД системы, кроме этого в БД формируются ин-

дексы, позволяющие значительно ускорить процесс поиска данных в СУБД.

За отображение загруженных данных, работу с аналитическими режимами отвечает блок визуальных интерфейсов.

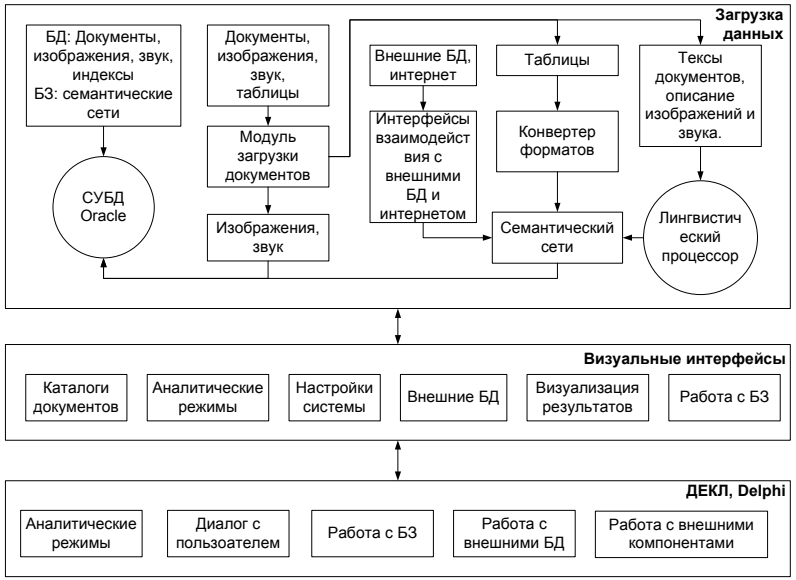


Рис. 3. Структурная схема СОПД

Внутренние механизмы системы представляют собой синтез двух языков: ДЕКЛ и DELPHI. ДЕКЛ осуществляет всю работу с РСС, начиная от извлечения РСС из БЗ, заканчивая их анализом и передачей результатов обработки в модули DELPHI на визуализацию. DELPHI, помимо визуализации, отвечает за взаимодействие со всеми внешними системами (операционная система Windows, СУБД и т.п.), и за передачу данных в ДЕКЛ.

Загрузка данных в хранилище

В общем случае данные поступают на загрузку после предварительной обработки, состоят из исходного вида документа (изображения, звука и т.д.), РСС документа, исходного текста документа, индексов, которые строятся на его основе, и каталогов основных объектов. Опи-

шем структуру БнЗ, который позволил бы хранить все типы данных, описанные выше.

Каталоги представляют собой список основных объектов, по которым проводится быстрый поиск по БЗ. Такими объектами могут быть, например, адрес, телефон, ФИО.

Для ускорения работы в базе отдельно хранится заголовок документа, поскольку чаще всего отображается содержание не документа, а его заголовок, который представляет собой первые несколько слов документа. Исходный текст документа, РСС документа, заголовок, тип документа (текст, биллинг, изображение, звук и т.д.) и исходный вид документа представляют собой единую сущность «Документ». Индексы представляют собой единую сущность «Слово», которая характеризуется самим словом, частотой и списком документов, в котором оно встретилось. Характеристика «частота» используется для того, чтобы показать сколько раз то или иное слово встречается во всей БЗ, соответственно, когда в результате удаления какого-то документа эта характеристика становится равной нулю, это слово можно удалить. Для создания экземпляра сущности «Слово», необходимо при загрузке каждого документа разбирать его, преобразовывать каждое слово в нормальную форму (единственное число, мужской род, именительный падеж для существительных) и уже после этого заполнять соответствующие таблицы.

Представим структуру БнЗ, в котором храниться вся информация:

Сущность	**Атрибут**
Документ	Номер документа
	Текст
	Семантическая сеть
	Заголовок
Слово	Номер слова
	Слово
	Частота
Набор документов	Номер набора
	Номер слова
	Номер документа
Каталог ФИО	Номер ФИО
	Номер документа
	ФИО

Каталог паспортов	Номер каталога паспортов
	Номер документа
	Серия и номер паспорта

Рис. 4. Структура БнЗ

Более подробно структура хранилища и механизмы работы с ним представлены в работе [Рабинович, 2004]. Далее подробнее опишем работу с каждым конкретным типом данных.

Обработка текстов на ЕЯ

Как уже отмечалось ранее, одним из основных компонентов системы является ЛП, цель которого — обработка текстов на ЕЯ. Система выделяет из текстов семантически значимую информацию: интересующие пользователя объекты, их количественные, качественные характеристики и связи. Например, это могут быть конкретные люди, их адреса, телефоны, организации, а также производства с указанием их месторасположения, состава выпускаемой продукции, их количества, качества и т.д. Под связями понимаются отношения (принадлежности, родственные), участие в одном действии, время, место события. Выделение как раз и осуществляется ЛП, который состоит из оболочки, управляемой лингвистическими знаниями [Кузнецов, 2003]. На выходе получается РСС, которая уже рассматривалась ранее. Система ориентирована на обработку сводок происшествий, сообщений средств массовой информации и т.д. Приведем пример обработки текста СМИ.

Текст СМИ: "Багдад. 9 июня. ИНТЕРФАКС/АФП — Два акта саботажа были совершены на нефтепроводах на севере Ирака ..."

Семантическая сеть данного документа выглядит следующим образом:

```
PLACE_(ГОРОД,БАГДАД/1+)
 ДАТА_(9,ИЮНЬ/2+)
 ОРГ_(ИНТЕРФАКС,АФП/3+)
 КОЛИЧ_(2,АКТ,САБОТАЖ/4+)
 PLACE_(СЕВЕР,ГОС-ВО,ИРАК/5+)
 СОВЕРШИТЬ(4-,НА,НЕФТЕПРОВОД,5-/6+)
 ....
```

Безусловно, эффективность распознавания ЕЯ лингвистическим процессором напрямую зависит от тех лингвистических знаний, которые в него вложены. Другими словами, ЛП может эффективно работать только с определенной предметной областью (ПО), на которую он настроен, и при попытке обработать им текст другой структуры (у сводок происшествий есть своя структура) или другой ПО, результат не будет таким «чистым», как это представлено выше. Для настройки ЛП на другую ПО или для улучшения качества распознавания ЕЯ в данной ПО необходимо изменять лингвистически знания.

Более подробно особенности обработки текстов естественного языка на основе технологии баз знаний представлены в работе [Кузнецов, 2003].

Отметим, что текст на ЕЯ часто хранится в печатном виде. Для того чтобы перевести его в электронный вид необходимо воспользоваться специальными программами для распознавания документов. Примером таких программ могут быть FineReader фирмы Abbyy, CuneiForm фирмы Cognitive Technologies, Caere OmniPage. Лучшей среди них по результатам различных испытаний является FineReader [ЭР4]. Каждая из компаний предлагает разработчикам набор специальных библиотек, которые можно встраивать в программные продукты. Недостатком такого решения для разработчиков является то, что все эти библиотеки закрыты для изменений.

В июне 2004 группа сотрудников Санкт-Петербургского государственного университета при поддержке компании Digital Design разработала систему распознавания графических изображений с открытыми исходными текстами. С результатами работы инициативной группы можно ознакомиться на сайте www.ocr.apmath.spbu.ru. Данная разработка далека от функционала и эффективности коммерческих систем, но при необходимости иметь открытый программный код ее достоинства становятся очевидны. Механизмы, применяемые в вышеописанных системах все еще не оптимальны. По всему миру в научно-исследовательских институтах (ИСА РАН, ИПИ РАН, ИПИИ г. Донецк и др.) и в коммерческих предприятиях ведутся разработки новых методов и алгоритмов в данной области [Белозерский, 2005]

Обработка таблиц и диаграмм

В настоящее время диаграммы привлекают внимание специалистов из различных областей знаний, начиная от искусственного интеллекта и обработки изображений до когнитологии и философии. Сформировалось новое направление, известное как диаграмматика или тео-

рия диаграмм. По этой тематике проходят международные конференции и публикуются спецвыпуски журналов [Andersen, 2000], [Hegarty, 2002], [Blackwell, 2001]. Далее представлен краткий анализ тех работ первых двух конференций по теории диаграмм, проходивших в 2000 и 2002 годах, в которых диаграммы рассматриваются и анализируются как знаковые формы представления знаний. При таком подходе к анализу диаграмм и таблиц, которые рассматриваются как частный случай диаграмм, можно наблюдать формирование трёх основных направлений в теории диаграмм:

1) исследование роли диаграмм в среде социальных коммуникаций;

2) распознавание диаграмм при анализе изображений страниц полнотекстовых документов;

3) индексирование и поиск диаграмм в базах данных и электронных библиотеках с учётом сочетания вербальной и образной информационных модальностей, которые им свойственны.

Подробнее остановимся на распознавании диаграмм при анализе изображений страниц документов. Этот процесс является ключевым этапом процесса преобразования документов из бумажной формы в электронную. Важно отметить, что в виде изображений, формул и диаграмм могут быть представлены те концепты, описание которых отсутствует в вербальных компонентах документа. Процесс преобразования документов включает в себя распознавание структуры документа, распознавание линейного текста и кодирование невербальных компонентов, включая диаграммы. Ниже приводится краткий обзор ряда подходов к распознаванию диаграмм при анализе изображений страниц документов в соответствии с работой [Blostein, 2000].

Проблема обработки диаграмм в процессе преобразования документов из бумажной формы в электронную привлекает внимание специалистов уже в течение десятков лет. Главными проблемами в распознавании диаграмм являются: большое разнообразие классов диаграмм и размытость границ классов, подклассов и более детальных уровней классификации; отсутствие средств описания синтаксиса и семантики диаграмм; необходимость обработки графических искажений.

Самое сложное заключается в том, что даже самый верхний уровень классификации диаграмм является весьма условным, так как зависит от построения той или иной системы образных знаков.

В настоящее время иногда используется типология диаграмм на основе предметных областей их использования, например, различаются инженерные, математические, музыкальные, химические и экономические диаграммы. Такой подход предложено называть предметной типо-

логией диаграмм, когда каждый вид относится к той или иной предметной области. Предметная типология диаграмм не отменяет необходимость классификации диаграмм относительно построенных и используемых систем образных знаков.

Важно отметить, что существующие методы распознавания диаграмм, как правило, основаны на распознавании графических примитивов, а не форм образных знаков. Методы распознавания разрабатываются как для нарисованных от руки диаграмм, так и для напечатанных диаграмм. Очевидно, что нарисованные от руки диаграммы достаточно трудно распознавать, так как используемые графические примитивы и их размещение подвержены большему разнообразию, чем в случае с напечатанными диаграммами.

Подавляющее большинство созданных и действующих систем распознавания диаграмм являются исследовательскими или заказными, ориентированными на бизнес-процессы конкретной компании. Основная трудность создания коммерческих систем распознавания заключается в том, что диаграмма в общем случае не является линейным и дискретным текстом. При этом отсутствует общепринятое формальное определение диаграммы. В системах диаграммы определяются с учётом контекста их использования. Поэтому в разных исследовательских системах могут существенно отличаться нотации формального определения диаграммы. С точки зрения определения, хранения и использования нотаций, существующие системы распознавания диаграмм можно разделить на три основные группы.

В первой группе используется программный подход, что говорит о внесении нотаций формального определения диаграммы в исходный код программы, т. е. используемые знания эксплицитно представляются в виде исходного кода, понятного только программисту. Поэтому получаемые в результате проектирования системы распознавания диаграмм этой группы трудно обслуживать и расширять.

Во второй группе используется частично декларативный подход, при котором нотации сначала фиксируются на некотором формальном языке, но полученные спецификации не являются полными. Иными словами, частично знания эксплицитно представляются только в виде исходного кода.

В третьей группе используется декларативный подход, при котором нотации сначала фиксируются на некотором формальном языке и полученные спецификации являются полными. При этом программный код используется как эквивалентная форма записи нотаций.

Описание нотаций формального определения диаграммы может быть достаточно объёмным. Поэтому в исследовательских системах часто реализуют в виде прототипа только подмножество формального определения диаграммы, так как успешная демонстрация прототипной системы является важным этапом в демонстрации жизнеспособности предлагаемого метода распознавания диаграмм. Однако те методы, которые демонстрируют работоспособность в имеющихся прототипных системах, не всегда распространяются на полное формальное определение диаграммы.

Далее проведем краткий обзор методов распознавания диаграмм в исследовательских системах. Это далеко не полный обзор, но он иллюстрирует разнообразие применяющихся методов.

Грамматики графов и графовые преобразования. Графы являются удобной структурой данных для методов распознавания диаграмм. Вершины могут представлять пикселы, примитивы, символы, синтаксические структуры диаграмм. Рёбра графа представляют отношения между вершинами. Графы могут быть обработаны с использованием графовых продукций. Это правила, которые позволяют преобразовывать один граф в другой, удалять подграф и заменять его на другой. Графовый подход применяется для распознавания принципиальных (коммутационных) схем, музыкальных записей и математических формул. Графовые грамматики применяются также для распознавания инженерных чертежей и таблиц. Большинство систем, разработанных на основе грамматики графов и графовых преобразований, являются демонстрационными. Однако модуль распознавания таблиц, разработанный на их основе, стал частью коммерческого продукта [Rahgozar, 1996].

Стохастические грамматики. Сканированные образы диаграмм часто содержат искажения. Они могут содержать, например, области с избытком типографской краски, что вызывает слияние графических примитивов и символов, или области с недостатком краски, что вызывает разрывы графических примитивов и символов. Для применения грамматик в условиях искажённых изображений диаграмм используются различные подходы. Они включают в себя грамматический разбор с коррекцией ошибок, что позволяет найти самое близкое формально корректное интерпретирование. Стохастические грамматики позволяют включать в нотации вероятностные оценки. Эти оценки используются для определения следующего шага грамматического разбора. Стохастические грамматики используются также для анализа и распознавания математических формул,

набранных типографским способом [Chou, 1989].

Методы, основанные на визуальных языках. Специалисты по визуальным языкам разрабатывают средства пространственного описания диаграмм, а также методы их распознавания, редактирования и конструирования, основанные на этих языках. Обзор методов визуальных языков опубликован в [Marriott, 1985]. Визуальные языки позволяет реализовывать грамматический разбор с учётом семантики пространственных отношений. Прежде чем применять эти языки в системах распознавания диаграмм, они должны быть к ним адаптированы. Некоторые из идей построения визуальных языков были использованы для создания системы распознавания рукописных математических формул. Для этого потребовалось расширить их формализм, так как для интерпретации рукописных символов необходимы достаточно сложные пространственные правила.

В заключение этого раздела необходимо отметить, что распознавание диаграмм является сложной проблемой в силу двух основных причин:

- диаграмма в общем случае не является линейным и дискретным текстом;
- отсутствует общепринятое формальное определение диаграммы.

По этим причинам существующие на сегодняшней день исследовательские системы распознавания диаграмм являются, как правило, предметно-ориентированными. Перспективным направлением в распознавании и кодировании диаграмм авторы обзора [Blostein, 2000] считают создание инструментальных средств для построения систем, ориентированных на разные предметные области и разные формальные определения диаграмм [Зацман, 2004].

Из этого можно сделать вывод, что универсальным, но не самым оптимальным, способом обработки диаграмм в СОПД был бы способ, при котором в соответствие каждой диаграмме ставилось бы ее естественно-языковое описание (ЕЯО). Соответственно, при загрузке документов с диаграммами, диаграммы поступали бы в систему в виде изображений, а их смысловое содержание было бы отражено в ЕЯО. ЕЯО, пройдя через ЛП, преобразовывалось бы в РСС и поступало в БЗ системы.

Анализу табличных структур частично посвящена работа [Рабинович, 2005 — Б]. Примером обрабатываемой структурированной информацией могут быть детализации телефонных переговоров и банковских переводов. В СОПД реализовано несколько режимов обработки подобной информации. Детализации телефонных переговоров и банковских переводов (биллингов) имеют в зависимости от компании — автора, в

которой ведется учет звонков или переводов, разные структуры. Эти структуры постоянно обновляются и добавляются, в них часто вводят сложные правила и зависимости, из-за чего практически невозможно создать инструментарий, позволяющий обрабатывать все возможные типы табличных структур.

Звонки с номера 0959222301 за интервал с 01.01.2002 00:00:01 по 06.08.2002 23:59:59

тел. — 0959222301 imsi — Амирасланов Азер Агали оглы лиц. счет — 1746762 Частное лицо

Mob_num	Num	Dur	Dir	Num_A	Cell_Id	
Bts_addr						
0959222301	**+70951079565**	**01.01.02**	**13:30:26**		**168**	**O**
Неизвестно						
0959222301	+70951713495	01.01.02	14:55:12		3	O
Неизвестно						
0959222301	+70951713495	01.01.02	14:55:50		40	O
Неизвестно						

В общем виде биллинги состоят из заголовка и основной части. В заголовке находится информация об интересующем пользователя объекте, и расположена она в произвольном порядке, в связи с чем в работе предложены следующие методы обработки заголовков:

- Распознавание лингвистическим процессором.
- Разбор заголовка таким образом, чтобы в заголовке искались только определенные объекты. Например, поиск телефонного номера осуществлять путем поиска нескольких идущих подряд цифр, поиск ФИО – поиском идущих подряд двух или трех слов, начинающихся с заглавной буквы.
- Разработка интерфейса, который позволил бы пользователю настраивать шаблон для каждого типа биллинга. Причем в этой настройке пользователь бы указывал, где находится тот или иной объект в тексте.

Для анализа основной части необходимо с помощью какого-либо интерфейса создать шаблон, в котором указывалось бы, с какого места в строке начинается то или иное поле, на какой строке заканчивается заголовок, какой разделитель используется для разделения параметров основной части биллинга. Кроме этого необходимо учесть то, что какое-либо поле может принимать специфические значения.

Из этого можно сделать вывод, что создание такого конвертера, который позволил бы осуществлять настройку на любой формат биллинга практически невозможно. Поскольку, во-первых, крайне сложно создать конвертер, обрабатывающий все доступные на сегодняшний день биллинги. А, во-вторых, каждая новая система может генерировать новый формат биллинга, который может иметь существенно отличающуюся структуру.

Поэтому, на сегодняшний день, возможно создать конвертор, который мог бы обрабатывать только фиксированное число входных форматов. На выходе же у него всегда будет готовая РСС, которая будут загружаться в БЗ системы.

Распознавание рукописных графических образов

Исследование методов и механизмов распознавания графических образов, представляющих собой рукописный текст или формы, стало интересовать ученых в 70-х годах прошлого столетия [Кузнецов, 1976; Kolers, 1968]. На сегодняшний день существуют программные комплексы, способные эффективно распознавать графические образы.

Так же есть комплексы, такие как FormReader фирмы Abbyy, UIMA фирмы IBM и др., специализирующиеся на обработке рукописных форм. Особенностью этих форм является то, что для каждого символа написанного от руки отведена специальная ячейка для облегчения процесса распознавания, в то время как при обычном письме слова представляют собой связанные между собой буквы.

Рис. 5. Вид обрабатываемых форм.

По мировой статистике 80% всех документов, использующихся в бизнесе, – это формы. Кроме того, формы используются при опросах и

анкетированиях, т.е. данные вид информации представляет собой один из потоков СОПД. Процесс обработки таких форм в СОПД может выглядеть следующим образом:

1) формы сканируются, т.е. из бумажного вида переводятся в электронный и сохраняются в виде изображения;

2) запускается модуль распознавания рукописных форм, например FormReader (существует возможность встраивания этого модуль в любые разработки);

3) распознанный текст подается на вход ЛП;

4) на выходе из ЛП получается РСС, которая помещается в БЗ системы.

К сожалению, структура таких форм, также как и в биллингах, может быть очень разнообразной. И применение ЛП во многих случаях может быть неоправданно и неэффективно, т.к. по своей структуре формы могут полностью или частично относиться к табличным структурам. Обработка такого вида информации, как формы, существенно расширило бы функционал и сферу применения СОПД.

Самое большое развитие в области распознавания рукописных графических образов на сегодняшний день получили программы распознавания рукописного текста в различных вариантах карманных компьютеров (Palm, Pocket PC). До недавнего времени широкое распространение имели программы, позволяющие распознавать отдельно введенные буквы в специальной области экрана такого компьютера: Graffiti 2 (Palm) и PenReader (Pocket PC). Существенными ограничениями такого рода программ являются большие требования к способу написания всех символов и раздельное их написание. Другими словами, для работы этих программ необходимо писать символы в определенном поле поочередно по жестким и фиксированным правилам.

Не так давно компания Paragon Software (SHDD) усовершенствовала версию системы распознавания рукописного ввода PenReader для Windows Mobile 2003/2003 SE — PenReader 2005 Professional Edition. В данной версии наряду с улучшенной системой графического распознавания рукописного ввода добавлен еще и орфографический модуль, содержащий встроенные словари русского, английского, немецкого и чешского языков. Модуль орфографии позволяет распознавать рукописный текст на новом уровне. В данной версии уже допускается достаточно свободное написание символов

Компания "Кварта Технологии" в 2005 г. выпустила пакет расширения Pocket PC Recognizer Pack, представляющий собой систему распознавания русского слитного рукописного ввода на КПК под управлени-

ем Microsoft Windows Mobile 2003 Second Edition for Pocket PC и старше. Данный продукт основан на технологии распознавания рукописного текста нового поколения riteScript от компании EverNote — одного из лидеров в области разработки средств распознавания образов, в том числе и рукописных текстов.

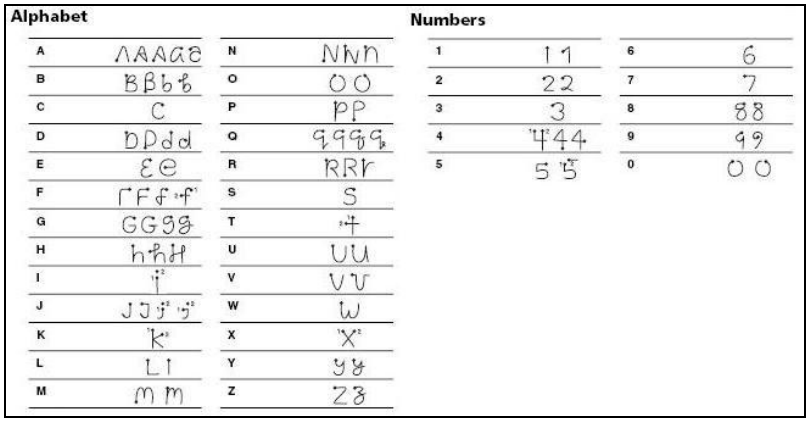


Рис. 6. Правила написания символов в Graffiti 2.

Такая система ввода является качественно новой для российского рынка, так как речь идет о распознавании слитного текста непосредственно в процессе его ввода — представленные ранее решения позволяли работать только с раздельным написанием букв. Второе важное отличие от других подобных продуктов состоит в возможности создавать приложения с интегрированной функцией распознавания рукописного ввода практически для любого почерка. При этом устраняется процесс предварительного “обучения” или “тренировки” пользователей. Технология riteScript позволяет распознавать рукописный текст, вводимый фразами в несколько строк, т. е. в наиболее естественном для пользователей режиме. И еще одно новшество заключается в том, что при подключении созданного “Квартой” механизма к операционной системе Windows обеспечивается его бесшовная интеграция с другими приложениями.

Адаптация технологии riteScript к русскому языку выполнялась совместными усилиями компаний “Кварта” и EverNote. В ходе этой работы формировалась база образцов почерков, создавались словари, отрабатывались алгоритмы лингвистической поддержки распознавания тек-

стов. Технологии этих компаний в области распознавания перьевого ввода широко известны во всем мире и, в частности, используются на лицензионной основе в продуктах Microsoft. Благодаря сотрудничеству с “Квартой” отечественные фундаментальные разработки становятся доступными и в нашей стране [Колесов, 2004].

Программа Pocket PC Recognizer Pack работает на любых устройствах под управлением Microsoft Windows Mobile 2003 Second Edition for Pocket PC и более поздних версиях. Объем занимаемой памяти — около 5 Мб. Продукт полностью интегрирован со всеми стандартными службами операционной системы и встроенных приложений, прост и удобен в использовании. Для ввода текста в этой программе используется специальная панель QPad. QPad позволяет более рационально организовать сам процесс рукописного ввода: текст можно вводить практически непрерывно, без интервалов ожидания результатов распознавания [ЭР5]. Данная разработка после некоторой модификации могла бы с определенной эффективностью распознавать рукописные графические образы в новых приложениях.

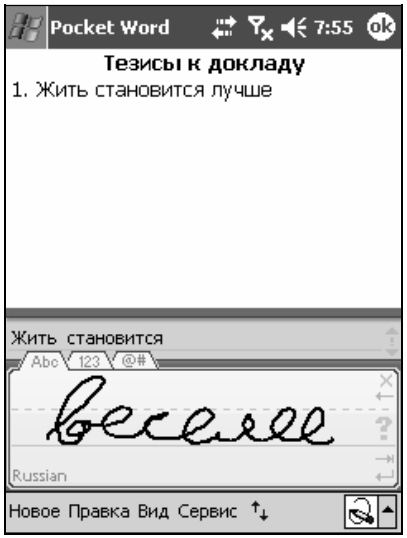


Рис. 7. Слитное написание в программе Pocket PC Recognizer Pack.

Возможность ввода рукописного текста также реализована в Windows XP. Как уже отмечалось ранее, Microsoft использует в своих операционных системах уже готовые решения. Здесь средство распознава-

ния рукописного текста позволяет вводить текст не с помощью клавиатуры, а путем его написания от руки. Текст можно вводить с помощью устройства рукописного ввода, такого как цифровое перо с планшетом, или путем перемещения мыши по коврику (при нажатой основной кнопке мыши) для написания слов. Введенный от руки текст преобразуется в машинописные знаки и вставляется в нужное место. В некоторых приложениях, например в Microsoft Word 2002, допускается вставка рукописного текста в рукописной форме. Рукописная форма отображает текст в точности так, как он был написан от руки. Модули распознавания рукописного текста привязаны к конкретным языкам. В настоящее время имеются модули распознавания рукописного текста Microsoft для четырех языков: китайского (упрощенное и традиционное письмо), английского, японского и корейского. Ведется разработка модулей для других языков.

Распознавание звуковой информации

Распознавание речи является необходимым компонентом интеллектуальных диалоговых систем, начинающих широко внедряться в бытовую технику и постепенно становящихся необходимой составляющей среды обитания человека («умный дом», «умный автомобиль» и т.д.). Процесс распознавания речи представляет собой сравнение входного словарного потока, произносимого на выбранном языке, с моделью данного языка (состоящего из словаря и правил образования предложений). Системы распознавания речи классифицируется по типу входного речевого потока (распознавание изолированных слов и распознаватели слитной речи) и по степени открытости (распознаватели, зависимые от диктора и не зависимые от диктора).

Наиболее сложным для реализации являются независимые от диктора распознаватели слитной речи. Они требуют больших вычислительных ресурсов и огромного объема памяти для хранения базы данных, включающих в себя как модели речевых фрагментов (слогов, слов и т.д.), так и разнообразные правила (грамматические, синтаксические, семантические и т.д.), позволяющие корректно сегментировать входной поток и идентифицировать отдельные слова и фразы. Обычно такие системы распознавания речи реализуются на мощных, высокопроизводительных вычислительных комплексах.

Наиболее простыми для реализации являются настроенные на ограниченный круг дикторов распознаватели изолированных слов. Они требуют меньших вычислительных ресурсов, так как модели слов оказываются менее объемными, а применение правил, позволяющих сегментировать входной речевой поток на слова и фразы, не требуется. Про-

цесс распознавания изолированных слов-команд заключается в сравнении и выборе наиболее близкого по произношению слова. При этом каждое слово хранится в словаре в виде некоторой реализации звуковой модели.

В настоящее время наиболее широко используются последовательности изолированных слов, независимых от диктора. Они реализуют методологию распознавания изолированных слов и по требуемым вычислительным ресурсам занимают промежуточное положение между предыдущими двумя системами [Рождественский, 2004].

Разработки в данной области ведут такие компании как: IBM — ViaVoice [ЭР6], Microsoft — MS Speech Server 2004 [Гуриев, 2004], SUN Microsystems – средства интегрированы в ОС [Гуриев, 2004], Siemens — Speech Sensitive Processing [ЭР7], Spirit [ЭР8] и т.д.

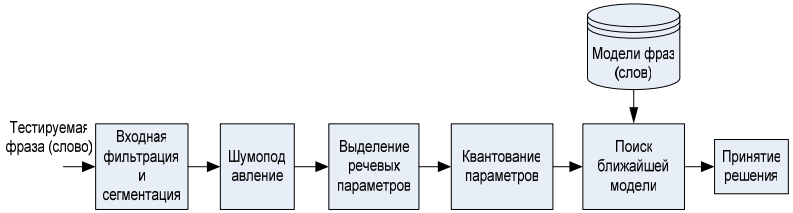


Рис. 8. Блок схема алгоритма распознавания речи

Многие из этих разработок могут быть выделены в отдельные модули с интерфейсами, позволяющими использовать их в разработке новых программных продуктов. Выделяется несколько общих свойств, характеризующих адаптивность модулей распознавания к конкретному пользователю: устройства ввода-вывода, используемые языки, воспроизведение речи и точность распознавания слов.

Кроме этого у некоторых компонентов (например, MS Speech Server 2004) имеется возможность создания нового профиля или настройки профиля на работу с индивидуальным стилем речи. Новый профиль распознавателя следует создавать при смене помещения, изменении уровня шума или в случае, если возле компьютера часто разговаривают другие люди. Профиль распознавателя позволяет разным пользователям совместно работать на одном компьютере, используя при этом разные конфигурации распознавания речи.

Обучение позволяет распознавателю речи адаптироваться к звуку голоса, произнесению слов, акценту, манере говорить и даже к новым словам или идиоматическим выражениям. Это делается с помощью

мастера настройки речевого ввода. Система также адаптируется к индивидуальным особенностям речи во время речевого ввода, т.е. с течением времени распознавание речи улучшается.

Технология распознавания речи используется в различных домашних, офисных и бизнес-приложениях, в которых требуется автоматическое распознавание речи системой: для голосового набора номера в устройствах громкой связи, для ввода PIN-кода для входа в систему, авторизации кредитных карт, работы с голосовыми меню и т.д.

Обработка статических и динамических изображений

В рамках СОПД может быть реализована обработка статических и динамических изображений. В теоретических основах информатики выделен раздел когнитивной компьютерной графики (ККГ), в котором описаны методы обработки изображений.

В предисловии к работе [Зенкин, 1991] известный специалист в области искусственного интеллекта Д. А. Поспелов сформулировал три основных задачи когнитивной компьютерной графики. Первой задачей является создание таких моделей представления знаний, в которых была бы возможность однообразными средствами представлять как объекты, характерные для логического мышления, так и образы-картины, с которыми оперирует образное мышление. Вторая задача — визуализация тех человеческих знаний, для которых пока невозможно подобрать текстовые описания. Третья — поиск путей перехода от наблюдаемых образов-картин к формулировке некоторой гипотезы о тех механизмах и процессах, которые скрыты за динамикой наблюдаемых картин.

В СОПД должны использоваться решения первой задачи ККГ. К сожалению, на сегодняшний день не существует решений, которые позволяли ли бы эффективно решать задачу автоматического построения ЕЯ описания графического образа. Методы ККГ используются в системах безопасности и решают достаточно «примитивные» задачи.

В настоящий момент обработка изображений в системе может быть реализована также как и графических образов рукописных текстов— с помощью обработки их ЕЯ описаний.

Чаще всего, при работе с динамическими изображениями речь идет о видео файлах. Видео представляет собой динамическое изображение со звуковым сопровождением. Т.е. обработка видео изображения представляет собой синтез методов когнитивной компьютерной графики и методов распознавания звуковой информации и представляет собой крайне сложный для реализации процесс.

Описание механизмов поиска дополнительной информации

Существует возможность поиска информации в таких источниках, как внешние базы данных (ВБД) или в Интернете. Для извлечения информации из Интернета была создана система дифференцированного извлечения информации Internet Text Miner (IT-miner).

Это система нового класса, которая ориентирована на широкие круги пользователей сети Internet. Она обеспечивает для них дифференцированный поиск в сети Internet необходимой информации, выделение из нее интересующих пользователя компонент, их содержательный анализ с выдачей пользователю результатов в наиболее удобном и сжатом виде, например, в виде рефератов или форм с заполняемыми полями [ЭР1].

Алгоритм работы системы, построенной с применением технологии IT-miner включает, прежде всего, разработку двух основных компонент.

Первая — система сбора, осуществляющая с помощью типовых средств нахождение соответствующих запросу информационных ресурсов и выделение из них ограниченных фрагментов текстового материала с заданным объемом и достаточно высокой плотностью вхождения ключевых слов. Примером таких поисковых систем могут быть yandex.ru, google.com и т.п.

Вторая компонента — ЛП, на вход которому поступают выделенные параграфы. Соответствующие тексты проходят все виды лингвистического анализа: лексический, морфологический и синтактико-семантический. В результате текст автоматически формализуется. Из него выделяются необходимые информационные объекты: лица, организации, объекты научной, образовательной или производственной или деятельности, даты, места расположения объектов, адреса и т.д., а также их свойства и связи, которые описаны в тексте с помощью глагольных форм или в виде анафорических ссылок. Все это представляется в виде РСС и запоминается в базе знаний.

Внешними могут быть любые доступные информационные БД. Основное требование, предъявляемой к ним, является необходимость их загрузки в ту же СУБД, в которой хранится БЗ системы.

Все базы данных можно условно поделить на базы с простой и сложной структурой. Простая структура БД – это БД, состоящая из одной таблицы, сложная – из нескольких взаимосвязанных таблиц. Рассмотрим в качестве примера подключение простой внешней базы базу МГТС. Структура такой БД выглядит так:

Табл. 2.

Структура таблицы с информацией по физическим лицам.

Телефон
Город
Улица
Дом
Буква
Дробь
Корпус
Строение
Квартира
Фамилия или название организации
И (инициалы)
О (инициалы)
Имя и Отчество (полностью)
Паспортные данные
Комментарии по абоненту
Комментарии по телефону

Табл. 3.

Структура таблицы с информацией по юридическим лицам.

Номер телефона
Принадлежность к организации

Какие шаги мы должны предпринять для подключения такой базы к системе?

1) Эта база должна быть «залита» в Oracle.

2) Необходимо настроить шаблоны взаимодействия между объектами системы и объектами базы МГТС. Перечислим те объекты, которые на данный момент доступны в системе: ФИО, Телефон, Паспорт, Дата рождения, Машина, Адрес.

3) Далее нам надо обеспечить сохранение этих шаблонов в системе.

4) После этого реализовываем механизм запроса информации по шаблонам во внешние базы.

5) Визуализируем результат запроса и/или сохраняем их.

Описанный выше механизм подразумевает загрузку внешних БД в СУБД системы. При этом эти БД должны иметь простую структуру.

Опишем возможный способ взаимодействия с БД, имеющими сложную структуру, загрузка которых в какую-либо СУБД невозможно по той или иной причине. Для подобных БД возможно частное решение, заключающееся в жесткой фиксации структуры БД и способов взаимодействия с нею в коде программы. Подобное решение требует большое количество временных затрат, оно не является универсальным, в отличие от механизма взаимодействия с простыми БД. Это решение является рациональным только в том случае, когда для конкретной области применения СОПД крайне необходимо иметь связь с какой-либо сложной БД этой области.

В общем же случае возможность взаимодействия СОПД с внешними БД и Интернет ресурсами существенно расширяет поисковые и информационные возможности системы.

Возможные области применения СОПД

СОПД можно применять в различных областях деятельности. Частично СОПД внедрена в ГУВД г. Москвы и МВД, где она получила название Логико-Аналитической Системы «КРИМИНАЛ». Соответственно область применения системы – криминалистика. СОПД выполняет следующие функции: обработка текстов на естественном языке, обработка таблиц (биллингов банковских переводов и телефонных переговоров), формирование РСС, формирование индексных файлов, аналитическая обработка, визуализация результатов аналитической обработки.

Кроме криминалистики, весь функционал СОПД можно использовать во многих других областях, например юриспруденция, медицина и т.д. Остановимся подробнее на примере медицины.

Входными потоками данных здесь могут быть тексты на ЕЯ (описание болезней, лекарств, медицинские карты и т.п.), изображения (фотографии операций, проявлений болезней, лекарств и т.п.), видео (учебные пособия, операции и т.п.), таблицы (стоимости лекарств в каталогах, статистика заболеваний и т.п.).

Все эти потоки данных при помощи описанных ранее механизмов загружаются ОБД системы.

На основании этих данных возможна разработка практически любых видов ЭС: энциклопедические системы, диагностические системы, обучающие системы, аналитические системы и т.п.

В качестве примера использования такого рода систем рассмотрим ситуацию, когда какому-либо человеку, не имеющему специального медицинского образования, необходимо узнать всю возможную инфор-

мацию о болезни. В СОПД вводится запрос на ЕЯ, результатом которого могут будут следующие данные: определение болезни, инструкция по лечению, список лекарства, стоимость лекарств, фотографии проявления болезней, иллюстрирующие болезнь видео изображения.

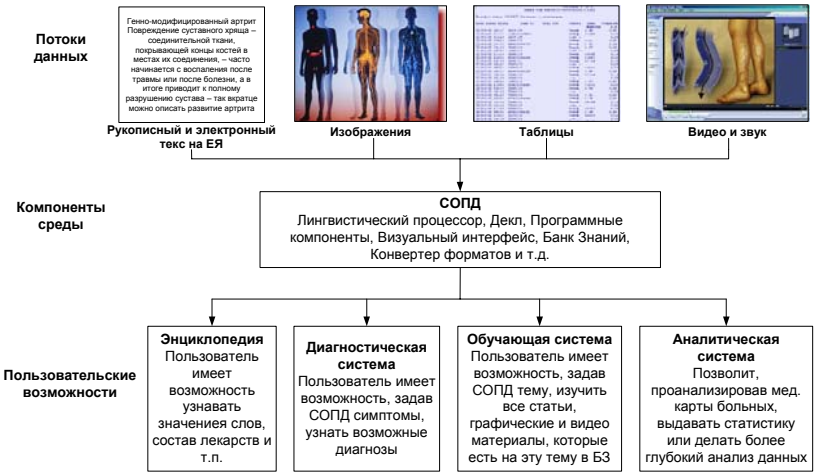


Рис. 8. Применение СОПД в медицинской области.

Заключение

СОПД представляет собой модель системы, позволяющей осуществлять сбор, хранение и обработку таких типов данных, как текст на естественном языке (ЕЯ), структурированная информация (таблицы и биллинги), графические образы рукописных текстов, статические и динамические изображения, звуковые ряды.

В статье представлен обзор существующих методов обработки графических образов рукописных текстов, таблиц, изображений и звуковых рядов, которые возможно применять в системе для автоматической обработки. Применение этих методов в системе носят пока скорее теоретический, нежели практический характер. Они не опробовались в промышленных системах, работающих с большими массивами информации. Их эффективность все еще очень низкая. Кроме этого перечисленные методы значительно повышают требования к производительности и приводят к значительному снижению скорости обработки. Поэтому в рамках СОПД эти методы пока не применяются.

Для этих типов данных (графические образы рукописных текстов,

таблицы, изображения, звуковые ряды) в СОПД человек-оператор на стадии загрузки конкретного экземпляра строит его естественно-языковое описание (ЕЯО). При загрузке документов изображения, звуковые ряды, таблицы, графические образы рукописных текстов поступают в систему в исходном виде, а их смысловое содержание отражается в ЕЯО. ЕЯО, пройдя через ЛП, преобразовывается в РСС и поступает в БЗ системы. Поиск связей и анализ осуществляется по РСС этих документов.

В качестве исходных данных так же описаны Интернет и внешние БД. Одним из ограничений системы является локализация ЛП под определенную предметную область. Для настройки ЛП под другую предметную область потребуется участие в этой работе специалиста по работе со знаниями.

Данная Система может быть полезна во многих областях, где требуется осуществить формализацию информации, обработку текстов на ЕЯ, обработку биллингов. Созданные режимы анализа накопленных знаний позволят решить ряд аналитических задач. Визуализация во всех режимах позволит наглядно отобразить полученные в результате аналитической обработки информации данные [Рабинович, 2002 – А, Б]. Кроме этого, при необходимости, возможно создание новых аналитических режимов, что существенно расширяет область применения Системы.

Литература

Публикации автора

Рабинович, 2003 — *Рабинович Б.И.* Аналитическая система обработки и управления структурированной информацией. Интеллектуальные технологии и система. Выпуск 5. – М.: Издательство ООО «Эликс+», 2003, стр. 284-296.

Рабинович, 2005А — *Рабинович Б.И.* Организация баз знаний в современных СУБД. Проблемы и методы информатики. II Научная сессия Института проблем информатики Российской академии наук. Москва, 18-22 апреля 2005 г. Тезисы докладов. – М.: Издательство, 2005. Стр. 165-168.

Рабинович, 2005Б — *Рабинович Б.И.* Система сбора и обработки разнородной информации «АНАЛИТИК». Интеллектуальные технологии и система. Выпуск 7. – М.: Издательство ООО «Эликс+», 2005, стр. 284-296.

Рабинович, 2004 — Рабинович Б.И. Хранение БЗ в современных СУБД.

Интеллектуальные технологии и система. Выпуск 6. – М.: Издательство ООО «Эликс+», 2004, стр. 173-186.

Рабинович, 2002А — *Рабинович Б.И., Гнидо Е.И., Кузнецов И.П., Мацкевич А.Г.* Временной анализ потоков событий в Логико-Аналитической системе «Аналитик». Научно-техническая конференция профессорско-преподавательского, научного и инженерно-технического состава. МТУСИ. Москва, 29-31 января 2002 г. Тезисы докладов. – М.: Издательство ООО «Инсвязьиздат», 2002. Стр. 409

Рабинович, 2002Б — *Рабинович Б.И., Гнидо Е.И., Кузнецов И.П., Мацкевич А.Г.* Частотный анализ биллингов телефонных переговоров в Логико-Аналитической системе «Аналитик». Научно-техническая конференция профессорско-преподавательского, научного и инженерно-технического состава. МТУСИ. Москва, 29-31 января 2002 г. Тезисы докладов. – М.: Издательство ООО «Инсвязьиздат», 2002. Стр. 409-410.

Используемые источники

Andersen, 2000 — *Andersen M., Cheng P., Haarslev V.* (Eds.) Theory and Applications of Diagrams // Proceedings of the First International Conference, Diagrams 2000, LNAI 1889 (Edinburgh, Scotland, UK. September 1-3, 2000). Berlin. Springer, 2000

Blackwell, 2001 — *Blackwell A.F.* Introduction: Thinking with Diagrams // Artificial Intelligence Review. 2001. V. 15. P. 1-3

Blostein, 2000 — *Blostein D., Lang E., Zanibbi R.* Treatment of Diagrams in Document Image Analysis. Anderson M., P. Cheng, and V. Haarslev (Eds.): Diagrams'2000, LNAI 1889. Berlin: Springer, 2000. P. 330-344

Chou, 1989 — *Chou P.* Recognition of Equations Using a Two-Dimensional Stochastic Context-Free Grammar, Proc. SPIE Visual Communications and Imge Processing IV, Philadelphia PA, 852-863, Nov. 1989

Couchman, 2002 — *Jason Couchman, Ulrike Schwinn,* Oracle 8i Certified Professional DBA – М.: Изд.-во «Лори», 2002 г.

Fastus, 1996 — FASTUS:a Cascaded Finite-State Trasducerfor Ex-

tracting Information from Natural-Language Text. AIC, SRI International. Menlo Park. California, 1996.

Hegarty, 2002 *Hegarty M., Meyer B,. Narayann N.H.* (Eds.) Diagrammatic Representation and Inference // Proceedings of the Second International Conference, Diagrams 2002, LNAI 2317 (Gallaway Gardens, Georgia, USA, April 18-20, 2002). Berlin: Springer, 2002

Kolers, 1968 *Kolers P., Dan M.,* Recognizing Patterns: Studies in Living and Automatic Systems – London: The M. I. T. Press, 1968.

Marriott, 1985 *Marriott K., Meyer B., Wittenburg K.* A Survey of Visual Language Specification and Recognition, in Visual Language Theory / Eds. K. Marriott, B. Meyer. Berlin: Springer, 1985. P. 5-85

Rahgozar, 1996 *Rahgozar M.A., Cooperman R.A* Graph-based Table Recognition System, Document Recognition III, SPIE Proceeding Series. V. 2660. 1996. P. 192-203

Белозерский, 2005 *Белозерский Л.А.* Базовые понятия теории и практики современного распознавания // Искусственный интеллект. — № 2 — Донецк: ИПИИ, 2005

Григорьев, 2002 *Григорьев Ю.А., Ревунков Г.И.,* Банки данных: Учеб. для вузов. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2002. – 320с.

Губин, 2004 *Губин А.В., Краюшкин Д.В., Кузьмин В.В.* Выбор технологии построения системы управления знаниями. Системы и средства информатики/Ин-т пробл. Информатики. Вып. 14. – М.: Изд-во «Наука», 2004 – стр. 145-146

Гуриев, 2004 *Владимир Гуриев.* День открытых систем. // Компьютерра — №35 – М.: ИД «Компьютерра», 2004 г. Стр. 2.

Зацман, 2004 *И. М. Зацман, О. А. Курчавова.* Лингвосемиотический подход к анализу диаграмм. Системы и средства информатики/Ин-т пробл. Информатики. Вып. 14. – М.: Изд-во «Наука», 2004 – стр. 170-185

Зенкин, 1991 *Зенкин А.А.* Когнитивная компьютерная графика/Под ред. Д.А. Поспелова — М.: Наука, 1991. 192с.

Колесов, 2004 — *Колесов А.* Русский рукописный ввод становится реальностью // PCWeek/RE — №9 – М: Издательство ЗАО "СК ПРЕСС", 2004.

Кузнецов, 1976 — *Кузнецов И.П.,* Кибернетические диалоговые системы. – М.: Издательство «Наука», 1976. 293 с.

Кузнецов, 1986 — *Кузнецов И.П.,* Семантические представления. – М.: Издательство «Наука», 1986. 290 с.

Кузнецов, 1999 — *Кузнецов И.П.* Методы обработки сводок с выделением особенностей фигурантов и происшествий. Труды международного семинара Диалог-1999 по компьютерной лингвистики и ее приложениям. Том 2. Протвино. – М.: Издательство «Наука», 1999.

Кузнецов, 2002 — *Kuznetsov Igor, Matskevich Andrey.* System for Extracting Semantic Information from Natural Language Text. Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Том 2. Протвино. – М.: Издательство «Наука», 2002.

Кузнецов, 2003 — *Кузнецов И.П.* Особенности обработки текстов естественного языка на основе Баз знаний. Системы и средства информатики/Ин-т пробл. Информатики. Вып. 13. – М.: Изд-во «Наука», 2003 – стр. 241-250

Мацкевич, 2003 — *Мацкевич А.Г., Кузнецов И.П.* Особенности организации базы предметных и лингвистических знаний в системе АНАЛИТИК. Труды международной конференции Диалог'2003. — М.: Изд-во «Наука», 2003.

Рождественский, 2004 — *Ю.В. Рождественский, Ю.Г. Дьяченко, Н.В. Морозов*. Проблема реализации распознавателя речи на рекуррентном процессоре. Системы и средства информатики/Ин-т пробл. Информатики. Специальный выпуск – М.: Изд-во «Аксон», 2004 – стр. 65-71

Тей, 1990 — *А.Тей, П. Грибомон, Ж. Луи.* Логический подход к искусственному интеллекту: от классической логики к логическому программированию: Пер с франц./Тейз А., Грибомон П., Луи Ж. И др. – М.: Мир, 1990. – 432 с.

Хомоненко, 2003 — *Хомоненко А, Гофман В., Мещеряков Е., Никифоров* В., Delphi 7/Под общ. Ред. А.Д. Хомоненко. –

	СПб.: БХВ-Петербург, 2004 г. Стр. 488
Шарнин, 1988	*Шарнин М.М, Кузнецов И.П.* Продукционный язык программирования ДЕКЛ. В сб. «Система обработки декларативных структур знаний Деклар-2». ИПИАН/СССР — М.: Изд-во «Наука», 1988 – стр. 134-152
ЭР1	Кафедра Вычислительной Математики и Программирования МТУСИ [Электронный ресурс] — Режим доступа: www.mtuci.ru/kaf14, свободный. — Загл. с экрана. — Яз. рус., англ.
ЭР2	Центр Информационных Технологий – Критерии выбора СУБД при создании информационных систем [Электронный ресурс]/ Статья. 2001. Аносов А.– Режим доступа: http://citforum.utmn.ru/database/ articles/criteria; свободный — Загл. с экрана. — Яз. рус., англ.
ЭР3	alphaWorks : Unstructured Information Management Architecture SDK : Overview [Электронный ресурс] – Режим доступа: http://www.alphaworks.ibm.com/ tech/uima; свободный — Загл. с экрана. — Яз. англ.
ЭР4	Тесты и награды [Электронный ресурс]. – Режим доступа: http://www.abbyy.ru/company/?param=4343; свободный — Загл. с экрана. — Яз. рус., англ.
ЭР5	Quarta Technologies [Электронный ресурс]. – Режим доступа: http//www.quarta.com; свободный — Загл. с экрана. — Яз. англ.
ЭР6	Новая система распознавания речи от IBM "к словам не цепляется" [Электронный ресурс]. – Режим доступа: http://itnews.com.ua/20029.html; свободный — Загл. с экрана. — Яз. рус., англ.
ЭР7	Siemens Hearing — Система Распознавания Речи, SSP [Электронный ресурс]. – Режим доступа: http://www.siemens-hearing.ru/client/page.asp?id=261; свободный — Загл. с экрана. — Яз. рус., англ.
ЭР8	Spirit Corp [Электронный ресурс]. – Режим доступа: http//www.spirit.com; свободный — Загл. с экрана. — Яз. рус., англ.

Е.В.Ртищева

АВТОМАТИЗАЦИЯ ПОСТРОЕНИЯ КУРСОВ ЛЕКЦИЙ ПО ПРОГРАММИРОВАНИЮ НА ОСНОВЕ ВЕБ-ТЕХНОЛОГИЙ*

Особенности проведения курсов по программированию и изучению программных пакетов

В настоящий момент в связи с широким распространением Интернет-технологий проектируется множество электронных учебных курсов с использованием этих технологий. Так как большинство программных средств для веб являются бесплатными, а браузер дает возможность просматривать любую информацию на любом компьютере, то наиболее целесообразно использовать их при создании учебных курсов.

Эти учебные курсы применяются:

– при дистанционном обучении

– как методические указания при выполнении практических и лабораторных работ

– как методический материал при чтении лекций и проведении практических занятий.

Естественно ничто не сможет заменить общение преподавателя со студентами, но большое время тратится на различные рутинные задачи, такие как выдача домашних заданий, списков литературы, изложения сведений о стандартах и ГОСТах. А это время могло бы быть потрачено более эффективно на консультации, разбор примеров и т. д. Так же важна оперативность обмена информацией между студентами и преподавателями. [1]

При изучении курсов программирования и других практических курсов, связанных с изучением программных пакетов непосредственное применение полученных данных на практике очень важно и увеличивает усвояемость материала. Обычно такие курсы строятся следующим образом: преподаватель объясняет материал в аудитории на доске, описывает команды, приводит примеры программ. Студенты все это пере-

* Руководитель работы к.т.н., доцент Ревунков Г.И.

писывают в тетрадь. А потом на практических занятиях вводят данные примеры в компьютер и компилируют программы или выполняют последовательности команд при изучении программных пакетов. Все это недостаточно эффективно, так как тратится много времени на переписывание примеров, что чревато появлением ошибок.

Наилучшим способом проведения курсов по программированию является совмещение практических занятий и лекций. То есть преподаватель объясняет какую-либо команду, оператор или функцию, а студенты сразу же выполняют этот пример на компьютере. Можно еще больше сократить время изучения темы, предоставляя студентам материалы лекций и практических занятий в электронном виде. Таким образом, можно изучить гораздо больше материала, разобрать больше примеров, ответить на вопросы студентов.

Но такой способ обучения достаточно дорог, так как каждый студент должен быть обеспечен компьютером, а так как компьютерные классы обычно состоят из 20-30 компьютеров, то потоки, при проведении лекций, надо делить на группы, что увеличивает количество часов отводимых для изучения дисциплины. Этот способ хорошо применять на курсах повышения квалификации и переобучения, а так же на втором высшем образовании по специальностям связанным с информатикой и вычислительной техникой. Так как группы там небольшие до 30 человек, обучение платное и поэтому степень оснащения компьютерных классов выше.

Но существует одна проблема: не всякий преподаватель даже эксперт в области программирования может знать веб-технологии на таком уровне, чтобы свободно ими пользоваться, создавая электронные курсы. И, кроме того, информация в этом курсе может очень быстро меняться, например, если преподаватель вводит новые примеры или изменяет старый или модифицирует пояснительные тексты в соответствии с пожеланиями студентов.

Следовательно, нужна система построения и управления курсом, которая бы позволяла:

– эффективно и быстро строить лекцию на основе шаблона без знания веб-технологий;

– быстро модифицировать примеры и тексты;

– иметь систему поиска или оглавления, с помощью которого можно быстро получить доступ к любому разделу курса из другого раздела.

Дополнительно можно реализовать систему таким образом, чтобы на сервере находились все читаемые курсы, и студенты обращались к ним

во время лекций, через локальную сеть кафедры и во время выполнения домашних заданий через Интернет.

То есть система управления курсами должна поддерживать возможность создания нескольких курсов несколькими разными преподавателями.

Модель контента курса лекций

Учебный курс состоит из определенного количества занятий. Эти занятия могут не являться лекциями или семинарами в чистом виде, так как там происходит смешение теоретического и практического материала, потому что после изучения некоторого теоретического элемента сразу же следует его практическое применение.

Система состоит из клиентской и администраторской частей.

В клиентской части при загрузке главной страницы появляется список читаемых курсов, при выборе курса появляется его оглавление (список всех занятий). Далее можно перейти на любое занятие. Каждое занятие является веб-страницей или множеством страниц, если контент занимает более 3-4 экранов, то его целесообразно разбить его на несколько страниц. В каждом занятии есть темы, примеры, практические задания для них тоже можно выделить отдельную страницу, если это необходимо.

В администраторской части преподавателю дается возможность после авторизации создавать, удалять и модернизировать учебные курсы. Каждый преподаватель может делать это только со своим курсом. Администратор имеет права на все курсы.

Выделим информационные сущности учебного курса. Под *информационной сущностью* будем понимать некий атомарный, логически завершенный элемент контента, объединяющий в себе атрибуты, которые описывают свойства данной сущности: преподаватель, курс, занятие, тема, пример, задание, литература (может быть несколько разных списков литературы: для темы, для занятия и для курса).

Выделим связи между объектами: тема *содержит* пример, преподаватель *читает* курс, курс *содержит* список литературы, тема *содержит* список литературы, занятие *содержит* список литературы. Определим атрибуты для сущностей (таблица 1).

Таблица 1.

Сущности и их атрибуты.

Сущность	Атрибуты
преподаватель	фамилия имя отчество контактные данные читаемые курсы

	имя пользователя пароль
Курс	название курса краткое описание оглавление курса данные преподавателя
занятие	название занятия оглавление занятия
тема	название теоретический материал порядок следования в структуре занятия номер примера и/или практического задания
пример	номер название содержание
задание	номер название содержание
литература	номер название содержание

Можно добавить еще некоторое количество атрибутов, например для "темы", номер списка литературы, практического задания.

На рисунке 1 показана структурная схема предметной области.

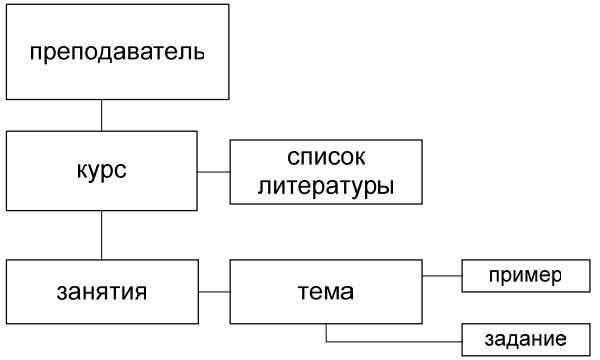


Рисунок 1. Структурная схема предметной области "курс".

Выделим ассоциирующие сущности. *Ассоциирующая сущность* – поименованная совокупность экземпляров семантических связей между данной сущностью и другими сущностями.

Выделим веб-страницы двух видов: информационные и индексные. *Информационные страницы* являются множеством экземпляров информационных сущностей (совокупности атрибутов с установленными значениями) и ассоциирующих сущностей, а *индексные страницы* являются множеством ассоциирующих сущностей.

Ассоциирующие сущности интерпретируются: как атрибуты информационных сущностей, как индексные страницы, как подмножество информационной страницы. В рассматриваемой модели некоторые ассоциирующие сущности являются атрибутами информационных: оглавление курса, оглавление занятия. А некоторые являются отдельными сущностями: список преподавателей, список курсов.

Пример ассоциирующей и информационной сущности показан на рисунке 2.

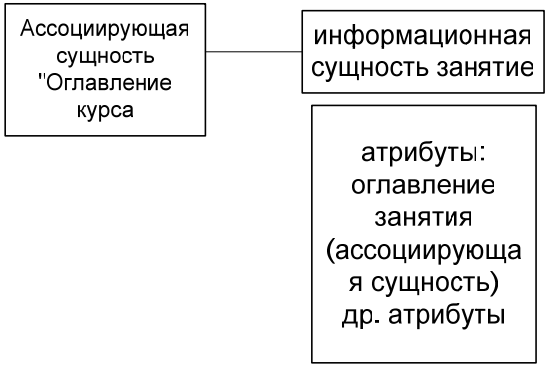


Рисунок 2. Ассоциирующая и информационная сущность.

Экземпляры связей являются гиперссылками. Входом гиперссылки является имя связи, выходом – экземпляр сущности.

В зависимости от желания разработчика этот список можно переделать, так, например, из сущности "Курс" можно убрать оглавление и таким образом создать новую ассоциирующую сущность "Оглавление курса". А сущность "Список курсов" можно сделать информационной, добавив в нее следующие атрибуты: название курса, краткое описание, данные преподавателя.

Отображение информационной модели на структуру страниц электронного курса

На сервере физически электронный курс является веб-сайтом и состоит из системы файлов, баз данных и программ, которые представляют собой различные веб-документы. [2]

Для преобразования информационной модели в электронный курс необходимо: отобразить информационные сущности в шаблоны страниц; отобразить информационные сущности в структуру базы данных; отобразить ассоциирующие сущности в панели навигации, оглавления и отдельные гипертекстовые ссылки.

Построение шаблонов страниц

С точки зрения модели шаблонов страниц электронный курс рассматривается как совокупность графических, текстовых и мультимедийных элементов. Шаблон задает положение элементов на странице, и их внешний вид (шрифт, цвет, размер и т. д.), а также внешний вид страницы.

Задача автоматизированного построения страницы состоит в следующем: нужно по вводимой пользователем текстовой, графической и мультимедиа информации построить страницу, содержащую введенную пользователем информацию и оформленную в соответствии с пожеланиями пользователя [3]. При реализации данной задачи человеком пользователем должны выполняться следующие задачи:

- выбор шаблона страницы;
- заполнение необходимых полей ввода данных (закачка рисунков, мультимедиа);
- доработка макета страницы, изменение, при необходимости, цветовой схемы, шрифтов, расположения объектов.

Автоматически в реализации этой задачи выполняются следующие операции:

- формирование страницы на основе выбранного пользователем шаблона и введенной информации;
- формирование цветовой схемы страницы;
- формирование шрифтовой схемы страницы.

Каждая страница строится на основе шаблона, а шаблон строится на основе сущности, то есть в шаблоне "Пример" обязательно должны присутствовать блоки, в которых выводятся "название примера" и содержание примера. Та часть содержания "Примера", которая содержит

программный код, выводится определенным образом: другим шрифтом или с подсветкой определенных элементов.

Следовательно, в шаблоне есть обязательная часть, с помощью которой выводятся экземпляры сущностей контента данной страницы, и есть необязательная, в которой задается расположение блоков и их формат. Поэтому можно говорить о *семействе шаблонов*, в которых обязательная часть одна и та же, а необязательная различается.

В системе построения электронных курсов могут существовать следующие семейства шаблонов: список курсов, оглавление курса, занятие, тема, пример, задание, список литературы.

Кроме того, в системе существует специальный модуль определения внешнего вида страниц, где задается цветовая и шрифтовая схемы всего электронного курса и отдельных страниц. А так же выбирается шаблон из семейства шаблонов отображения сущности.

Для индексных страниц существуют специальные семейства шаблонов, которые отображают в себя ассоциирующие сущности, их внешний вид тоже определяется разработчиком.

Если ассоциирующая сущность является атрибутом информационной или входит в состав информационной страницы, то в шаблоне предусматривается специальный блок для отображения гиперссылок.

Реализация системы

При реализации системы выполняются следующие действия: строится база данных, проектируется администраторский интерфейс, проектируются шаблоны, разрабатываются программные модули загрузки и обновления информации на странице.

При проектировании базы данных из информационных сущностей получаем реляционную модель базы данных, причем атрибуты сущностей модели контента не обязательно будут совпадать с атрибутами сущности реляционной модели.

Можно объединить информационные сущности "Пример", "Тема", "Задание" в одну сущность реляционной модели (рис. 3) .

Еще одним вариантом реализации является возможность добавления атрибутов к сущностям реляционной модели при создании курса. То есть у каждой сущности имеется базовый набор атрибутов, а разработчик добавляет свои, изменяя, таким образом, таблицы базы данных.

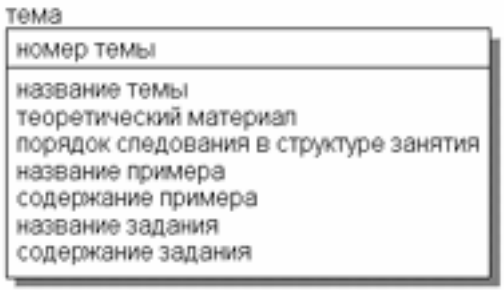


Рисунок 3. Сущность "Тема" в реляционной модели.

Заключение

В работе рассмотрены этапы построения системы управления электронными курсами, в соответствии с особенностями преподавания. Так же определена модель контента предметной области, которая может быть дополнена и расширена, если это требуется. В соответствии с моделью контента рассмотрен способ отображения сущностей в шаблон страницы. А так же предложен вариант реализации системы.

Разработанная модель контента может быть применена при разработке электронных курсов в других областях образования, таких как экономика, естественные, науки, гуманитарные науки и т. д.

Литература

1. *Виноградова М.В.* Использование интернет-технологий для автоматизации учебного процесса в очных вузах // Интеллектуальные технологии и системы: Сборник статей аспирантов и студентов. Сост. и ред. Ю.Н. Филиппович – М., 2001. – Вып. 3.– С. 159-181.

2. *Терно. Ю. А.* Построение информационной модели WEB-сайта на основе семантического структурирования. http://joker.botik.ru/show_thesis.php3? year=2000&name=terno

3. *Таль Н.В. Волков А.Л.* Автоматизированное проектирование визитных карточек // Интеллектуальные технологии и системы: Сборник статей аспирантов и студентов. Сост. и ред. Ю.Н. Филиппович — М, 2001. – Вып. 3.– С. 130-141.

А.А.Самочетов

ДИСТРИБУТИВНО-СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ

Все представленные в реферативном обзоре публикации были разделены на две группы. В первой группе “Методы и алгоритмы дистрибутивно-статистического анализа текстов” представлены статьи, основной темой которых является описание самого метода дистрибутивно-статистического анализа, а также сравнение различных алгоритмов его применения. Во вторую группу “Использование дистрибутивно-статистического метода и тезаурусов в автоматизированных информационных системах” вошли те статьи, которые в большей степени освещают вопросы о том, как результаты дистрибутивно-статистического метода могут применяться в АИС. Во вторую часть реферативного обзора также вошли статьи связанные не только с использованием результатов дистрибутивно-статистического метода, но и статьи, посвященные вопросам применения тезаурусов, так как дистрибутивно-статистический метод является одним из возможных методов построения тезаурусов.

Методы и алгоритмы дистрибутивно-статистического анализа текстов

Москович В.А. **Дистрибутивно-статистический метод построения тезаурусов: современное состояние и перспективы. Ч.1 и Ч.2 // Научно-техническая информация, сер.2, 1972, №3, с.12-22, №4, с. 15-25.**

В начале статьи дается общий обзор работ, посвященных проблеме дистрибутивно-статистического анализа текстов (до начала 70-х годов), освещается также история вопроса, дается оценка сделанным достижениям в этой области. Далее приводится описание различных методик дистрибутивно-статистического метода, отличающихся использованием разных коэффицентов связи между словами. Вместе с описанием коэффициентов даются также результаты сравнительного анализа, позволяющие сделать вывод в пользу той или иной формулы. Отдельный большой раздел посвящен интерпретации связей между словами, выявляемых дистрибутивно-статистическим методом. В этом разделе выделены и показаны на многочисленных примерах различные типы синтаг-

матических и парадигматических отношений между словами, которые могут быть выявлены на основе дистрибутивно-статистического метода. Там же вводится понятие интервала текста, определяется оптимальная его длина для извлечения наиболее ценной информации из текста. Отдельно в статье также освещены частные вопросы дистрибутивно-статистического метода – разметка текста, группировка слов в семантические поля, использование дистрибутивного метода для выявления парадигматических связей. В конце статьи приводится описание практического использования дистрибутивно-статистического метода на примере реально действующих автоматизированных информационно-поисковых систем и делается заключение о перспективности использования этого метода для построения тезауруса и о возможности автоматизации этого метода с помощью ЭВМ.

Шайкевич А.Я. **Дистрибутивно-статистический анализ в семантике. // Принципы и методы семантических исследований. М., 1976, с. 353-378.**

В статье автор выделяет три основные области изучения семантики, для которых может использоваться дистрибутивно-статистический анализ текстов – создание процедур семантического открытия, исследование макросемантики, семантические системы языков прошлого. Для исследования семантики автор предлагает использовать следующие формальные статистические методы: исследование распределения элементов в тексте в целом, исследование распределения элементов в тексте с учетом позиции занимаемой элементом в тексте (позиционный анализ), исследование взаимного распределения элементов в тексте (собственно дистрибутивно-статистический метод). Далее дается описание каждого статистического метода, приводятся многочисленные примеры. Для взаимного распределения элементов в тексте (в том случае, если частоты появления элементов достаточно малы) делается предположение, что распределение может соответствовать распределению Пуассона. На основе распределения Пуассона предлагается оценка коэффициента связи между элементами текста. В статье также рассматривается ряд вопросов связанных с изучением семантической информации, получаемой на разных интервалах текста, а также с исследованием семантических карт.

Н.С. Иванова. **К вопросу об автоматическом построении тезауруса. // Научно-техническая информация, сер.2, 1969, №6, с. 17-20.**

В статье рассмотрена возможность построения тезауруса на основе статистического метода, который основан на вычислении коэффициента совместной встречаемости некоторых слов в заданном интервале текста

по некоторой формуле. Сравниваются результаты применения двух статистических методов для построения тезауруса – метода, предложенного Кембриджским отделом лингвистических исследований, и метода, предложенного Шайкевичем. На одном и том же текстовом материале (тексты патентных формул) проводились расчеты по разным коэффициентам совместной встречаемости и на разных интервалах текста. По результатам исследования можно сделать вывод, что для построения тезауруса метод, предложенный Шайкевичем, оказался более предпочтительным.

Маршакова И.В. **Построение информационно-поискового тезауруса методом дистрибутивно-статистического анализа. // Научно-техническая информация, сер.2, 1977, №5, с. 11-15.**

В статье рассмотрен пример для построения информационно-поискового тезауруса, выполненного на основе дистрибутивно-статистического метода, предложенным Шайкевичем. Эксперимент проводился по текстам рефератов (по тематике “Газовые горелки”) на основе двух интервалов – минимального (одно слово влево и одно вправо от исходного слова) и среднего – (от 50 до 500 слов в тексте). При проведении эксперимента для получения списка устойчивых словосочетаний использовался статистический метод, основанный только на подсчете совместной встречаемости двух слов, а для выявления парадигматических связей использовался дистрибутивный метод, основанный также на учете данных о встречаемости данных слов с другими словами в заданном интервале текста. По результатам исследования были получены частотный словарь ключевых слов, список устойчивых словосочетаний, статьи информационно-поискового тезауруса, семантические поля. Примеры результатов исследования также приводятся в статье и в приложении. В заключении статьи делается вывод о возможности использования дистрибутивно-статистического метода для построения тезауруса (и, как следствие, использования тезауруса в различных поисковых задачах), а также о целесообразности автоматизации алгоритма дистрибутивно-статистического метода с помощью ЭВМ.

Пекар В.И. **Дистрибутивная модель сочетаемостных ограничений глаголов. // Материалы конференции «Диалог-2004. Компьютерная лингвистика и ее приложения». Под ред.Нариньяни А.С. Москва: Наука, 2004. - http://www.wlv.ac.uk/~in8113/papers/dialog04_pekar.pdf**

В статье рассматривается возможность применения дистрибутивно-статистического метода для определения сочетаемостных ограничений глагола. Предлагаемый алгоритм дистрибутивно-статистического метода состоит в выделении из корпуса текстов существительных, наиболее

близких по некоторому подсчитанному коэффициенту к тому существительному, которое употреблялось при исследуемом глаголе. Далее на основе полученного класса существительных и найденных коэффициентов можно было оценить вероятность совместимости глагола с каждым из этих существительных. Выделение существительных, принадлежащих одному классу, выполнялось по дистрибутивному методу: вначале по некоторой формуле из корпуса текстов определяются признаки для данного существительного (другие слова, употребляемые при данном существительном), затем по некоторой метрике дистрибутивной схожести определялась близость существительных по полученным признакам. Статья в основном посвящена сравнению различных формул, используемых для выделения признаков некоторого существительного. Для этой цели в статье предлагались такие статистические формулы, как “хи-квадрат”, точный тест Фишера, Gain Ratio. По результатам проведенного эксперимента делается вывод, что из предложенных формул для отбора признаков точный тест Фишера является наиболее предпочтительным.

Использование дистрибутивно-статистического метода и тезаурусов в автоматизированных информационных системах.

Королев Э.И. **Применение дистрибутивно-статистического метода в лингвистическом обеспечении автоматизированных информационных систем. // Научно-техническая информация, сер.2, 1977, №1, 27-31.**

В статье рассматриваются вопросы, связанные с возможностями использования дистрибутивно-статистического метода на разных этапах построения и эксплуатации информационных систем. В статье выделены три аспекта возможного использования дистрибутивно-статистического метода для ИПС, причем основное внимание уделено первым двум – для построения тезаурусов, для расширения поискового предписания за счет ассоциированных терминов, для объединения документов, содержащие ассоциированные по смыслу термины. В статье дается определение дистрибутивного и статистического метода, определяются достоинства и недостатки этого метода. По результатам ряда проведенных работ, на которых ссылается автор, делается вывод об ограниченности возможностей дистрибутивно-статистического метода для решения поставленных задач. Тем не менее, в статье также отмеча-

ется, что дистрибутивно-статистический метод в принципе может использоваться для отладки и пополнения тезаурусов, которые были предварительно построены специалистами, а также для расширения поискового предписания, хотя и не превращает этот процесс в полностью автоматизированную процедуру.

Пекар В.И. **Автоматическое пополнение специализированного тезауруса. // Материалы конференции «Диалог-2002» т.2. Прикладные проблемы, М. 2002, с. 413-421.**

Статья посвящена проблеме автоматического пополнения тезауруса новыми лексическими единицами, извлекаемыми из текста. Рассматриваются два алгоритма классификации – метод ближайших соседей и метод, основанный на учете таксономической организации тезауруса. Оба алгоритма для пополнения тезауруса новым словом используют в той или иной степени дистрибутивный метод для определения семантической близости нового слова по отношению к одному или нескольким членам синонимической группы в тезаурусе. Основное отличие алгоритмов состоит в том, что второй алгоритм, в отличие от первого, дополнительно также учитывает отношения между отдельными классами в тезаурусе (например, связанными отношениями род-вид). В статье также описан эксперимент по классификации слов и приведены его результаты, которые показали, что второй алгоритм в ряде случаев дает более точные результаты, чем первый.

Браславский П.И. **Автоматические операции с запросами к машинам поиска Интернета на основе тезауруса: подходы и оценки. // Материалы конференции «Диалог-2004. Компьютерная лингвистика и ее приложения». Под ред.Нариньяни А.С. Москва: Наука, 2004.** - http://www.dialog-21.ru/Archive/2004/Braslavskij.htm

В статье рассматриваются предпосылки использования автоматических операций с запросами к машинам поиска (МП) Интернета на основе тезауруса. В рамках предложенного подхода описаны операции перевода запроса, расширения запроса на основе шаблона, построения запроса на основе пути между двумя концепциями, а также ослабления запроса. Приведены примеры запросов, сформированных на основе тезауруса предметной области «Автоматический оптический контроль печатных плат», и соответствующие отклики МП Яндекс и Google. Обсуждаются результаты и даются рекомендации по практическому применению предложенных методов.

Харин Н.П. **Поиск документов по запросу, расширенному автоматически построенными парадигматическими отношениями. // Новости искусственного интеллекта, 2003, №6**

В статье рассматриваются различные способы повышения эффективности поиска: нечеткий поиск (широко используемый в Интернет); ассоциативный поиск, когда отношения между терминами являются зафиксированными в тезаурусе. Предлагается новый способ поиска – динамический ассоциативный поиск, когда ассоциативные отношения устанавливаются не заранее, а после выдачи части документов, наиболее соответствующих запросу. После получения части документов, наиболее релевантных запросу, в них, на основе статистики совместной встречаемости, определяются слова, наиболее близко и наиболее часто используемые со словами запроса. В дальнейшем, в процессе динамического ассоциативного поиска, предполагается, что эти слова также будут участвовать в поиске новых документов. В статье также описывается эксперимент с проведением обычного (нечеткого) поиска и динамического ассоциативного поиска в заданном массиве документов. На основе результатов эксперимента делается вывод о целесообразности использования автоматически построенных ассоциативных отношений в системах с нечетким поиском.

Ермаков А.Е., Плешко В.В. **Семантическая сеть в глазах аналитика. // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов - Москва, 2002. - http://www.rco.ru/article.asp?ob_no=25#pic1**

Статья посвящена вопросам анализа и визуализации содержания текста с помощью семантической сети. Дается определение семантической статьи, выделяется ряд задач, которые аналитик может решить при помощи семантической сети – исследование тематического состава большого количества (коллекции) документов и выделение ключевых тем, поиск новой информации, связанной с темой, выявление новых скрытых связей, которые не были зафиксированы в семантической сети. В статье также приводится обобщенное описание интерфейсных средств визуализации семантических сетей, даются примеры того, в каких случаях семантическая сеть может заменить или дополнить содержание полной коллекции документов. В заключение отмечаются перспективы возможного дальнейшего развития технологий семантических сетей, связанных с фактографическим поиском.

А.Ю.Филиппович, А.М.Серебряков

ГЕНЕТИЧЕСКИЕ АЛГОРИТМЫ. ПРИМЕР РЕАЛИЗАЦИИ И ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ*

Постановка задачи

В рамках проекта «Автоматизированная система научных исследований динамики ассоциативно-вербальной модели языкового сознания русских как индикатора образа России в новейшей истории и современности» при поддержке гранта РГНФ № 06-04-03803в создается подсистема «Анализа ассоциативно-вербальной сети (АВС)». В нее входит модуль «Поиска ассоциативных цепочек», который предназначен для исследования связей и поиска путей между ассоциациями с помощью различных методов. АВС представляет собой взвешенный граф, вершинами которого являются слова естественного языка (стимулы и реакции), дугами – различные связи, которые возникают между ними, а весами дуг - частота проявления связи. На первом этапе исследования топологии сети можно отказаться от анализа типов связей и их направленности, упростив тем самым математические и программные методы. Для простоты расчетов также будем обозначать конкретные стимулы и реакции с помощью числовых идентификаторов.

В данной статье будем рассматривать задачу поиска кратчайшего пути между двумя вершинами АВС. Например, для графа на Рис 1. кратчайшим путем из вершины 1 в вершину 5 является путь через вершину 4. Рядом с дугами графа указаны стоимости прохождения по этой дуге.

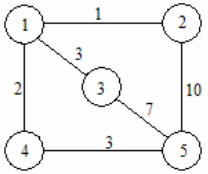


Рисунок 1. Взвешенный ненаправленный граф.

* Статья подготовлена совместно с научным руководителем — А.Ю.Филипповичем.

Решать данную задачу будем с помощью *генетических алгоритмов*. Рассмотрим систему для решения подобных задач, разработанную в среде Delphi (Рис 2.)

Ниже будут рассмотрены примеры на языке Object Pascal, с помощью которых можно будет реализовать систему, использующие генетические алгоритмы. Будут проведены исследования по оптимизации ГА, насколько зависит эффективность ГА от той или иной реализации *генетических операторов*, от разных значений вероятностей из применения. Так же рассмотрим насколько эффективно решение данной задачи с помощью ГА при тех или иных начальных условиях. Но сначала опишем ограничения, которые накладываются на данную задачу рассматриваемой системой.

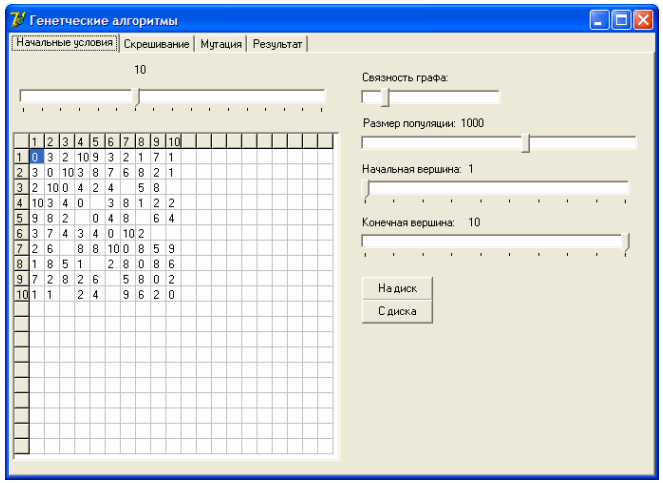


Рисунок 2. Общий вид окна системы.

Количество вершин графа может варьироваться от четырех до двадцати. Размер *популяции* может меняться от одной особи до ста тысяч. Поиск может производиться между двумя любыми вершинами. Длина хромосомы постоянна, количество генов в ней меньше количества вершин в графе на два, т.к. вершины, между которыми ищется кратчайший путь, не меняются во время поиска. Гены в хромосоме имеют десятичный формат для более простого понимания кода программы, и каждый из них представляет собой номер вершины. Длины дуг (или стоимость прохождения по дуге) могут принимать значения от 1 до 10, при этом простой на вершине не стоит ничего. Между двумя вершинами может вообще не быть пути, условимся тогда что в этом случае расстояние между вершинами

равно 1000001. Такое большое число гарантирует что хромосома, имеющая в своем составе такой путь, будет отсеяна во время *селекции*. В рассматриваемой системе мы может регулировать *связность графа*, т.е. процент «бесконечных вершин».

В данной задаче генетический алгоритм будет стараться минимизировать *целевую функцию*. Целевая функция в данной задаче – это сумма длин всех дуг пути, представляемого хромосомой.

Скрещивание.

В данной системе количество *точек скрещивания* можно варьировать. См. Рис 3.

Рисунок 3. Многоточечное скрещивание.

Процедура, которая осуществляет скрещивание двух хромосом, может иметь такой вид:

```
type Hromosom = record
      Gens: array of byte;
      f: real;
      end;

type MatrType = array of array of real;

var
  Find1, Find2: integer;
  Matrix: MatrType;

procedure CroosPair(RandFlag: boolean; pointsCount,
                      len: integer; h1,h2:hromosom;
                      var rez1,rez2:hromosom);
type rec=record
      num:integer;
      f:boolean;
     end;
var
 arr: array of rec;
 i, j, p, tlen: integer;
```

```
 flag: boolean;
 rand, points: integer;
begin
 tlen:=len-3;
 for i:=1 to (len-2) do
  begin
   arr[i].num:=i;
   arr[i].f:=false;
  end;
 if RandFlag then points:=random(pointscount)+1
  else points:=pointsCount;
 for i:=1 to points do
  begin
   p:=0;
   j:=1;
   rand:=(trunc(random(tlen))+1);
   while (j<=rand) do
    begin
     inc(p);
     while (arr[p].f) do inc(p);
     inc(j);
    end;
   dec(tlen);
   arr[p].f:=true;
  end;
 flag:=true;
 for i:=1 to (len-2) do
  begin
   if flag then
    begin
     rez1.Gens[i]:=h2.Gens[i];
     rez2.Gens[i]:=h1.Gens[i];
    end
   else
    begin
     rez1.Gens[i]:=h1.Gens[i];
     rez2.Gens[i]:=h2.Gens[i];
    end;
   if arr[i].f then
    if flag then flag:=false else flag:=true;
  end;
 rez1.f:=Matrix[Find1,rez1.Gens[1]];
 rez2.f:=Matrix[Find1,rez2.Gens[1]];
 for i:=1 to (len-3) do
  begin
   rez1.f:=rez1.f+Matrix[rez1.Gens[i],rez1.Gens[i+1]];
   rez2.f:=rez2.f+Matrix[rez2.Gens[i],rez2.Gens[i+1]];
  end;
   rez1.f:=rez1.f+Matrix[rez1.Gens[len-2],Find2];
   rez2.f:=rez2.f+Matrix[rez2.Gens[len-2],Find2];
end;
```

Данный листинг представляет собой процедуру для скрещивания двух хромосом в любом количестве точек. Первым параметром является флаг, который показывает, будет ли выбрано количество точек скрещивания случайно или оно будет задано вторым параметром. Третий параметр – длина хромосомы. Следующие два - хромосомы, которые требуется скрестить, а последние два – результат скрещивания. Глобальные данные: Find1 и Find2 вершины путь между которыми ищется; Matrix – симметричная матрица чисел, которая является матрицей весов дуг графа, поиск в котором ведется.

Arr – это вспомогательный массив для записи мест где будет происходить скрещивание. Все места скрещивания выбираются случайным образом и распределены равномерно. После выбора точек скрещивания, происходит обмен генами между хромосомами и в конце рассчитываются новые значения целевых функции.

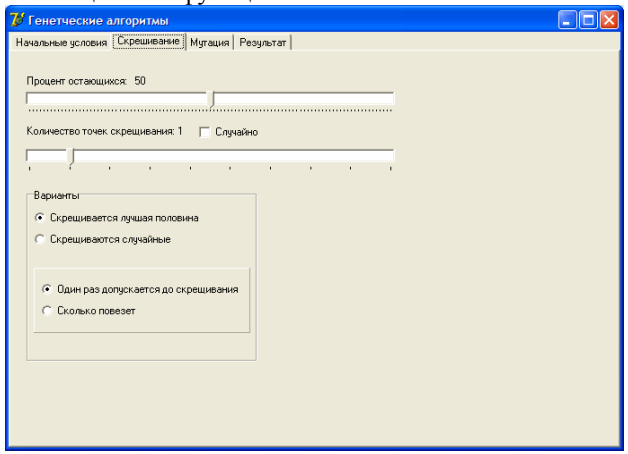


Рисунок 4. Настройка параметров скрещивания.

В данной системе доступны четыре разных способа отбора особей для скрещивания:

1. Скрещиваются две случайно выбранные особи поколения, и одна особь может скрещиваться несколько раз.
2. Скрещиваются две случайно выбранные особи поколения, и одна особь может скрещиваться только один раз.
3. Скрещиваются две особи из лучшей половины поколения, и одна особь может скрещиваться несколько раз.
4. Скрещиваются две особи из лучшей половины поколения, и одна особь может скрещиваться только два раза.

Рассмотрим каждый из этих вариантов и попытаемся выявить наиболее подходящий. Для того чтобы в эксперимент не вмешивались посторонние факторы, не будем производить мутации и сохранение наиболее приспособленных особей из предыдущих поколений зададим равным 1%. Чтобы эксперимент смог завершиться, зададим достаточно большой размер популяции – 1000 особей, количество генов – 10, и сделаем граф полно-связным. Также не будем учитывать случаи попадания в *локальный экстремум*, т.к. они пока нас не интересуют, и потом, при добавлении других операторов, появляться не будут. Количество точек скрещивания будем задавать два раза: сначала равное единице, а затем случайное. Это позволит определить влияет ли этот параметр на эффективность того или иного варианта. Оценивать эффективность того или иного метода будем по количеству новых поколений, которое придется создать для того, чтобы получить результат.

Первый вариант при одноточечном скрещивании фактически показал результат в 327 итерации, при случайном количестве точек скрещивания – 207, но при этом очень часто «проваливался» в локальный экстремум. Вариант 2, кроме того, замедлил примерно на 40%. процесс генерации нового поколения из-за контроля на отсутствие повторяемости, показал следующие результаты: 211 в первом и 190 во втором случаях. Третий вариант скрещивания дал результат в 62 итерации в первом случае и 51 во втором. Четвертый вариант оказался самым эффективным с результатом в 50 новых поколений с одноточечным скрещиванием, со случайным количеством точек скрещивания получилось в среднем 45 поколений.

Самым выгодным оказался вариант номер 4. Также эксперимент показал, что количество точек скрещивания качественно не влияет на выбор варианта в данных условиях.

Рассмотрим теперь эксперименты со случайным количеством точек скрещивания хромосом. Очень важной характеристикой процесса является то, что из-за случайности процесса решение часто не бывает найдено и попадает в локальный экстремум. Но если необходимая комбинация генов совпала, то решение находится быстрей из-за более сильной динамики процесса.

Теперь, при тех же начальных условиях эксперимента, выберем вариант 4, и будем варьировать количество точек скрещивания. В результате получим зависимость представленную на графике Рис 5.

Данное исследование показало довольно интересные результаты. Во-первых, скрещивание работает более эффективно если количество точек скрещивания четное, т.е. хромосома, как минимум, сохранят свой первый и последний ген. Так же из графика видно, что чем больше количество

точек скрещивания, то тем меньше количество итераций. Это происходит из-за того, что процесс приобретает все большую случайность. Про вариант с 7-ю точками стоит сказать отдельно. Для того чтобы добиться верного решения от него требуется упорство, т.к. процесс постоянно показывает более длинный путь, чем нужно. Но те редкие случаи, что были, дают возможность говорить о низком значении количества итераций. В данном случае происходящий процесс наименее похож на скрещивание, а носит чрезвычайно случайных характер, поэтому в случаях с меньшим количеством точек скрещивания «попадания» в локальный экстремум не наблюдалось.

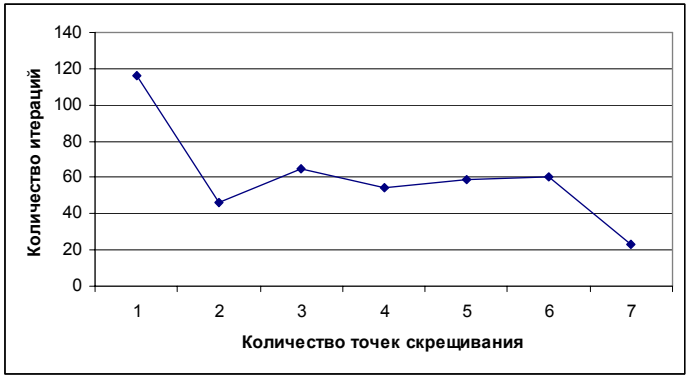


Рисунок 5. Зависимость количества итерации от количества точек скрещивания.

По итогам этого раздела будем считать, что наилучшим вариантом для скрещивания является вариант 4, а количество точек скрещивания нужно принимать равным двум.

Мутация

Мутация – второй по важности генетический оператор. С помощью мутации создаются такие цепочки генов, которые не входили в первое поколение, но были необходимы для получения правильного решения. В тоже время она создает бесполезные и «вредные» цепочки, что затрудняет поиск решения. Необходимо правильно пообобрать настройки мутации, чтобы она и помогала и мало мешала поиску решения. В рассматриваемой системе можно выбрать из четырех типов мутации: Инверсия вершины; Замена вершины на случайную; Замена вершины на соседнюю; Случайный способ из представленных выше.

Также задается вероятность, с которой к той или иной особи будет применена мутация. Внутри хромосомы мутирует случайная вершина.

Процедура мутации поколения может иметь такой вид:

```
type Population = array of Hromosom;

procedure Mutation(var ppl: population; amount, len,
  typem: integer; p: real);
var i, j, r: integer;
begin
  for i := 1 to amount do
    begin
      j := random(len - 2) + 1;
      if (random(1000001) / 10000) <= p then
        begin
          if typem = 4 then typem := random(3) + 1;
          if typem = 1 then
            ppl[i].Gens[j] := len + 1 - ppl[i].Gens[j]
          else if typem = 2 then
            ppl[i].Gens[j] := random(len) + 1
          else

            if j = 1 then
              ppl[i].Gens[j] := ppl[i].Gens[j + 1]
            else if j = (len - 2) then
              ppl[i].Gens[j] := ppl[i].Gens[j - 1]
            else
              begin
                r := random(2) + 1;
                if r = 1 then
                  ppl[i].Gens[j] := ppl[i].Gens[j + 1]
                else
                  ppl[i].Gens[j] := ppl[i].Gens[j - 1];
              end;
          ppl[i].f := Matrix[Find1, ppl[i].Gens[1]];
          for j := 1 to (len - 3) do
            ppl[i].f := ppl[i].f +
              Matrix[ppl[i].Gens[j], ppl[i].Gens[j +
              1]];
          ppl[i].f := ppl[i].f + Matrix[ppl[i].Gens[len
            - 2], Find2];
        end;
    end;
end;
```

Ppl – популяция, в которой необходимо провести мутацию, amount – количество особей в популяции, len – количество генов в хромосоме, typem – тип мутации согласно списку выше, p – вероятность мутации хромосомы в процентах.

Попробуем оценить какая самая оптимальная вероятность мутации. Для этого необходимо так подобрать параметры задачи, чтобы при поиске решения популяция вырождалась к неверному решению, имеющему более длинный путь, чем возможно было бы. Для этого необходимо задать достаточно небольшой размер популяции, чтобы необходимые сочетания генов встречались реже. Выборку самых лучших особей, проводить не будем, т.к. она значительно сокращает время поиска решения.

Выберем полносвязанный граф с десятью вершинами, размер популяции - 100, количество генов в особи – 10. Будем оценивать вероятность мутации по количеству создаваемых поколений для получения решения и по количеству неверных решений, если для каждого исследования требовалось получить десять верных решений. В результате получилась зависимость, представленная на рис. 6.

При данных начальных условиях получили результаты такие, что неизменно при увеличении вероятности мутации количество поколений, которые необходимо создать для получения решения неизменно растет. При вероятности мутации в 6% число неверных решений наиболее низкое. Значит, эта вероятность мутации оптимальна в данных условиях. При вероятности мутации в 30% и более количество неверных решений уменьшается, это происходит из-за того, что такая вероятность уже почти всегда обеспечивает необходимое сочетание генов, но время на получение решения уже сильно увеличивается.

Теперь рассмотрим эффективность того или иного способа мутации, из представленных в системе. При инверсии вершин, т.е. получении нового гена по номеру путем вычитания текущего номера из максимального, количество неверных решений при верных тридцати было равно 15. При мутации на случайную вершину – 10, такой результат лишь подчеркивает то, что мутация - процесс случайный. Чем более равномерное распределение, тем эффективнее мутация. При замене вершины в пути на следующую или предыдущую количество неверных решений получилось равным семи, значит, при выбрасывании вершины из пути, появляются правильные пути чаще. Но при уменьшении связности графа это преимущество нивелируется.

Элитизм

Элитизм в генетических алгоритмах – выбор или приведение решение к заведомо лучшему результату. Элитизм «помогает» алгоритму быстрее прийти к верному решению. В рассматриваемой системе он представлен механизмом сохранения лучших особей из предыдущего поколения. Некоторое количество (задается процентным соотношением относительно общего числа особей в популяции) особей, целевая функция которых ми-

нимальна, входят в следующее поколение без изменений. Этот метод позволяет выделять наилучшие варианты путей. Проведем исследование - при каком проценте сохранения лучших путей решение находится максимально быстро (Рис. 7.):

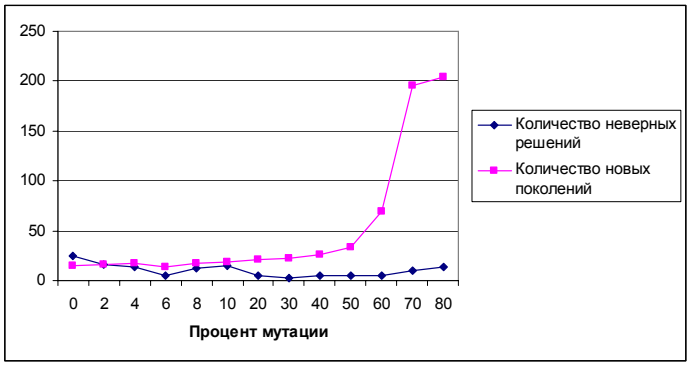


Рисунок 6. Исследование вероятности мутации.

Рисунок 7. Зависимость быстроты поиска решения от количества оставшихся вершин.

По графику видно, что элитизм начинает действовать с 30% сохранения особей и затем его влияние не изменяется практически до конца. Начиная с 80% количества новых особей уже не хватает, и решение ищется долго. Также при выборе количества сохраняемых вершин, необходимо учитывать, что чем больше вершин без изменения переходят из поколения в поколение, то тем больше неверных решений может возникнуть.

А.В.Сиренко

БАЗА ДАННЫХ ЛИНГВОКУЛЬТУРНОГО ТЕЗАУРУСА РУССКОГО ЯЗЫКА

Постановка задачи. Новые типы словарей

Естественный язык долгое время был для лингвистов неприкосновенным объектом изучения, и работы в области языковедения носили описательных характер. Таким образом, язык рассматривался как некий самостоятельно развивающийся объект изучения. Рассматривались различные факты языка, проводилась систематизация, выделение его структур, схем функционирования. С развитием вычислительной техники, постановкой новых задач и накоплением знаний о языках появилась необходимость в лингвистических объектах, имеющих новые свойства. Позволяющих с другой стороны взглянуть на естественный язык. Не только на сложившиеся внешние признаки его функционирования, но и на принципы развития языка. Такими объектами могут быть порождающие синтезаторы, грамматики нового типа и т.д. [Караулов, 1981, с.19]

В чем отличие этих новых лингвистических объектов от прежних? Очевидно, что словари строятся на эмпирической основе. То есть для их составления необходимо обработать некий массив языковых данных, например, в текстовом виде. Если же исходных данных в необходимом количестве нет, то составление словаря по сложившейся методике невозможно. Создание же порождающего механизма, метода лингвистического конструирования могло бы помочь иначе взглянуть на эту задачу. Кроме того, язык не является самостоятельным независимым объектом. Он функционирует и развивается в рамках окружающей его действительности и, с одной стороны, является ее отражением, а с другой, формирует наше представление о ней, так называемое «языковое сознание». Язык в данном случае – инструмент, посредством которого мы можем взглянуть на «сознание», на восприятие мира отдельным субъектом. Очевидно, что язык динамичен, но мы будем рассматривать относительно стабильную форму его существования в виде текстов.

Мы подошли к понятию Языковой Картины Мира. Она состоит из элементов, отражающих знания об окружающем мире (Единицы Знания

о Мире — ЕЗМ). Эти элементы могут представлять собой различные языковые структуры: определения, пословицы, прецедентные тексты. Совокупность таких элементов создает разнородную, подчас противоречивую картину мира. Языковое сознание имеет двойственную структуру, являя собой некий механизм соединения знаний о языке со знанием о мире. Языковое Сознание (далее ЯС) можно представить в виде структуры:

Языковое Сознание = Единица Знаний о Мире + Языковая Единица

В этой зависимости языковое сознание представляет собой когнайзер, в котором происходит постоянное преобразование Единиц Знаний о Мире в Языковые Единицы и наоборот. Задачей, которая ставится при выполнении описываемой работы, в конечном итоге, является моделирование процессов, протекающих в когнайзере, то есть перехода от ЕЗМ к ЯЕ и наоборот. Необходимо выяснить, как этот переход осуществить, возможно, ли это и равноценен ли переход в ту, или другую сторону? Работу когнайзера по переходу от слова (знака) к знанию будем называть активным режимом работы, а от знания к знаку – пассивным [Караулов, Филиппович 2005, с.5-6].

Активный и пассивный эксперимент

Ранее был проведен ассоциативный эксперимент. В нем респонденту называлось некое слово (стимул) и его задачей было назвать приходящее при этом в голову слово. Отвечать было необходимо не задумываясь [Черкасова, 2004]. В результате формировалась пара слов «стимул-реакция», которые можно рассматривать как отражение Языкового Сознания, где стимул выступает в роли Языковой Единицы, а реакция в роли Единицы Знаний о Мире. Причем, как показала практика, связь между стимулом и реакцией двунаправлена. Если стимул «слон» вызывает реакцию «хобот», то, как правило, имеет место и обратная связь. Стимул «хобот», скорее всего, имеет среди множества своих слов-реакций «слон». Для этого необходимо проведение достаточно полного опроса. Таким образом, здесь не важно, что мы выберем за Языковую Единицу, а что за Единицу Знания о Мире. Ассоциативный эксперимент отражает активный режим работы когнайзера.

Для рассмотрения пассивного режима работы когнайзера используются материалы кроссвордов. В кроссворде человеку предлагается смысл разгадываемого слова в той или иной форме (дефиниции, синонима и т.д.). Этот смысл, чаще всего в соединении с самим разгадываемым словом, представляет собой не что иное, как элементарное знание о мире, а именно о понятии, которое необходимо разгадать.

Например:

1) Оттенок на фоне какого-нибудь цвета – ОТЛИВ (Дескрипция).
2) Применяют, когда штурм не удался – ОСАДА (Антоним).
3) Очень мелкие цветные бусинки – БИСЕР (Дескрипция).

Таким образом, материалы кроссвордов представляют в своеобразной форме Языковую Картину Мира. И, если предположения о пассивном и активном режимах работы когнайзера верны, то результаты разгадывания кроссворда должны найти свое подкрепление и в активном режиме работы когнайзера – ассоциативном эксперименте.

Знания об окружающем мире можно представить в виде так называемых Фигур Знания [Караулов, 2004]. Это элементарные когнитивные единицы, которыми мы оперируем в повседневной жизни. Они содержат как интенсиональные характеристики, а именно Знак, Способ, Смысл, описанные выше, так и экстенсиональные, отражающие положение Фигуры знания в общем пространстве знаний. Такими параметрами являются когнитивная область и функция. Схематичное представление фигуры знания можно увидеть на рис.1

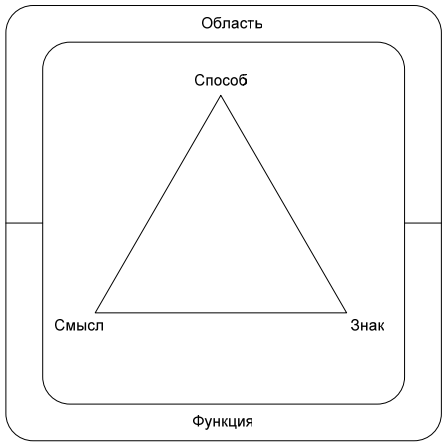


Рис.1 Фигура Знания

Все знания мы можем разбить на категории, отражающие определенные стороны окружающего мира, на области. Знак может быть связан с Областью многообразными ассоциативными связями. К чему относится, с чем связано, частью чего является – вот далеко не полный список возможных вариантов. Строго говоря, составление списка возможных областей и разбиение фигур знания по ним — процесс субъек-

тивный, и каждый исследователь сделает это по-своему. В процессе наполнения базы данных эксперимента постепенно формируется множество возможных областей. При обработке достаточно большого количества понятий это множество перестает расти. Очевидно, одно и то же понятие может принадлежать нескольким областям, что свидетельствует о том, что области могут частично перекрываться, и разбиение на них весьма условно.

В активном режиме когнайзер переходит от Знака к Смыслу. При этом определяется Способ, которым Знак раскрывается через Смысл. В пассивном же режиме изначально имеется Смысл и, таким образом заданный, Способ. Имея эти два параметра, когнайзер переходит к Знаку. В обоих режимах экстенсионал фигуры знания определяется после интенсионала. Казалось бы, связи между экстенсиональными и интенсиональными параметрами быть не должно, однако, некоторые зависимости можно проследить. Например, зависимости между Областью и Способом. В некоторых областях очевидно предпочтение определенным Способам выражения смысла. Например, Способ «Элемент множества» очень популярен в Географии. Также среди фигур знания, принадлежащих области «Авиация», многие имеют Функцию «Ретушь», так как знания в этой области не всегда важны в обыденной жизни. В области «Быт» наблюдаем обратную тенденцию.

Попробуем на примере проследить связь между активным и пассивным режимом в связи знака с областью. Важно заметить, что в ассоциативно-вербальной сети присутствует столь великое множество пар стимулов-реакций, что если мы будем искать связь между словами через сколь угодное множество промежуточных переходов, то на практике можем столкнуться с перемещением в огромном числе ветвлений и цепочек, где практически между любыми словами находится путь. Наша же задача состоит в нахождении минимальных путей. Рассмотрим следующую единицу знания пассивного режима когнайзера:

1. Последняя буква кириллицы. 2. Фрейм. 3. ИЖИЦА. 4. Язык. 5. Рцт.

Рассмотрим обратный ассоциативный словарь, организованный от реакции к стимулу, включающий в себя все содержимое ассоциативного словаря. Ищем пары, в которых ИЖИЦА является реакцией. Таких пар не нашлось, значит, из смысловой формулы подбираем возможный аналог, это знак БУКВА. В обратном ассоциативном словаре БУКВА является реакцией на множество стимулов, среди которых есть стимул «язык», что подтверждает связь знака с областью в активном и пассивном режиме.

Исходные данные. Первичная обработка

Исходные данные для эксперимента представлены, как уже описывалось выше, данными ассоциативного эксперимента, с одной стороны, и материалами кроссвордов, с другой.

Ассоциативный эксперимент существует в виде базы данных в формате Paradox. Материалы кроссвордов же, было необходимо привести к форме, удобной для последующего использования.

Для этого сначала было необходимо выбрать технологическую базу выполнения эксперимента, то есть выбрать среду разработки программного обеспечения и систему управления базой данных, поскольку работа с информацией предполагала выполнение различных запросов. Выбор пал на использование Borland Delphi , входящий в состав пакета Borland Developer Studio 2006, так как он удобен для быстрой разработки приложений и имеет широкий набор средств для обращения к различным СУБД (Системам Управления Базами Данных). Базу данных предполагалось строить на основе MS Access. Причин для этого несколько:

1) MS Access установлен на подавляющем большинстве персональных компьютеров, что позволит работать с базой данных практически везде, то есть любому исследователю.

2) База данных Access представляет собой отдельный файл формата mdb и при переносе программного обеспечения эксперимента этот файл также может быть перемещен, без необходимости проведения каких-либо наладочных работ. Это дает мобильность.

3) Эта СУБД входит в пакет MS Office, поэтому имеет возможность импорта и экспорта данных в различном виде. Например, импорта данных в виде файла MS Excel, или экспорта базы данных в СУБД MS SQL Server, что важно как предположительный вариант модернизации базы данных в случае необходимости улучшения ее характеристик.

4) Access имеет собственные средства для обработки данных, выполнения запросов, генерации отчетов и выборок.

После выбора СУБД исходная таблица, представленная в текстовом документе MS Word, была сперва переведена в MS Excel, а затем перенесена в базу данных Access.

Таблица имеет следующий вид:

Код	Знак	Смысл	Область	Способ	Функция

Будем именовать эту таблицу далее таблицей фигур знания.

Необходимо внести пояснения по ее составу.

КОД имеет числовой формат и служит для идентификации записей.

ЗНАК содержит в текстовом виде слово, которое в кроссворде требуется отгадать.

СМЫСЛ представляет собой текст, на который ориентируется отгадывающий человек, то есть разъяснение понятия, представленное в той или иной форме.

ОБЛАСТЬ является перечислением тех областей знания, к которым относится данное понятие. Каждое понятие принадлежит как минимум одной области. Области в поле перечислены через запятую, пробелы либо другие небуквенные символы.

СПОСОБ представляет собой структуру смысла, через который загадано слово. Это может быть дефиниция, прецедентный текст, метафора или множество других способов.

ФУНКЦИЯ определяет так называемую «ценность» понятия. С точки зрения ценности различают знание-рецепт и знание-ретушь [Караулов, Филиппович, 2005, с.14]. Знание-рецепт несет в себе важную, существенную для исследователя информацию, необходимую ему в повседневной жизни. Знание-ретушь содержит дополнительные, менее существенные данные, которые в некоторой степени можно не рассматривать при анализе текста. Исследователь, определяя, к какому типу знание относится, руководствуется собственными представлениями о необходимости этого знания для него самого, для его понимания окружающего мира.

Рассмотренная выше таблица имеет ряд недостатков. Например, поле СПОСОБ может принимать одно из некоторого набора возможных значений. На этапе разработки количество вариантов способов не известны. Запись в поле СПОСОБ наименования способа ведет не только к избыточности объемов хранимых в памяти данных, но и является потенциальным источником ошибок. Поэтому возможные области выделяются в отдельную таблицу, а в первоначальной таблице записывается лишь код записи таблицы способов.

Аналогично можно поступить с полем ФУНКЦИЯ.

Особенностью областей с позиции строения базы данных является то, что, во-первых, области образуют некоторое конечное множество, во-вторых, любое понятие может принадлежать одной либо нескольким областям. Поэтому использование той же схемы выделения отдельных таблиц, как в случае с полями СПОСОБ и ФУНКЦИЯ, невозможно. Была смоделирована связь многие-ко-многим с помощью промежуточной таблицы, в которой каждая запись представляет собой сочетание кода

записи таблицы данных и кода записи таблицы областей. Таким образом, одной записи исходной таблицы может быть поставлено в соответствие несколько областей.

Структура таблиц. Их заполнение

Итоговая схема БД представлена на рисунке 1.

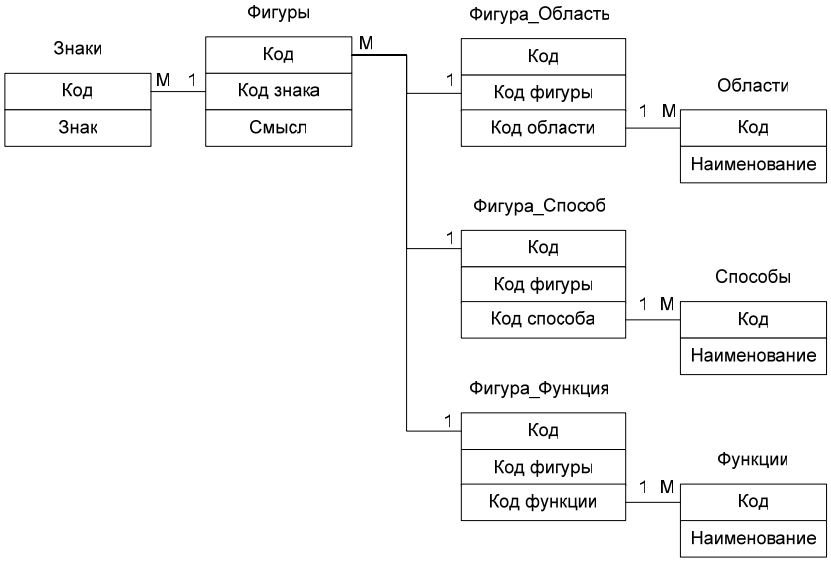


Рис.1

Заполнение таблиц Знаки, Фигуры, Фигура_Область, Фигура_Способ, Фигура_Функция, Области, Способы по исходной таблице не может выполняться полностью в автоматическом режиме. Допустим, в поле Область исходной таблицы мы встречаем обозначение, которое не присутствует среди имеющихся в таблице областей наименований. Либо нам необходимо добавить новую область и создать связь, либо это одна из имеющихся областей, просто иначе обозначенная, в наименовании которой допущена ошибка. Поиск же в БД имеющихся областей и предложение их пользователю можно автоматизировать.

Для автоматизации разбора исходной таблицы была разработана форма, изображенная на рисунке 2.

В таблице представлены еще не обработанные записи. Ниже в текстовых полях представлен состав активной записи, которая будет обработана при нажатии кнопки «Разобрать».

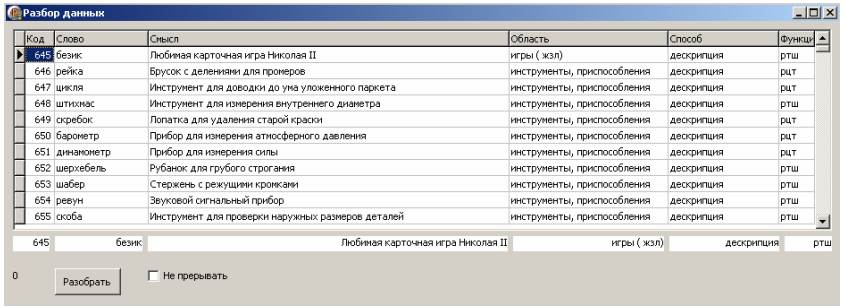


Рис.2

Разбор очередной записи исходной таблицы и заполнение результирующих таблиц представляет собой последовательность из следующих этапов:

1) Если среди Знаков уже есть содержимое поля «Знак» исходной таблицы, Фигура знания связывается с имеющимся знаком. Иначе создается новый Знак.

2) Затем среди Областей осуществляется поиск элементов, наименования которых встречаются в поле «Область» исходной таблицы. В случае если в записи исходной таблицы всего одна область, и наблюдается полное совпадение с какой-либо областью таблицы областей, связь устанавливается автоматически, без участия пользователя. Иначе предлагается вручную выбрать необходимые области, в том числе с возможностью создать новые. Пример такого диалога можно увидеть на рис. 3.

3) Производится разбор поля Способ, аналогично полю Область.

Стоит заметить, что если осуществить разбор поля автоматически не удается, пользователю предлагается выбрать Способ из имеющихся в базе данных, либо ввести новый.

4) Заполнение поля Функция.

Обработав таким образом все записи исходной таблицы, переходим к схеме данных, представленной на рис.1.

Главная форма программного обеспечения позволяет просматривать эти таблицы в удобной форме, а также переходить к редактированию. Ее внешний вид показан на рис.3.

В таблице «Понятие» перечислены наименования всех понятий, представленных в БД, согласно одноименной таблице. Таблица «Смысл» перечисляет все записи таблицы «Фигуры_Знания», которые соответствуют текущей записи в таблице понятий.

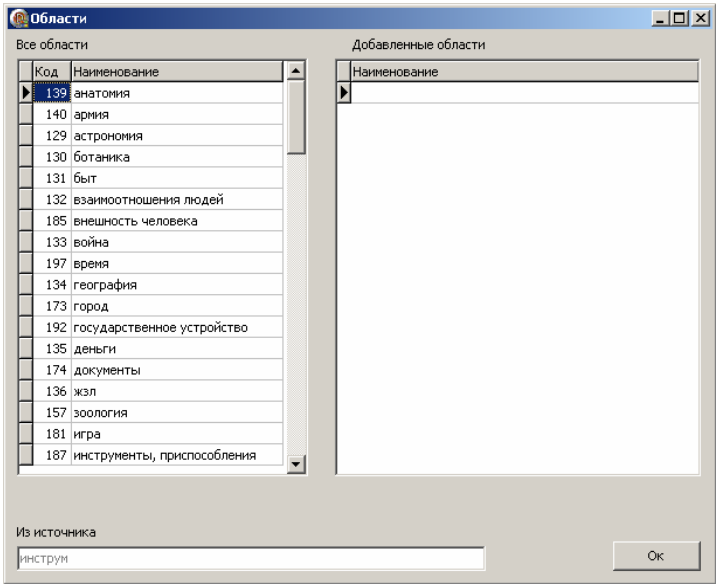

Области
Все области
Код
Наименование
139 анатомия
140 армия
129 астрономия
130 ботаника
131 быт
132 взаимоотношения людей
185 внешность человека
133 война
197 время
134 география
173 город
192 государственное устройство
135 деньги
174 документы
136 жзл
157 зоология
181 игра
187 инструменты, приспособления
Добавленные области
Наименование
Из источника
инструм
Ок

Рис.3

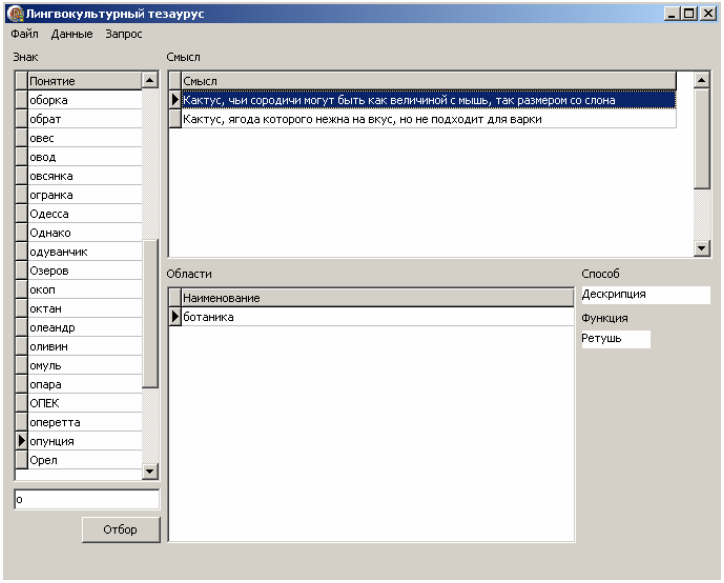

Лингвокультурный тезаурус
Файл
Данные
Запрос
Знак
Понятие
оборка
обрат
овес
овод
овсянка
огранка
Одесса
Однако
одуванчик
Озеров
окоп
октан
олеандр
оливин
омуль
опара
ОПЕК
оперетта
опунция
Орел
о
Отбор
Смысл
Смысл
Кактус, чьи сородичи могут быть как величиной с мышь, так размером со слона
Кактус, ягода которого нежна на вкус, но не подходит для варки
Области
Наименование
ботаника
Способ
Дескрипция
Функция
Ретушь

Рис.4

Таблица «Области» и текстовое поле «Способ» отображают Области и Способ, связанные с текущей фигурой знания.

Поиск по базе данных

Для поиска информации, как одиночных, так и групп записей в соответствии с различными требованиями была разработана форма, представленная на рисунке 5.

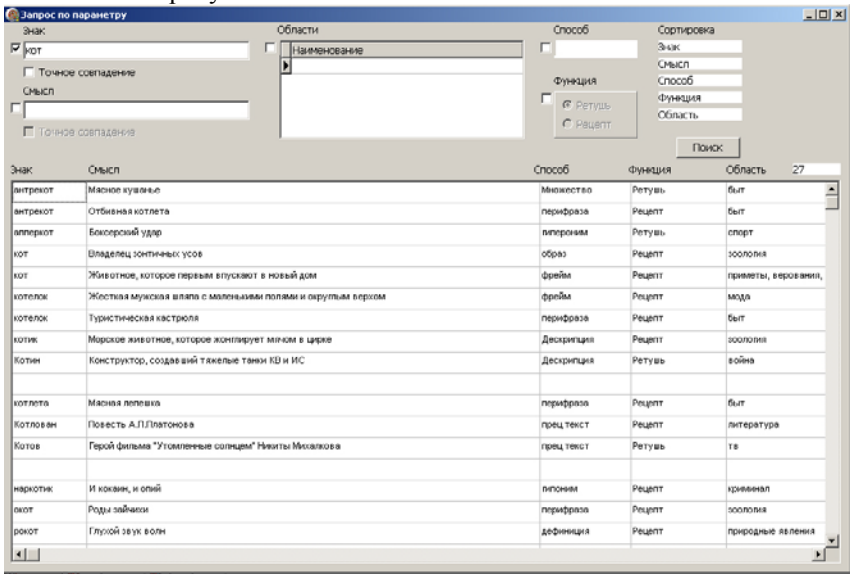


Рис.5

В верхней части формы пользователь задает параметры поиска. Для текстовых полей, таких как Знак и Смысл можно установить положение переключателя «Точное совпадение». В случае если переключатель установлен в положение «Нет», результаты поиска будут содержать фигуры знания, Знак или Смысл которых содержат в качестве подстрок заданные значения. Например, на рисунке 5 в качестве Знака указано значение «кот». Среди результатов можно видеть фигуры знания со знаком «Котлован», «рокот» и множество других, в количестве 27.

Рядом с каждым параметром стоит переключатель, который позволяет исключить его из рассмотрения. В результате установки параметров поиска и нажатия кнопки «Поиск» формируется SQL-запрос, включающий также сортировку. Таким образом, результаты запроса не нуждаются в дальнейшей обработке.

Заключение

Полученная схема базы данных представляет собой третью нормальную форму в терминах построения баз данных, что обеспечивает не только минимизацию ресурсов на хранение информации за счет ухода от избыточности, но и уменьшение риска ошибок. Она облегчает построение запросов, что позволит выполнять необходимые операции при построении и дальнейшем функционировании лингвокультурного тезауруса. Применение полноценной среды разработки позволяет использовать как работу с данными посредством SQL-запросов, так и самостоятельно описывать обработку записей БД, что может быть полезно в сложных операциях. Программное обеспечение не зависит от одной конкретной базы данных. Можно использовать несколько источников данных, в том числе неоднородных по своему типу. Разработанная база данных позволяет модификацию себя с сохранением работоспособности и расширение за счет добавления новых таблиц либо новых полей в старых таблицах.

Литература

Караулов, 1981 *Ю.Н.Караулов.* Лингвистическое конструирование и тезаурус русского языка. –М.: Издательство «Наука», 1981.

Караулов, 2005 *Ю.Н.Караулов, Ю.Н.Филиппович.* Лингвокультурологический тезаурус русского языка. –М.: 2005.

Караулов, 2004 *Ю.Н.Караулов.* Концептография языковой картины мира. Статья 1. Первый этап «восхождения» к образу мира: от элементарных фигур знания к предметно-референтным областям культуры// Проблемы прикладной лингвистики. Выпуск 2. Сборник статей./ Отв. ред. Н.В.Васильева. –М.: «Азбуковник», 2004. – 400с.

Черкасова, 2004 *Г.А.Черкасова.* Формальная модель ассоциативного исследования.// Проблемы прикладной лингвистики. Выпуск 2. Сборник статей./ Отв. ред. Н.В.Васильева. –М.: «Азбуковник», 2004. – 400с.

Анна Филиппович

АНАЛИЗ ТИПОГРАФСКИХ ТЕКСТОВЫХ ШРИФТОВ ВТОРОЙ ПОЛОВИНЫ XVIII ВЕКА

Общее описание шрифтов

Во второй половине XVIII века графика гражданского типографского шрифта определялась типографиями Академии наук (АН), Московского университета (МУ) и некоторых других, например Императорской, Московского кадетского корпуса, Вейтбрехта, Шнора, «Селивановского и товарища», Лопухина, Анненкова, Мейра, Гиппиус, Хр. Клаудия, Зеленкова и др. Но вплоть до конца века все типографии находились под влиянием типографии АН: «они заимствовали у нее не только шрифты и типографское оборудование, но и квалифицированные типографские кадры» [цит. по Шицгал, 1985, с. 52].

Ассортимент шрифтов этого периода времени представлен в следующих источниках: [Пробная книга…, 1748], [Ведомость…, 1796], [Показание, 1788]; типографии Московского университета: [Образцы…, 1796], [Образцы…, 1808], [Образцы…, 1815], [Образцы…, 1826]; некоторых других типографий: [Образцы…, 1805-А], [Образцы…, 1805-Б], [Образцы…, 1807], [Collection …, 1810] и др.

В начале XVIII века шрифты (азбуки) назывались по их назначению: например куплементальная (размер шрифта «Геометрии» и «Комплементов»), коронационная (им было набрано издание «Коронация Елизаветы…» 1744 года) [Шицгал, 1959]. Названия шрифтов середины XVIII века заимствованы из Германии: «*Доппель Цицеро*», «*Миттель Антиква*», «*Гробе Цицеро*» и др. Данные названия даны в соответствие с размером шрифта: например, Петит – 8 пунктов (п.), Корпус – 10 п., Цицеро – 12 п., Миттель – 14 п., Канон – 42 п. [Волкова, 2002, с.110-112] и т.д. Однако размеры не соответствуют применявшимся позднее, например «*Гробе Цицеро*» по размеру ближе к современному корпусу (10 п.) [Шицгал, 1985].

Шрифты 50–80-х годов XVIII века

Шрифты типографии Академии наук характерные для середины XVIII века представлены в [Пробная книга..., 1748]. Графика этих

шрифтов, во-первых, отражает характерные особенности Петровского времени (первой половины XVIII в.), во-вторых, новые тенденции. Обратимся ко второй группе шрифтов, графика которых характерна для второй половины XVIII века.

Текстовые шрифты, отражающие графику середины XVIII века, это прямые: *Парагон Антиква, Гробе Цицеро, Корпус Антиква* и *Нонпарель Антиква*; курсивные начертания: *Парагон Антиква* и *Цицеро Антиква.* Кроме этих шрифтов использовался шрифт *Миттель Антиква.* Он появился в 1733 году и «был широко распространен в изданиях гражданской печати в Петербурге и Москве вплоть до восьмидесятых годов XVIII века» [цит. Шицгал, 1974].

По особенностям шрифтового оформления издания, выпущенные типографией Академии наук с 40-х до 80-х годов XVIII века, можно разделить на две группы.

В первую группу необходимо отнести литературу, посвященную торжествам в честь императрицы Елизаветы (рис. 1), примером может служить [Коронация Елисаветы, 1744] и некоторые официальные издания, например [Привелегия..., 1765].

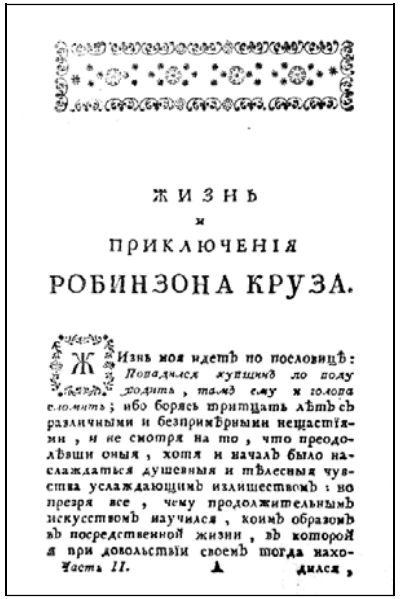

Когда неисповѣдными судьбами и неизреченнымъ промысломъ всѣхъ благъ подателя Творца Бога 25. Ноября 1741. года ЕЯ ИМПЕРАТОРСКОЕ ВЕЛИЧЕСТВО, всемилостивѣйшая Государыня, ко общей всѣхъ вѣрныхъ подданыхъ радости, въ царствующемъ градѣ Санктпетербургѣ благополучно на Всероссійскій Императорскій родительскій престолъ вступила, и потомъ высочайшее Свое намѣреніе объявить соизволила, дабы по обычаю Императорскихъ своихъ предковъ и прочихъ Христіанскихъ Потентатовъ коронованіе, и святое муропомазаніе, восприять въ царствующемъ

А перво-

Рис. 1. Начальная страница из книги [Коронация Елисаветы, 1744].

Эти издания, как правило, имели большой или средний форматы. Их титулы были гравированными или включали почти полный набор образцов шрифтов типографии. Текст же набирался, как правило, крупным шрифтом — *Парагон Антиква*, прямым и курсивным начертанием (рис. 3, а). Стиль оформления изданий торжественный, пышно-декоративный близкий к стилю барокко.

Ко второй группе изданий, особенно распространенных в 60–70-х годах, следует отнести большую часть «изящной» литературы (рис. 2), например [Дефо, 1762-1764].



ЖИЗНЬ
и
ПРИКЛЮЧЕНІЯ
РОБИНЗОНА КРУЗА.

Жизнь моя идетъ по пословицѣ: *Повадился кувшинъ по воду ходить, тамъ ему и голова сломить*; ибо борясь тридцать лѣтъ съ различными и безпримѣрными нещастіями, и не смотря на то, что преодолѣвши оныя, хотя и началъ было наслаждаться душевныя и тѣлесныя чувства услаждающимъ излишествомъ: но презря все, чему продолжительнымъ искусствомъ научился, коимъ образомъ въ посредственной жизни, въ которой я при довольствіи своемъ тогда нахо-

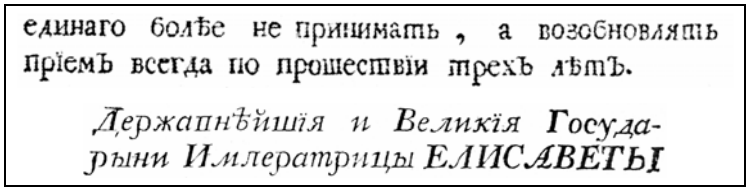
Часть II. А дился,

Рис. 1. Начальная страница из книги [Дефо, 1762-1764, ч. II].

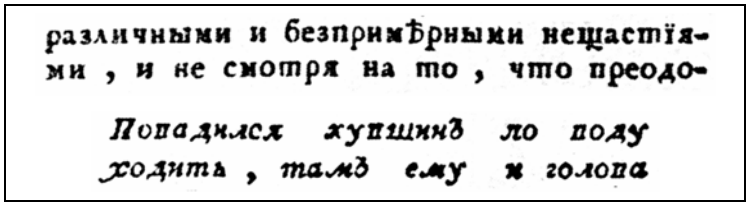
Шрифтовое оформление здесь сочеталось с легким узором наборного орнамента в основном цветочных мотивов. Книги эти, в противоположность изданиям первой группы, выходили в малых форматах. Используется шрифт *Гробе Цицеро* (рис. 3, б). Этот шрифт особой популярностью пользовался в царствование Екатерины II и даже употреблялся в научных изданиях, например, [Новиков, 1773]. Шрифтом *Гробе Цицеро* часто набирались журналы, в том числе первый русский журнал «Ежемесячные сочинения к пользе и увеселению служащих» (1755) и

петербургские журналы Новикова («Трутень», 1769-1770; «Живописец», 1772-1773; «Кошелек», 1774).

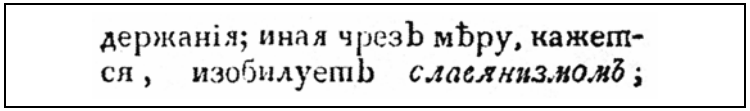
а)



б)



в)



г)

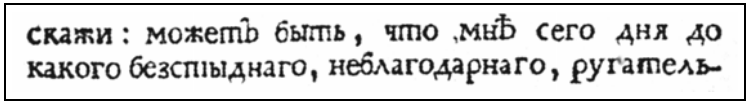
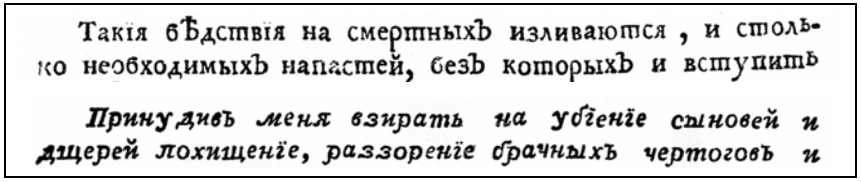


Рис. 3. Образцы шрифтовых гарнитур 50–80-х годов:
а) *Парагон Антиква* типографии АН (из [Привилегия..., 1765]); б) *Гробе Цицеро* типографии АН; в) *Миттель* типографии МУ, заимствованный у типографии АН (из [Бакаревич, 1880]); г) *Миттель Антиква* типографии АН (из [Житие..., 1740]);

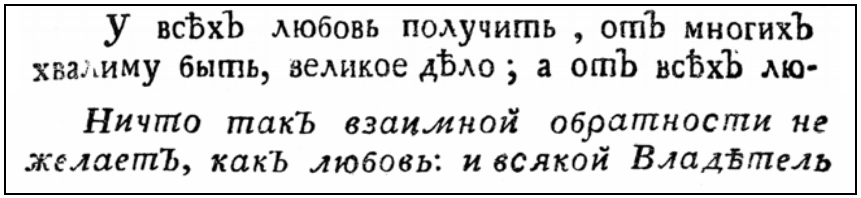
В типографии Московского университета, основанной в 1755 году использовались русские шрифты заимствованные из Академии наук, а также изготовленные в собственной словолитне. К 1761 году в типографии имелись петербургские текстовые шрифты (т. е. типографии Академии наук): *Парагон Антиква* с курсивом (№ 10), *Гробе Цицеро Антиква, Цицеро Антиква* с курсивом (№ 13) и *Боргес Антиква* (№ 15). Из шрифтов, отлитых в типографии Московского университета, были *Миттель Антиква* с курсивом (№ 12), *Цицеро Антиква* (№ 13), *Корпус Антиква* с курсивом (№ 14). Анализ перечисленных шрифтов показыва-

ет, что московские литеры были сделаны по образцу академических, но несколько уступали им по качеству рисунка (рис. 3, в, г). Наиболее популярными шрифтами типографии Московского университета в 50–60-х г. являются *Парагон Антиква* и *Миттель Антиква*.

а)



б)



в)

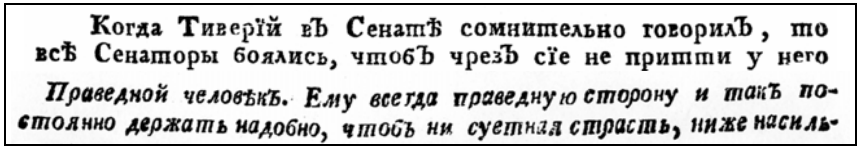


Рис. 4. Образцы старых шрифтов 1788 года типографии АН:
а) *Тонкое Цицеро* и *Цицеро Курсив*; в) *Миттель*; г) *Корпус*.

Шрифты конца 80-х – 90-х годов XVIII века

Новые шрифты 80-90 гг. типографии АН представлены в [Показание, 1788], охарактеризуем их (рис. 5). Текстовые образцы шрифтов, так называемые ординарные, представлены четырьмя размерами: *Терция Прямая* с курсивом (примерно равен 14 пунктам), *Ординарный Миттель на Терцию* с курсивом (равен 12 пунктам), *Гробе Цицеро на Миттель* (равен 11 пунктам), *Корпус на Цицеро* с курсивом (равен 10 пунктам). Новые образцы отличаются от старых размером очка, который зависел от емкости шрифта, и рисунком — он стал строже, тоньше и светлее за счет уменьшения толщины основных штрихов. Существенно изменились в новом рисунке шрифта курсивы. В них отсутствуют, в частности, буквы с разным наклоном. Прописное начертание курсива *Гробе Цицеро* построено на основе гравированных образцов сороковых годов

XVIII века [Шицгал, 1985]. Таким образом, новые шрифты были более строгими по начертанию и являются переходной формой от существовавшего ранее пышно-декоративного стиля к классическому, используемому в начале XIX века.

а)

О смертный! покорись противностямъ
своимъ, сноси безъ слабости, что духъ ни
Знать силу и время вещей. Всякая на-
туральная вещь до толь ростетъ, лока въ

б)

Съ командиромъ симъ въ сраженье
И въ кровавой идемъ бой,
Лишь бы только намъ случилось
Тѣхъ злодѣевъ поимать,

в)

Аптекарь дѣлаетъ лѣкарства только по пред-
писанїю врача; и по тому не должно давашься
Взоръ есть нѣкоторая тайная сила
высочества природнаго, а не отъ прикрасы или

г)

Хочу я у тебя спросить мой другъ совѣта,
И жду на свой вопросъ рѣшительнаго ошвѣта.
Но естлижъ я, женясь, ленять на то самъ стану,
Слокойстеа я лишусь; и въ цвѣтѣ лѣтъ увяну.

Рис. 5. Образцы новых шрифтов 1788 года типографии АН:
а) Терция Прямая; б) Ординарный Миттель на Терцию;
в) Гробе Цицеро на Миттель (Дашковская гарнитура);
г) Корпус на Цицеро.

Издания типографии Академии наук конца XVIII века набирались строгими по рисунку шрифтами (*Терция*, *Одинарный Миттель на Тер-*

цию, Гробе Цицеро на Миттель прямого и курсивного начертания), часто без виньеток и особых украшений [Шицгал, 1959]. К этим изданиям следует отнести [Бюффон, 1789], [Марк Витрувий, 1790], [САР, 1789-1794].

С 1780 по 1789 год, когда арендатором типографии МУ был Н.И. Новиков, его издания часто набирались шрифтами типографии АН (примером служит [Букварь, 1780]), в том числе новых образцов 1788. Однако, начиная с восьмидесятых годов, особенно в конце XVIII века, отдельные варианты шрифтов университетской типографии приобрели своеобразный характер, резко отличающий их от шрифтов типографии АН. Это и привело к тому, что в России с конца XVIII века установилось два типа шрифтов: петербургский и московский.

В [Образцы..., 1796] текстовые шрифты МУ по рисунку можно разделить на три группы. К первой группе относятся шрифты, которые сделаны по академическому образцу: *Гробе Цицеро* и *Терция Прямая* (№ 16). Контрастные по типу, они более строги по рисунку, чем академические. Ко второй группе относятся шрифты, которые представляют собой модификацию академических образцов в сторону приближения их рисунка к шрифтам старого стиля: *Терция Прямая* (№ 14), *Прямой Миттель на Терцию* (№ 18), *Длинный Большой Миттель* (№ 20), *Длинное Цицеро* (№ 29) и другие. К третьей группе относится новый тип текстового шрифта, так называемого «Круглого» начертания, ставшего прототипом московского рисунка начала XIX века (рис. 6.)

а) 

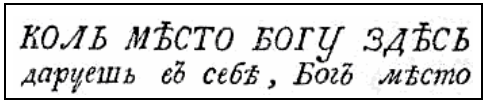
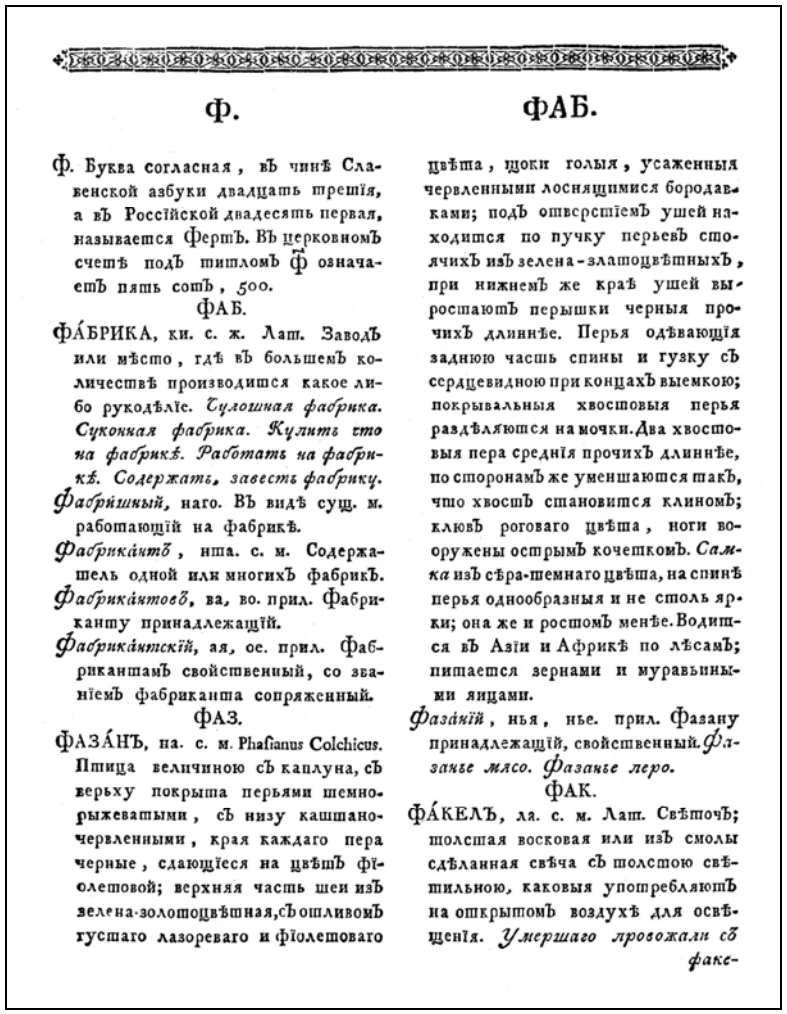
б) 

Рис. 6. Образцы шрифтов «круглого» начертания 1796 года типографии МУ: а) Толстый круглый Миттель №22; б) Курсив Миттель № 23.

Дашковская гарнитура

Характерным для второй половины XVIII века изданием типографии Академии наук является Словарь Академии Российской 1789-1794 гг. (САР) – см. рис. 7.



Ф. ФАБ.

Ф. Буква согласная, въ чинѣ Славенской азбуки двадцать третія, а въ Россійской двадесять первая, называется Ферть. Въ церковномъ счетѣ подъ титломъ ф̄ означаетъ пять сотъ, 500.

ФАБ.

ФА́БРИКА, ки. с. ж. Лат. Заводъ или мѣсто, гдѣ въ большемъ количествѣ производится какое либо рукодѣліе. *Суконная фабрика. Суконная фабрика. Купить что на фабрикѣ. Работать на фабрикѣ. Содержать, завесть фабрику.*

Фабри́шный, наго. Въ видѣ сущ. м. работающій на фабрикѣ.

Фабрика́нтъ, нта. с. м. Содержатель одной или многихъ фабрикъ.

Фабрика́нтовъ, ва, во. прил. Фабриканту принадлежащій.

Фабрика́нтскій, ая, ое. прил. Фабриканта свойственный, со званіемъ фабриканта сопряженный.

ФАЗ.

ФАЗА́НЪ, на. с. м. Phasianus Colchicus. Птица величиною съ каплуна, съ верьху покрыта перьями темно-рыжеватыми, съ низу кашанно-червленными, края каждаго пера черные, сдающіеся на цвѣтъ фіолетовой; верхняя часть шеи изъ зелена-золотоцвѣтная, съ шлевомъ густаго лазореваго и фіолетоваго цвѣта, щоки голыя, усаженныя червленными лоснящимися бородавками; подъ отверстіемъ ушей находится по пучку перьевъ стоячихъ изъ зелена-златоцвѣтныхъ, при нижнемъ же краѣ ушей выростаютъ перышки черныя прочихъ длиннѣе. Перья одѣвающія заднюю часть спины и гузку съ сердцевидною при концахъ выемкою; покрывальныя хвостовыя перья раздѣляются на мочки. Два хвостовыя пера среднія прочихъ длиннѣе, по сторонамъ же уменшаются такъ, что хвостъ становится клиномъ; клювъ роговаго цвѣта, ноги вооружены острымъ кочетковъ. *Самка* изъ сѣра-темнаго цвѣта, на спинѣ перья однообразныя и не столь ярки; она же и ростомъ менѣе. Водится въ Азіи и Африкѣ по лѣсамъ; питается зернами и муравьиными яицами.

Фаза́ній, нья, нье. прил. Фазану принадлежащій, свойственный. *Фазанье мясо. Фазанье перо.*

ФАК.

ФА́КЕЛЪ, ла. с. м. Лат. Свѣточъ; толстая восковая или изъ смолы сдѣланная свѣча съ толстою свѣтильною, каковыя употребляютъ на открытомъ воздухѣ для освѣщенія. *Умершаго провожали съ факе-*

Рис. 7. Страница текста САР, т. 6.

В его оформлении используются несколько шрифтовых гарнитур. Согласно документу [Показание, 1788] в качестве основного шрифта для набора текста используется шрифт «*Гродъ Цицеро на Мит-*

тель» (рис. 5, а и рис. 9, а) прямого и курсивного начертания. Назовем данную гарнитуру *Дашковской*. В некоторых научных статьях и в переиздании САР эта гарнитура называется *Елизаветинской*. Однако такое название не является корректным, т.к. шрифтовая гарнитура с таким названием существует. Данная гарнитура была разработана в начале XXI века дизайнером Л.Кузнецовой на основе рисунков дореволюционных шрифтов словолитного заведения Осипа Лемана в Санкт-Петербурге [ParaType]. Особенностью Елизаветинской гарнитуры является высокая контрастность букв и усложненное начертание ряда букв, сходное с русским шрифтом середины XVIII в. (З, Э, б, з, ц, щ, э) [Добкин, 1985] – см. рис. 8.

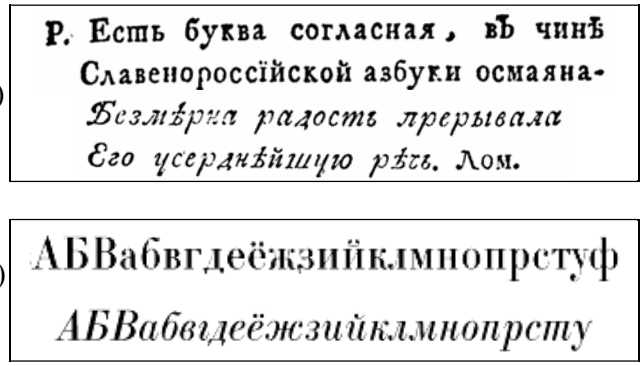


Рис. 8. Сравнение Дашковской и Елизаветинской гарнитур:
а) Дашковская; б) Елизаветинская.

Кроме основной Дашковской гарнитуры в САР используются другие шрифтовые гарнитуры. Для набора колонтитулов, скорее всего, используется гарнитура *Герцiя Прямая*, она же видимо использовалась для набора предисловия (рис. 9, б). В тексте встречаются латинские названия растений и животных, данные специфической гарнитурой (рис. 9, в). Показание к Словарю набрано меньшим кеглем. Возможно это Дашковская гарнитура меньшего кегля, о чем свидетельствует встречаемый курсив. В прямом начертании этого кегля есть сходства с гарнитурой *Корпус на Цицеро*. Толкования слов Церковнославянской лексики содержат написание этих слов в старой форме кириллицы (рис. 9, г).

В титульных надписях и в разделах «Предисловие», «Изъяснение», «Господа, члены Российскую Академии составляющие» и др. используется множество крупных гарнитур разных размеров. Из тех что встре-

чаются в [Показание, 1788] это гарнитуры «*Клейне каноно*», «*Доллель Цицеро*», «*Гроо Секунда*», «*Герція Курсивъ*».

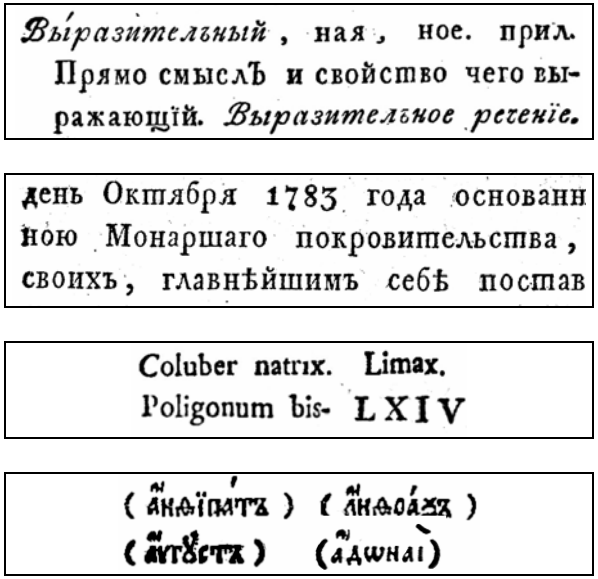


Рис. 9. Образцы шрифтовых гарнитур САР:
а) основной текст САР [САР, т. 5]; б) Предисловие [САР, т. 1];
в) латинские названия растений [САР, т. 5, 6]; г) церковнославянизмы [САР, т. 1].

Шрифты начала XIX века

В начале XIX века появилось большое количество новых типографий и новых шрифтов. Получил распространение так называемый «Круглый» тип шрифта. В целом следует отметить, что графика русского гражданского шрифта этого периода относится к новому классическому стилю. Однако в типографии Академии наук и некоторых других попрежнему применялись шрифты 1788 года. Об этом свидетельствует издание [Зябловский, 1807], [САР, 1806-1822], набранные Дашковской гарнитурой.

Выводы

В результате исследования шрифтового оформления книг второй половины XVIII – начала XIX века было выявлено следующее. Наи-

большее распространение в этот период времени получили шрифты типографий Академии наук и Московского университета, либо сделанные по их образцу. Анализ шрифтов показывает, что московские литеры были сделаны по образцу академических и сначала уступали им по качеству рисунка.

В 50-80 гг. XVIII века широкое распространение получили шрифты: *Парагон Антиква, Гробе Цицеро* и *Миттель Антиква.* Первый использовался для изданий, посвященных торжествам, второй – для «изящной» литературы, научных изданий и журналов.

Издания типографии Академии наук 80-90х годов набирались строгими по рисунку шрифтами, оформление их также было строгим, без особых украшений. В качестве одного их характерных изданий последней четверти XVIII века следует назвать Словарь Академии Российской 1789-1794 гг. А наиболее распространенным шрифтом типографии АН является *Дашковская гарнитура* и *Терция Прямая*.

В 80-х годах многие издания МУ набирались шрифтами типографии АН. В конце XVIII века установилось два типа шрифтов: петербургский и московский. Текстовые шрифты МУ этого периода времени по рисунку можно разделить на три группы: шрифты, сделанные по академическому образцу; модификация академических образцов в сторону приближения их рисунка к шрифтам старого стиля; шрифты «круглого» начертания. Последний стал прототипом московского рисунка шрифта начала XIX века, получившим в дальнейшем широкое распространение.

Литература

Collection …, 1810 — Collection de differents genres d’ecriture pour server de modele. М.: тип. Н. Всеволожского, 1810.

ParaType — ParaType: коллекция кириллических и национальных шрифтов [Электронный ресурс] / — Режим доступа: http://fonts.ru/, свободный. — Яз. Рус. англ.

Бакаревич, 1880 — Бакаревич М. Н Разговоры о физических и нравственных предметах. М., Университетская типография у Ридигера и Клаудия, 1800.

Букварь, 1780 — Букварь для употребления российского юношества. М., Университетская типография у Н. Новикова, 1780

Бюффон, 1789 — Всеобщая и частная естественная история графа де Бюффона. 1789.

Ведомость…, 1796 — Ведомость имеющихся при императорской Академии наук в типографии разных сортов литер. Спб. 1765.

Волкова, 2002 — Волкова Л.А., Решетникова Е.Р. Технология обработки текстовой информации. Часть I. Основы технологии издательских и наборных процессов. Издание второе, исправленное и дополненное: Учебное пособие. М.: Изд-во МГУП, 2002. 306 с.

Дефо, 1762-1764 — Даниель Дефо. Жизнь и приключения Робинзона Круза. 1762-1764.

Добкин, 1985 — Добкин С.Ф. Оформление книги. Редактору и автору. // Типографский шрифт. Современный шрифтовой ассортимент для металлического (отливного) набора. [Электронный ресурс] / 2-е издание, переработанное и дополненное. М.: «Книга», 1985. — Режим доступа: http://obzor.com.ua/dtp/book-design/index.html

Житие…, 1740 — Житие и дела Марка Аврелия. СПб. Тип. Академии наук, 1740.

Зябловский, 1807 — Зябловский Е. Краткое землеописание Российского государства в нынешнем его состоянии для пользы учащихся. СПб. При Имп. АН., 1807.

Коронация Елисаветы, 1744 — Коронация Елисаветы… СПб. Тип. Академии наук, 1744

Марк Витрувий, 1790 — Марк Витрувий. Об архитектуре. Спб., тип. Академии наук, 1790.

Новиков, 1773 — Н.И. Новиков. Древняя Российская Идрография. 1773.

Образцы…, 1796 — Образцы литер, находящихся в типографии Московского университета во время содержания оной Христианом Ридегером и Христофором Клаудием. М., 1796

Образцы…, 1805-А — Образцы литер российских, французских, немецких и прочего, находящегося в типографии Платона Бекетова в Москве. М. 1805

Образцы…, 1805-Б — Образцы литер в типографии Дубровина и Мерзлякова. М. 1805

Образцы…, 1807 — Образцы церковных и гражданских гартовых

	азбук Московской синодальной типографии. М. 1807
Образцы…, 1808	Образцы литер российских, французских, немецких, греческих, еврейских и арабских и прочего, находящегося в типографии Московского университета. М. 1808
Образцы…, 1815	Образцы литер российских, французских, немецких, греческих, еврейских и арабских и прочего, находящегося в типографии Московского университета. М. 1815
Образцы…, 1826	Образцы шрифтов типографии Московского университета. М. 1826.
Показание, 1788	Показание вновь сделанных сего 1788 года при Императорской Академии Наук и в книжной ея типографии имеющихся разных российских и иностранных письмян.
Привилегия…, 1765	Привилегия и устав императорской Академии трех знатнейших художеств. 1765
Пробная книга…, 1748	Пробная книга всем азбукам, знакам и типографским украшениям, которые при императорской типографии Академии наук находятся. СПб, 1748
САР, 1789-1794	Словарь Академїи Россїйской. Т. 1-6. – СПб. 1789-1794.
САР, 1806-1822	Словарь Академии Российской, по азбучному порядку расположеный: в 6 ч. СПб, 1806-1822.
Шицгал, 1959	Шицгал А.Г. Русский гражданский шрифт 1708-1958. – М. 1959
Шицгал, 1974	Шицгал А.Г. Русский типографский шрифт. Вопросы истории и практика применения. – М. 1974.
Шицгал, 1985	Шицгал А.Г. Русский типографский шрифт. Вопросы истории и практика применения. – М. Изд-во «Книга», 1985.

Методические работы

С.А.Кесель

АУТЕНТИФИКАЦИЯ ПОЛЬЗОВАТЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ OTP-ТОКЕНОВ КОМПАНИИ VASCO

Методические указания к лабораторной работе по дисциплине «Защита информации в АСОИУ»

Цель работы.

Ознакомиться с принципами аутентификации пользователей в системах обработки информации на основе одноразовых паролей с использование OTP-токенов компании Vasco.

Порядок выполнения работы.

1. Изучить теоретическую часть.
2. Последовательно выполнить все задания к лабораторной работе.
3. Проверить правильность выполнения заданий.
4. Оформить отчет по лабораторной работе.

Задания к лабораторной работе.

1. Ознакомиться с техническими характеристиками OTP-токенов DigipassGO3 и Digipass 250 .
2. Зайдите на демонстрационную часть сайта компании Vasco http://www.vasco.com/products/applicationguide.html.
3. Запустите тест аутентификационных методов «только ответ» и «запрос-ответ с использованием моделей токенов DigipassGO3 и Digipass 250.

Содержание отчета.

1. Название и цель работы.
2. Задание.
3. Краткое описание моделей токенов и задачи аутентификации пользователей.
4. Протоколы проведенных тестов.

Теоретическая часть

1. Аутентификация с использованием одноразовых паролей

В основе всех методов аутентификации с использованием пароля лежит предположение о том, что только законный пользователь знает свой пароль. При этом идентификатор пользователя злоумышленник зачастую может легко узнать — особенно, если в качестве идентификатора пользователя используется не назначаемый пользователю ID (случайная строка символов, состоящая из букв и цифр), а так называемое *"имя пользователя"*. Так как обычно *"имя-пользователя"* — это различные варианты комбинации имени и фамилии пользователя, то определить его не составляет большого труда. Соответственно, если злоумышленнику удастся узнать еще и пароль пользователя, то ему будет легко представиться этим пользователем.

Недостатки использования многоразовых паролей. Насколько бы не был пароль засекреченным, узнать его иногда не слишком трудно. Злоумышленник может сделать это, используя различные способы атак, основные из которых приведены ниже:

1. *Кража парольного файла.* Злоумышленник может прочитать пароли пользователя из парольного файла или резервной копии.

2. *Атака со словарем.* Злоумышленник, перебирая пароли, производит в (копии) файле паролей поиск, используя слова из большого заранее подготовленного им словаря. Злоумышленник зашифровывает каждое пробное значение с помощью того же алгоритма, что и программа регистрации.

3. *Угадываение пароля.* Исходя из знаний личных данных жертвы, злоумышленник пытается войти в систему с помощью имени пользователя жертвы и одного или нескольких паролей, которые она могла бы использовать

4. *Социотехника.* Объектами этой атаки могут быть пользователи или администраторы. В первом случае злоумышленник представляется администратором и вынуждает пользователя или открыть свой пароль, или сменить его на указанный им пароль. Во втором случае злоумышленник представляется законным пользователем и просит администратора заменить пароль для данного пользователя.

5. *Принуждение.* Для того чтобы заставить пользователя открыть свой пароль, злоумышленник использует угрозы или физическое принуждение.

6. *Подглядывание из-за плеча.* Расположенный рядом злоумышленник смотрит за тем, как пользователь вводит свой пароль.

7. *Троянский конь.* Злоумышленник скрытно устанавливает программное обеспечение, имитирующее обычную регистрационную программу, но собирающее имена пользователей и пароли при попытках пользователей войти в систему. Также злоумышленник может скрытно установить на компьютер пользователя аппаратное средство — sniffer клавиатуры, собирающее информацию, которую вводит пользователь при входе в систему.

8. *Трассировка памяти.* Злоумышленник использует программу для копирования пароля пользователя из буфера клавиатуры.

9. *Отслеживание нажатия клавиш.* Для предотвращения использования компьютеров не по назначению некоторые организации используют программное обеспечение, следящее за нажатием клавиш. Злоумышленник может для получения паролей просматривать журналы соответствующей программы.

10. *Регистрация излучения.* Существуют методы, с помощью которых злоумышленник может перехватывать информацию с монитора путём регистрации излучения, исходящего от электронно-лучевой трубки. Кроме того, существуют способы регистрации не только видеосигналов.

11. *Анализ сетевого трафика.* Злоумышленник анализирует сетевой трафик, передаваемый от клиента к серверу, для извлечения из него имен пользователей и их паролей.

12. *Атака на "золотой пароль".* Злоумышленник ищет пароли пользователя, используемые им в различных системах — домашняя почта, игровые сервера и т.п. Есть большая вероятность того, что пользователь использует один и тот же пароль во всех системах Для каждой из этих атак есть методы защиты. Но большинство из этих защит обладают различными недостатками — некоторые виды защит достаточно дороги (к примеру, борьба с побочным электромагнитным излучением и наводками оборудования), другие создают неудобства для пользователей (правила формирования пароля, использование длинного пароля). Один из вариантов защит от различных атак на аутентификацию по паролю — это переход на аутентификацию с использованием *одноразовых* паролей.

Идея использования одноразовых паролей. Идея применения схем одноразовых паролей явилась заметным шагом вперед по сравнению с использованием фиксированных паролей. При фиксированном пароле узнавший его злоумышленник может повторно его использовать с целью выдачи себя за легального пользователя. Частным решением этой

проблемы как раз и является применение одноразовых паролей: каждый пароль в данном случае используется только один раз.

Одноразовые пароли (One-Time Passwords, OTP) — динамическая аутентификационная информация, генерируемая для единичного использования с помощью аутентификационных OTP-токенов (программных или аппаратных). OTP неуязвим для атаки сетевого анализа пакетов, что является значительным преимуществом перед запоминаемыми паролями. Несмотря на то, что злоумышленник может перехватить пароль методом анализа сетевого трафика, поскольку пароль действителен лишь один раз и в течение ограниченного промежутка времени, у злоумышленника в лучшем случае есть весьма ограниченная возможность представиться пользователем посредством перехваченной информации. Для того чтобы сгенерировать одноразовый пароль, необходимо иметь OTP-токен.

OTP-токен — мобильное персональное устройство, принадлежащее определённому пользователю, генерирующее одноразовые пароли, используемые для аутентификации данного пользователя. Другим важным преимуществом применения аутентификационных устройств (OTP-токенов) является то, что многие из них требуют от пользователя введения PIN-кода. Введение PIN-кода может потребоваться:

- для активации OTP-токена;
- в качестве дополнительной информации, используемой при генерации OTP;
- для предъявления серверу аутентификации вместе с OTP.

В случае, когда дополнительно используется еще и PIN-код, в методе аутентификации используются два типа аутентификационных фактора (на основе "обладания чем-либо" и на основе "знания чего-либо"). В этом случае данный метод будет относиться к двухфакторной аутентификации.

Варианты реализации одноразовых паролей. Существует два основных варианта реализации схемы одноразовых паролей: использование разделяемого списка и вычисление OTP на стороне клиента.

Разделяемый список является простейшей схемой применения одноразовых паролей. В этом случае пользователь и проверяющий используют последовательность секретных паролей, где каждый пароль используется только один раз. Естественно, данный список заранее распределяется между сторонами аутентификационного обмена. Модификацией этого метода является использование таблицы вопросов и ответов, содержащей вопросы и ответы, использующиеся сторонами для проведения аутентификации, причем каждая пара должна применяться

один раз. Существенным недостатком этой схемы является необходимость предварительного распределения аутентифицирующей информации, а после того как выданные пароли закончатся, пользователю необходимо получать новый список. Такое решение, во-первых, не удовлетворяет современным представлениям об информационной безопасности, поскольку злоумышленник может просто-напросто украсть или скопировать список паролей пользователя, после чего легко им воспользоваться. Ну а во-вторых, постоянно ездить за новыми списками паролей вряд ли кому-либо понравится. Вместе с тем, в настоящее время разработано несколько методов реализации технологии одноразовых паролей, исключающих указанные недостатки. В их основе лежит вычисление OTP на стороне клиента и использование различных криптографических алгоритмов.

Суть метода *вычисления OTP на стороне клиента* состоит в следующем:

1. Клиент и сервер имеют разделяемый "секрет".

2. В процессе аутентификации клиент "доказывает" обладание "секретом", не разглашая его.

При аутентификации используются криптографические алгоритмы:

Функции хэширования. Симметричные алгоритмы (криптография с одним ключом) — в этом случае пользователь и сервер аутентификации используют один и тот же секретный ключ.

Асимметричные алгоритмы (криптография с открытым ключом) — в этом случае в устройстве хранится закрытый ключ, а сервер аутентификации использует соответствующий открытый ключ. Обычно в аутентификационных токенах используются *симметричные* алгоритмы. Устройство каждого пользователя содержит уникальный персональный секретный ключ, используемый для шифрования некоторых данных (в зависимости от реализации метода) для генерации OTP. Этот же ключ хранится на аутентификационном сервере, который производит аутентификацию данного пользователя. Сервер шифрует те же данные и сравнивает два результата шифрования: полученный им и присланный от клиента. Если результаты совпадают, то пользователь проходит процедуру аутентификации.

Аутентификационные токены, использующие симметричную криптографию, могут работать в *асинхронном* или *синхронном* режиме. Соответственно методы, используемые аутентификационными токенами, можно разделить на две группы. Асинхронный режим — метод "запрос-ответ".. Синхронный режим: метод "только ответ"; метод "синхронизация по времени"; метод "синхронизация по событию".

Метод "запрос-ответ". В методе "запрос-ответ" OTP является ответом пользователя на случайный запрос от аутентификационного сервера (рис. 1).

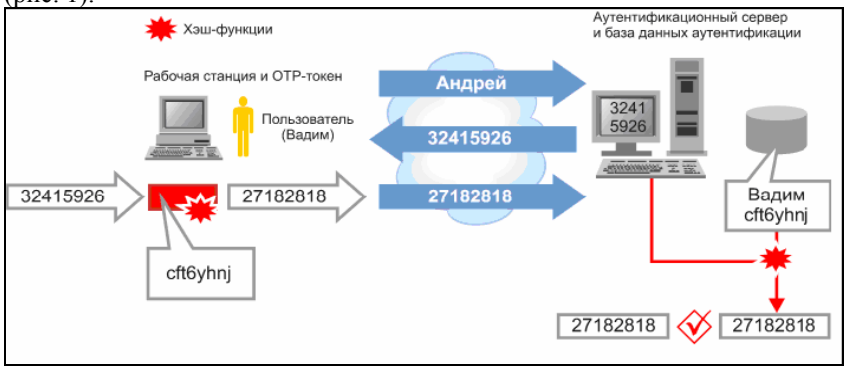


Рис. 1. Метод "запрос-ответ"

Пример прохождения пользователем процедуры аутентификация при использовании аутентификационным токеном метода "запрос-ответ":

1. Пользователь (Вадим) вводит своё имя пользователя на рабочей станции.
2. Имя пользователя передаётся по сети в открытом виде.
3. Сервер аутентификации генерирует случайный запрос ("32415926").
4. Запрос передаётся по сети в открытом виде.
5. Вадим вводит запрос в свой OTP-токен.
6. Аутентификационный токен шифрует запрос с помощью секретного ключа Вадима, в результате получается ответ ("27182818"), и он отображается на экране.
7. Вадим вводит этот ответ на рабочей станции.
8. Ответ передаётся по сети в открытом виде.
9. Аутентификационный сервер находит запись пользователя (Вадима) в аутентификационной базе данных и с использованием хранимого им секретного ключа пользователя зашифровывает тот же запрос.
10. Сервер сравнивает представленный ответ от пользователя ("27182818") с вычисленным им самим ответом ("27182818"). В случае совпадения значений аутентификация считается успешной.

Метод "только ответ". В методе "только ответ" аутентификационное устройство и сервер аутентификации генерируют "скрытый" запрос,

используя значения предыдущего запроса. Для начальной инициализации данного процесса используется уникальное случайное начальное значение, генерируемое при инициализации аутентификационного токена. Схема прохождения пользователем процедуры аутентификации приведена на рис. 2

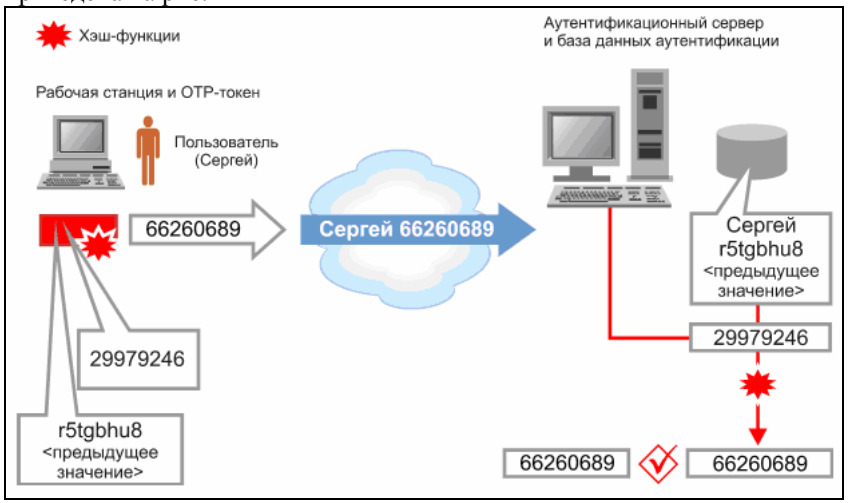


Рис. 2. Метод "только ответ"

Пример прохождения пользователем процедуры аутентификация при использовании аутентификационным токеном метода "только ответ":

1. Пользователь (Сергей) активизирует свой OTP-токен, который вычисляет и отображает ответ на«скрытый» запрос.

2. Пользователь (Сергей) вводит своё «имя пользователя» и этот ответ ("66260689") на рабочей станции.

3. Имя пользователя (Сергей) и ответ ("66260689") передаются по сети в открытом виде.

4. Сервер находит запись пользователя (Сергея), генерирует такой же скрытый запрос и шифрует его с использованием секретного ключа Сергея, получая ответ на свой запрос.

5. Сервер сравнивает представленный ответ от пользователя ("66260689") с вычисленным им самим ответом ("66260689").

Метод "синхронизация по времени". В методе "синхронизация по времени" аутентификационное устройство и аутентификационный сервер генерируют OTP на основе значения внутренних часов. Аутентифи-

кационный токен может использовать не стандартные интервалы времени, измеряемые в минутах, а в специальном длинном интервале времени, обычно он равняется 30 секундам. Процедура аутентификации показана на рис. 3.

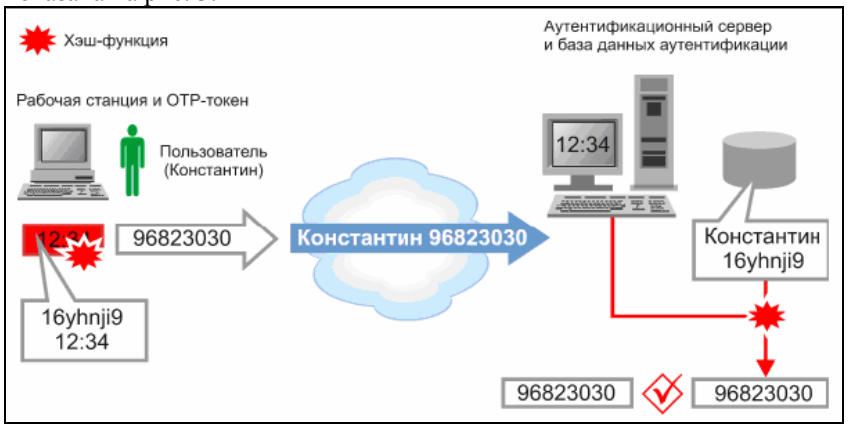


Рис. 3. Метод "синхронизация по времени"

Пример прохождения пользователем процедуры аутентификация при использовании

аутентификационным токеном метода "синхронизация по времени":

1. Пользователь (Константин) активизирует свой OTP-токен, который генерирует OTP ("96823030"), зашифровывая показания часов с помощью своего секретного ключа.
2. Пользователь (Константин) вводит своё имя пользователя и этот OTP на рабочей станции.
3. Имя пользователя и OTP передаются по сети в открытом виде.
4. Аутентификационный сервер находит запись пользователя (Константина) и шифрует показание своих часов с использованием хранимого им секретного ключа пользователя, получая в результате OTP.
5. Сервер сравнивает OTP, представленный пользователем, и OTP, вычисленный им самим.

Метод "синхронизация по событию". В методе "синхронизация по событию" аутентификационный токен и аутентификационный сервер ведут учет количества раз прохождения процедуры аутентификации данным пользователем, и на основе этого числа генерируют OTP (рис.4).

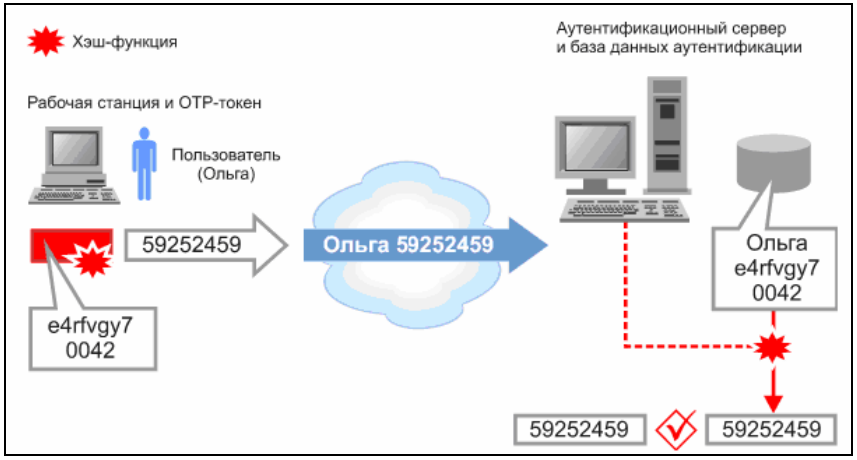


Рис. 4. Метод "синхронизация по событию"

Пример прохождения пользователем процедуры аутентификация при использовании аутентификационным токеном метода "синхронизация по событию":

1. Пользователь (Ольга) активизирует свой OTP-токен, который генерирует OTP ("59252459"), зашифровывая значения количества раз прохождения процедуры аутентификации данным пользователем, с помощью своего секретного ключа.
2. Пользователь (Ольга) вводит своё имя пользователя и этот OTP на рабочей станции.
3. Имя пользователя и OTP передаются по сети в открытом виде.
4. Аутентификационный сервер находит запись пользователя (Ольги) и шифрует значение количества раз прохождения процедуры аутентификации данным пользователем с использованием хранимого им секретного ключа пользователя, получая в результате OTP.
5. Сервер сравнивает OTP, представленный пользователем, и OTP, вычисленный им самим.

Группа OATH и система HOTP. Система одноразовых паролей HOTP (HMAC-based One-Time Password System) была разработана в 2005 году в рамках инициативы группы открытой аутентификации OATH (Open AuTHentification) и описана в документе RFC 4226. Данная система основана на концепции OTP-аутентификации с синхронизацией *по событию* (см. п. "Метод "синхронизация по событию"). Для генера-

ции одноразового пароля используется алгоритм HMAC (Hashed Message Authentication Code).

Этот алгоритм публичен и доступен для изучения любыми специалистами. Он обеспечивает возможность использования с ключами широкого спектра программного обеспечения, в том числе и серверного.

Система HOTP предусматривает возможность задания "окна" попыток аутентификации и синхронизацию сервера аутентификации с OTP-токеном после успешного прохождения аутентификации.

Значение одноразового пароля вычисляется по формуле:

HOTP(K,C) = Truncate(HMAC-SHA-1(K,C)), где:

K — секретный ключ;

C — счётчик числа прохождений процедуры аутентификации;

HMAC-SHA-1 — процедура генерации HMAC, основанная на функции хэширования SHA-1;

Truncate — процедура усечения 20-тибайтного значения HMAC-SHA-1 до 4-х байт.

Область применения одноразовых паролей. Одноразовые пароли широко применяются в следующих областях:

1. Аутентификация с устройств, не имеющих возможности подключения смарт-карт / USB-ключей (КПК, смартфоны и др.).
2. Аутентификация со временных рабочих мест, где нет возможности устанавливать ПО и среда не является контролируемой.
3. Аутентификация при удаленном доступе через незащищенные каналы связи (при возможном перехвате аутентификационной информации)
4. Аутентификация по голосовому каналу (например, удаленное управление банковским счетом).

2 Устройства строгой аутентификации пользователей Digipass компании VASCO Data Security

Одним из наиболее уязвимых мест при подключении пользователя к защищаемым ресурсам является аутентификация пользователей с применением статических паролей. При использовании статических паролей возникают многочисленные угрозы, связанные с возможностью перехвата пароля и несанкционированного входа в систему с его использованием. Проникновение в систему в этом случае крайне сложно обнаружить и распознать, так как в данном случае факт несанкционированного доступа неотличим от подключения легального пользователя, что приводит к невозможности принять эффективные меры по предупреждению таких вариантов взлома системы.

Возможность перехвата пароля имеется в как в частных (локальных) сетях, и в сетях общего пользования. Особенно велика возможность перехвата паролей в системах, использующих Интернет в качестве среды передачи данных. Интернет, как среда передачи данных, создает дополнительные возможности за счет проникновения на компьютер пользователя программ – троянов и такой угрозы, как фишинг.

Так например, при банковских операциях, проводимых с помощью Интернет-банкинга, надежная аутентификация пользователей становится одной из основных проблем при удаленном управлении банковским счетом.

Самой распространенной мерой защиты является периодическая смена паролей, которые либо создаются системным администратором, либо самими пользователями. К сожалению, эта мера далеко не всегда достигает поставленной цели. В том случае, если пароли создаются администратором, необходим защищенный от возможного перехвата канал передачи новых паролей пользователям. Более того, если пароль создается с помощью программных средств и является бессмысленным набором символов, букв и цифр, то, как правило, пользователи записывают пароль и хранят его в удобном для себя месте. В том случае, если создание нового пароля доверяется пользователю, то большинство из них используют свои персональные данные, что не обеспечивает высокой стойкости данных паролей.

Еще одним распространенным методом защиты от перехвата паролей является использование одноразовых паролей, который пользователь получает в виде книжки с готовыми паролями, каждый из которых может быть использован во время одного сеанса работы с системой. Во время следующего сеанса будет использован следующий пароль и так далее. Основным недостатком данного решения является высокая стоимость печати и распространения таких одноразовых паролей.

Система DigiPass предназначена для строгой аутентификации[13] пользователей и находит широкое применение в локальных сетях, системах удаленного доступа, электронной и мобильной коммерции, банковских системах дистанционного обслуживания. Оборудование

[13] под строгой аутентификацией подразумевается использование любых двух из трех следующих признаков пользователя:

1. Нечто, что имеется у пользователя – карта доступа, ключ.
2. Нечто, что известно только пользователю (пароль, PIN-код).
3. Нечто, что присуще только пользователю – отпечатки пальцев, радужная оболочка (биометрия).

DigiPass используется конечными пользователями – корпоративными и частными клиентами банков, обеспечивая надежную авторизацию и защиту от таких угроз, как фишинг и троянские программы при использовании удаленного доступа к управлению банковским счетом.

Системы DigiPass можно разделить на две большие группы: аппаратные средства генерации одноразовых паролей и программные средства, обеспечивающие эмуляцию аппаратных средств для различных операционных систем

Аппаратные средства DigiPass серии DigiPass GO1, GO3, GO5, GO6 предназначены исключительно для генерации одноразовых паролей.

DigipassGO1 DigipassGO3 DigipassGO5 DigipassGO6

Устройства DigiPass 250, 260 и DigiPass Pro 300 обеспечивают как генерацию одноразовых паролей (в двух режимах), так и вычисление Message Authentication Code (MAC), позволяющего проверить отсутствие изменений в передаваемой в ходе транзакции информации.

Digipass 250 Digipass 260 Digipass Pro 300

Устройства DigiPass 550, 560, 580 в функциональном отношении полностью аналогичны вышеуказанному оборудованию DigiPass Pro 300, но имеют более удобный для пользователя интерфейс и обеспечивают поддержку сообщений на любом языке.

DigiPass 550 DigiPass 560 DigiPass 580

Особенностью устройств Digipass 800,810, 820 и 850 является то, что данные устройства предназначены для выполнения всех вышеописанных функций и являются автономными кардридерами для микропроцессорных карт стандарта EMV. Преимущество устройств этой серии в том, что персонализация устройства на этапе производства отсутствует, а для генерации одноразовых паролей используется персональная информация, размещенная в памяти микропроцессорной карты. Тем самым создается возможность для быстрой поставки оборудования заказчику и возможность безболезненной передачи оборудования третьим лицам. Digipass 850 может быть использован как в автономном, так и в подключенном режиме, используя для взаимодействия с компьютером шину USB.

Digipass 800 **Digipass 810** **Digipass 820** **Digipass 850**

Для проверки одноразовых паролей и MAC на серверной стороне размещается программная компонента решения, позволяющая проверить их подлинность. Для аутентификации пользователей в локальных сетях компания VASCO предлагает готовые решения на основе продукта Vacman middleware, для продуктов электронной коммерции имеется набор процедур (Vacman Controller), которые интегрируются в соответствующие программные продукты поставщика решения.

Строгая аутентификация пользователей достигается за счет применения одноразовых паролей в двух основных режимах: запрос-ответ и одноразовый пароль, генерируемый по датчику реального времени.

В первом случае пользователю, при попытке подключения к ресурсам с ограниченным доступом предлагается ввести некоторое числовое значение в имеющееся у него устройство семейства DigiPass и, получив от устройства ответ, отображаемый на дисплее токена, ввести его в систему аутентификации.

Во втором случае вместо статического пароля для аутентификации пользователя используется одноразовый пароль, который предоставляется в распоряжение пользователя токеном DigiPass.

Принцип действия устройств DigiPass основан на генерации одноразовых паролей с применением симметричного криптографического алгоритма для кодирования уникальной информации, содержащейся в токене, в сочетании с датчиком реального времени, который также вхо-

дит в состав оборудования токена. В качестве алгоритма шифрования используется 3DES (AES для некоторых моделей). Результат шифрования отображается на дисплее токена и вводится вручную в систему аутентификации.

Программное обеспечение серверной части решения производит на основании имени пользователя генерацию одноразового пароля и сравнивает полученное от пользователя значение с вычисленным сервером одноразовым паролем. Далее, на основании результатов сравнения, программное обеспечение возвращает результат аутентификации в виде «да-нет». Основываясь на результатах аутентификации, следующий по иерархии уровень программного обеспечения позволяет принять решения о допуске или отказе в допуске пользователя к ресурсам вычислительной системы.

Оборудование семейства Digisass не предусматривает возможности какого-либо изменения функциональности устройств конечными пользователями и имеет встроенные механизмы защиты от любых попыток получения доступа к системе генерации паролей. Любая попытка вскрытия устройства приводит к уничтожению ключа шифрования и невозможности дальнейшего использования данного устройства.

Устройства Digipass серии 800 не содержат каких-либо криптографических ключей и используют для аутентификации ключи, размещенные на карте стандарта EMV в соответствии со спецификацией EMV CAP, являясь по существу интерфейсом между картой и пользователем. Для данного типа устройств предусмотрена конструкция, позволяющая пользователю определить по внешнему виду устройства факт попытки физического вторжения в устройство.

Рекомендуемая литература

1. *Б. Шнаер.* Прикладная криптография. Протоколы, алгоритмы, исходные тексты на языке Си. М.:ТРИУМФ,2003. — 816с.

2. Математические и компьютерные основы криптологии: Учеб. пособие/Ю.С.Харин и др. Минск: Новое Знание,2003. — 382с.

3. *Ю.В. Романец и др.* Защита информации в компьютерных системах и сетях Изд. 2-е. М: Радио и связь, 2001. — 376с.

Ю.Н.Филиппович, А.Ю.Филиппович, Г.А.Черкасова

СОЗДАНИЕ БАЗЫ ДАННЫХ СЛОВАРЯ АКАДЕМИИ РОССИЙСКОЙ 1789–1794 ГГ.

Методические указания к курсовой работе по дисциплине «Семиотика информационных технологий»

Цель курсовой работы:

Приобретение навыков разработки лингвистических баз данных.

Задачи курсовой работы:

— изучение материалов лекционных и практических занятий по данной дисциплинам «Семиотика информационных технологий» и «Лингвистическое обеспечение АСОИУ»;
— приобретение практических навыков работы с текстовыми процессорами и СУБД;
— приобретение навыков анализа и изучение методов и приемов исследования ЕЯ описания предметной области (ПО) в процессе выполнения заданий курсовой работы;
— приобретение знаний и навыков по оформлению результатов анализа и исследования ПО при оформлении курсовой работы.

Задания курсовой работы:

1. Подготовка словарных статей к вводу в базу.
2. Декомпозиция словарных статей.
3. Разработка лингвистической базы данных.
4. Разработка технологического процесса заполнения базы данных и автоматизированный ввод словарных статей.
5. Исследование эффективности технологического процесса заполнения базы данных.

Порядок выполнения курсовой работы:

1. Получить у преподавателя вариант курсовой работы (скопировать текстовые файлы).
2. Последовательно выполнить задания 1–5 курсовой работы.
3. Оформить результаты курсовой работы.
4. Апробировать результаты курсовой работы у преподавателя.

Пояснения, рекомендации и требования к выполнению курсовой работы.

После получения варианта курсовой работы рекомендуется внимательно прочитать и изучить словарные и методические материалы. Хорошее их знание в дальнейшем позволит качественно выполнить задания курсовой работы. Рекомендуется выполнять задания курсовой работы в указанной последовательности.

Задание 1.

Подготовка словарных статей к вводу в базу данных состоит в их записи в таблицы Microsoft Word (2003). При подготовке нужно выполнить следующие действия:

1. Установить шрифты;
2. Выделить цветом элементы словарных статей: желтым – заголовочные слова, зеленым – грамматические характеристики и пометы, бирюзовым – цитаты, красным – шифр источника (см. пример 2);
3. Заполнить таблицу 1 «Заголовочные слова»: <№ столбца>, <№ словарной статьи>, <заголовочное слово>, <словарная статья полностью> (см. пример 3).
4. Заполнить таблицу 2: <№ столбца>, <№ словарной статьи>, <заголовочное слово>, <№ значения>,<значение> (см. пример 4).

Задание 2.

Декомпозиция словарных статей состоит в переносе ее элементов в таблицы 3 и 4, имеющие следующие структуры (см. примеры 5 и 6):

Таблица 3. «Грамматические характеристики заголовочных слов» —
<№ столбца>, <№ словарной статьи>, <заголовочное слово>, <варианты>,<Часть речи>
<Род>, <Число>, <Вид>, <Помета>.

Таблица 4. «Характеристики значений» —
<№ столбца>, <№ словарной статьи>, <заголовочное слово>, <№значения>
<Толкование>, <*Примеры*>, <*Цитата*>, <Источник>.

Задание 3.

Перед началом разработки лингвистической базы данных отрезка САР необходимо ознакомиться с методическими материалами: 1) М.И.Чернышева. «Состав и структура Словаря академии Российской». (файл Чернышева.doc. 2) Предисловие. … (файл Предисловие.doc). Для более глубокого понимания проблемы создания баз данных гнездовых словарей рекомендуется ознакомиться со статьей А.Н.Тихонов. «Пер-

вый гнездовой словарь и его роль в развитии русской лексикографии» (файл Тихонов.doc). Структура разработанной базы данных должна отражать особенности заданного отрезка САР. Среда разработки базы данных Microsoft Access (2003).

Задание 4.

При разработке технологического процесса заполнения базы данных должны быть выделены и описаны этапы, процедуры и отдельные операции. Технологический процесс должен быть представлен в виде схемы с пояснениями.

Задание 5.

Исследование эффективности технологического процесса ввода информации состоит в определении основных характеристик — ресурсных затрат и ограничений на выполнение этапов, процедур и отдельных операций. Данное исследование может проводиться в процессе реализации предыдущих заданий. Результаты могут быть получены методами сплошного или выборочного исследования.

Оформление курсовой работы.

Результаты выполнения курсовой работы должны быть оформлены в виде технической документации, состоящей из двух документов: расчетно-пояснительной записки и приложения.

Расчетно-пояснительная записка должна быть оформлена в соответствии с требованиями, предъявляемыми к оформлению отчетов по НИР. При этом должны выполняться требования соответствующих ГОСТ.

Объем расчетно-пояснительной записки не регламентируется. Работа сдается в напечатанном виде и в виде файла форматов *.doc на CD. Текстовая часть должна содержать ссылки на используемую литературу с указанием страниц издания. Список используемой литературы приводится в соответствии с требованиями ГОСТ. Текст оформленной курсовой работы не должен содержать орфографических ошибок. Текст должен быть отформатирован с параметрами: отступы — нулевые, красная строка — 1 см, межстрочный интервал — одинарный, выравнивание — по ширине. Текст должен содержать переносы. Все страницы текста должны содержать верхний колонтитул с фамилией и шифром группы исполнителя работы, а также нумерацию страниц с размещением над колонтитулом по центру. Размер основного шрифта — 12.

В отчете рекомендуется иметь следующие разделы:

Введение. Пять разделов, заголовки которых должны соответствовать формулировкам первых пяти заданий курсовой работы. Заключение. Литература.

В разделе "*Введение"* перечисляются цели и задачи курсовой работы, дается краткое описание отрезка САР.

В *пяти разделах*, приводятся результаты выполненных заданий.

В "*Заключении"* приводятся выводы по результатам исследований, выполненных в курсовой работе.

В разделе "*Литература"* указываются использованные источники.

Титульный лист курсовой работы должен содержать:

в верхней части листа

Московский государственный технический университет им. Н.Э. Баумана
кафедра "Системы обработки информации и управления"

в середине листа

Создание базы данных
Словаря Академии Российской 1789–1794 гг.
Том X.
Столбцы XXX–XXX.

Папка: САР XXX_XXX.

Расчетно-пояснительная записка
курсовой работы по дисциплине
"Семиотика информационных технологий"
студента группы ИУ5-GG.
Фамилия Имя Отчество.

Преподаватель: к.т.н.,доц. Ю.Н.Филиппович

в нижней части листа

Москва, 2006 г.

Апробация курсовой работы.

Для апробации курсовой работы у преподавателя необходимо сдать для анализа и оценки: расчетно-пояснительную записку, рабочие материалы курсовой работы (исходные, промежуточные и результирующие файлы данных), выданный текстовый материал.

Расчетно-пояснительная записка сдается в стандартных полиэтиленовых файлах. Рабочие материалы курсовой работы сдаются на подписанных CD. Все файлы должны быть собраны в папку с именем ФА-

МИЛИЯ. В папке должен быть файл с именем Readme.txt, с описанием всех файлов рабочих материалов курсовой работы.

Оценка курсовой работы осуществляется по следующим правилам: “удовлетворительно” — при условии полного и качественного выполнения одного первого пункта;

“хорошо” — при условии полного и качественного выполнения пунктов 1–2; “отлично” — при условии полного и качественного выполнения всех пунктов задания.

В случае сдачи работы после срока ее оценка снижается на один балл, но не ниже «удовлетворительно».

Срок сдачи курсовой работы 12 неделя семестра (24–29 апреля).

Примеры

Пример 1.

Исходный текст отрезка САР 1789–1794 гг..

2

ЗАБОБОНЫ, бонъ. с. ж. множ. Бред-|ни, сказки по большей частпи основан-|ныя на суевѣрїи, и не доспойныя |вниманїя умнаго человѣка. *Сей ли |лутъ спасенїя, яко ломрагенъ забо-|бонами не знаешь, что ты долженъ |еси Богу, государю, отечеству?* Ѳеоф. |Прокоп.

Ты жилъ ло тѣмъ законамъ,|Которые лисалъ;|Смѣялся забабонамъ. М. Лом.

Забабонщина, щины. с. ж. простонар. |Рѣчи и разсказы пустыя, основанныя |на суевѣрїи. *Какой забабонщинѣ ты |вѣришь!*

ЗАБОТА, пы. с. ж. Попеченїе о чемъ,|пщанїе, спаранїе соединенное съ без-|покойспвомъ и печалїю. *Домашнїя|заботы сокрушили его здоровье. Весь|вѣкъ живу въ заботахъ.*

Межъ тѣмъ на Горизонтъ заря восходитъ,|И смертныхъ бдѣнїя, и тьмы заботъ возводитъ.

В. Пепр. Ен.

3

Заботливый, вая, вое. *Заботливъ*,|ва, во. прил. Спарапельный, попечи-|пельный; копорый о чемъ, или о комъ|печепся, забопипся. *Заботливая хо-|зяйка. Заботливъ о лодчиненныхъ.*

Заботливо. нар. Спарапельно, попечи-|пельно. *Этотъ хозяинъ заботливо|за всемъ смотритъ.*

Заботливость, спи. с. ж. Свойспво по-|печипельнаго, спарапельнаго, пща-|пельнаго человѣка; попечипель-|носпь. *Съ его заботливостїю всякой|трудъ преодолѣть можно.*

Заботный, ная, ное. *Заботенъ*, пна, пно. |прил. Съ хлопопами соединенный. |*Заботная должность. Заботной день.*

Заботно. нар. Хлопопно, многодѣльно,|суепливо. *Всякому крестьянину въ|лѣтнюю лору заботно.*

Забогу, пишь, забопилъ, пипь. гл. д. |просп. Причиняю кому забопу, хло-|попы. *Напрасно ты меня такою без-|дѣлицею заботишь.*

Забогусь, ся, пишься, забопился, забо-|пипься. гл. возв. Спараюся, пекуся о|комъ или о чемъ; забопу имѣю. *Забо-|гусь о воспитанїи дѣтей. Добрый че-|ловѣкъ о другихъ заботится.*

4

Беззаботливый, вая, вое. *Беззаботливъ*, |ва, во. прил. Безпечный, забопливоспи |неимѣющїй; небрегущїй о помъ, чпо |надобно. *Такъ беззаботливъ, что ни |за чемъ не смотритъ.*

Беззаботливо. нар. Безъ хлопопъ, без-|печно. *Беззаботливо вѣкъ провож-|даетъ.*

Беззаботливость, спи. с. ж. просп. |Безпечноспь, нерадивоспь. *Ты своею |беззаботливостїю проронилъ ето |мѣсто.*

Беззаботный, ная, ное. *Беззаботенъ*,|пна, пно. прил. 1) Безпечный, нера-|дивый. *Беззаботный человѣкъ себѣ|вреденъ, другимъ безполезенъ.* 2)|Безпечальный, свободный опъ вся-|кихъ хлопопъ, попеченїй.|*Ты* (кузнечикъ) *скачешь и поешь, | свободенъ, беззаботенъ.* М. Л.

Беззаботно. нар. 1) Безпечно. 2) Безпе-|чально. *Живу беззаботно.*

О Б Е З З А Б О Ч И В А Ю, ешь, забопилъ, бочу,|бочивать, забопить. гл. д. проспо-|нар. Свобождаю кого опъ заботы, опъ|хлопотъ. *Ежели ты дѣло ето воз-|мешь на себя, такъ совершенно меня|обеззаботишь.*

Обеззабочиваюсь, ешься, бочуся, забо-|пился, обеззабопиться. гл. возвр. |проспонар. Освобождаюсь опъ хло-|потъ. *Пристроивъ дѣтей къ мѣсту|обеззаботился.*

Пример 2.
Примеры выделения цветом элементов словарных статей.

ЗАБОБОНЫ, бонъ. с. ж. множ. Бред-|ни, сказки по большей частии основан-|ныя на суевѣріи, и не досгойныя |вниманія умнаго человѣка. *Сей ли |лутъ спасенія, яко ломрачень забо-|бонами не знаешь, что ты долженъ |еси Богу, государю, отечеству?* Феоф. |Прокоп.

Ты жилъ ло тѣмъ законамъ,|Которые писалъ;|Смѣялся забабонамъ. М. Лом.

Беззаботный, ная, ное. *Беззаботенъ*,|тна, тно. прил. 1) Безпечный, нера-|дивый. *Беззаботный человѣкъ себѣ|вреденъ, другимъ безполезенъ.* 2)|Безпечальный, свободный опъ вся-|кихъ хлопотъ, попеченій.|*Ты* (кузнечикъ) *скачешь и поешь, | свободенъ, беззаботенъ.* М. Л.

О Б Е З З А Б О Ч И В А Ю, ешь, забопилъ, бочу,|бочивать, забопить. гл. д. проспо-|нар. Свобождаю кого опъ заботы, опъ|хлопотъ. *Ежели ты дѣло ето воз-|мешь на себя, такъ совершенно меня|обеззаботишь.*

Пример 3.
Результирующая таблица 1. «Заголовочные слова»

№ столбца	№ словарной статьи	заголовочное слово	словарная статья полностью
2	1	ЗАБОБОНЫ	ЗАБОБО’НЫ, бонъ. с. ж. множ. Бред-|ни, сказки по большей часпи основан-|ныя на суевѣрїи, и не доспойныя |вниманїя умнаго человѣка. *Сей ли |путь спасенїя, яко ломрасенъ забо-|бонами не знаешь, что ты долженъ |еси Богу, государю, отечеству?* Θеоф. |Прокоп. *Ты жилъ ло тѣмъ зако-намъ,|Которые писалъ;|Смѣялся забобонамъ.* М. Лом.
2	2	*Забобонщи-на*	*Забобо□нщина*, щины. с. ж. проспонар. |Рѣчи и разсказы пуспыя, основанныя |на суевѣрїи. *Какой забобонщинѣ ты |вѣришь!*
2	3	ЗАБОТА	ЗАБО’ТА, пы. с. ж. Попеченїе о чемъ,|пщанїе, спаранїе соединенное съ без-|покойспвомъ и печалїю. *Домашнїя|заботы сокрушили его здоровье. Весь|вѣкъ живу въ заботахъ. Межъ тѣмъ на Горизонтъ заря вос-ходитъ.|И смертныхъ дѣнїя, и тьмы заботъ возводитъ.* В. Пепр. Ен.
3	1	*Заботливый*	*Забо□тливый*, вая, вое. *Забо□тливъ*,|ва, во. прил. Спараπельный, попечи-

			|тельный; копорый о чемЪ, или о комЪ|печется, забопится. *Заботливая хо-|зяйка. Заботливъ о подчиненныхъ.*
3	2	*Заботливо*	*Забо□тливо*. нар. Спарапельно, попечи-|тельно. *Этотъ хозяинъ заботливо|за всемъ смотритъ.*
3	3	*Заботливость*	*Забо□тливость*, спи. с. ж. Свойспво по-|печипельнаго, спарапельнаго, пща-|тельнаго человѣка; попечипель-|носпь. *Съ его заботливостїю всякой|трудъ преодолѣть можно.*
3	4	*Заботный*	*Забо□тный*, ная, ное. *Забо□тенъ*, тна, тно. |прил. Съ хлопотами соединенный. |*Заботная должность. Заботной день.*
3	*5*	*Заботно*	*Заботно*. нар. Хлопотно, многодѣльно,|суепливо. *Всякому крестьянину въ|лѣтнюю пору заботно.*
3	*6*	*Забочу*	*Забочу*, тишь, забопилъ, тить. гл. д. |просп. Причиняю кому забопу, хло-|поты. *Напрасно ты меня такою без-|дѣлицею заботишь.*
3	7	*Забочусь*	*Забочусь, ся*, тишься, забопился, забо-|титься. гл. возв. Спараюся, пекуся о|комЪ или о чемЪ; забопу имѣю. *Забо-|чусь о воспитанїи дѣтей. Добрый че-|ловѣкъ о другихъ заботится.*
4	1	*Беззаботливый*	*Беззаботливый*, вая, вое. *Беззаботливъ*, |ва, во. прил. Безпечный, заботливоспи|неимѣющїй; небрегущїй о томЪ, чпо |надобно. *Такъ беззаботливъ, что ни |за чемъ не смотритъ.*
4	2	*Беззаботливо*	*Беззаботливо*. нар. БезЪ хлопотЪ, без-|печно. *Беззаботливо вѣкъ провож-*

			|дает*ъ*.
4	3	*Беззаботливость*	*Беззаботливость*, спи. с. ж. просп. |Безпечность, нерадивость. *Ты своею|беззаботливостїю лроронилъ это|мѣсто*.
4	4	*Беззаботный*	*Беззаботный*, ная, ное. *Беззаботенъ*,|тна, тно. прил. 1) Безпечный, нера-|дивый. *Беззаботный человѣкъ себѣ|вреденъ, другимъ безполезенъ*. 2)|Безпечальный, свободный отъ вся-|кихъ хлопотъ, попеченїй.|*Ты* (кузнечикъ) *скачешь и поешь,	свободенъ, беззаботенъ*. М. Л.
4	5	*Беззаботно*	*Беззаботно*. нар. 1) Безпечно. 2) Без-пе-|чально. *Живу беззаботно*.
4	6	О Б Е З З А Б О Ч И В А Ю	О Б Е З З А Б О Ч И В А Ю, ешь, забопилъ, бочу,|бочивать, забопить. гл. д. проспо-|нар. Свобождаю кого отъ заботы, отъ|хлопотъ. *Ежели ты дѣло его воз-|мешь на себя, такъ совершенно меня|обезпокоишь*.
4	7	*Обезпокоиваюсь*	*Обезпокоиваюсь*, ешься, бочуся, забо-|пился, обезпокоиться. гл. возвр. |проспонар. Освобождаюсь отъ хло-|потъ. *Пристроивъ дѣтей къ мѣсту|обезпокоился*.

Пример 4.

Результирующая таблица 2. «Значения заголовочных слов».

№ столбца	№ словарной статьи	заголовочное слово	№ значения	значения
2	1	ЗАБОБОНЫ	1	Бред-|ни, сказки по большей часпи основан-|ныя на суевѣрїи, и не доспойныя |вниманїя умнаго человѣка. *Сей ли |лутъ сласенїя, яко ломраченъ забо-|бонами не знаешь, что ты долженъ |еси Богу, государю, отечеству?* Ѳеоф. |Прокоп. *Ты жилъ ло тѣмъ законамъ,|Которые лисалъ;|Смѣялся забабонамъ.* М. Лом.
2	2	*Забабонщина*	1	Рѣчи и разсказы пусныя, основанныя |на суевѣрїи. *Какой забабонщинѣ ты |вѣришь!*
2	3	ЗАБОТА	1	Попеченїе о чемъ,|пщанїе, спаранїе соединенное съ без-|покойствомъ и печалїю. *Домашнїя|заботы сокрушили его здоровье. Весь|вѣкъ живу въ заботахъ.* *Межъ тѣмъ на Горизонтъ заря восходитъ,|И смертныхъ дѣнїя, и тьмы заботъ возводитъ.* В. Пепр. Ен.
3	1	*Заботливый*	1	Спарапельный, попечи-|тельный; копорый о чемъ, или о

				комЪ|печется, заботится. *Заботливая хо-|зяйка. Заботливъ о подчиненныхъ.*
3	2	*Заботливо*	1	Спарапельно, попечи-|тельно. *Этотъ хозяинъ заботливо|за всемъ смотритъ.*
3	3	*Заботливость*	1	Свойство по-|печительнаго, спарапельнаго, пща-|тельнаго человѣка; попечитель-|ность. *Съ его заботливостїю всякой|трудъ преодолѣть можно.*
3	4	*Заботный*	1	Съ хлопотами соединенный. |*Заботная должность. Заботной день.*
3	5	*Заботно*	1	Хлопотно, многодѣльно,|суетливо. *Всякому крестьянину въ|лѣтнюю пору заботно.*
3	6	*Забочу*	1	Причиняю кому забопу, хло-|поты. *Напрасно ты меня такою без-|дѣлицею заботишь.*
3	7	*Забочусь*	1	Спараюся, пекуся о|комЪ или о чемЪ; забопу имѣю. *Забо-|чусь о воспитанїи дѣтей. Добрый человѣкъ о другихъ заботится.*
4	1	*Беззаботливый*	1	Безпечный, забопливо-спи|неимѣющїй; небрегущїй о томЪ, что |надобно. *Такъ беззаботливъ, что ни |за чемъ не смотритъ.*
4	2	*Беззаботливо*	1	БезЪ хлопотЪ, без-|печно. *Беззаботливо вѣкъ провож-|даетъ.*
4	3	*Беззаботливость*	1	Безпечность, нерадивость. *Ты своею|беззаботливостїю проро-*

				нилъ ето\мѣсто.
4	4	*Беззаботный*	1	Безпечный, нера-\дивый. *Беззаботный человѣкъ себѣ\вреденъ, другимъ безполезенъ.*
4	4	*Беззаботный*	2	Безпечальный, свободный отъ вся-\кихъ хлопотъ, попеченій.*Ты* (кузнечикъ) *скачешь и поешь, \ свободенъ, беззаботенъ.* М. Л.
4	5	*Беззаботно*	1	Безпечно.
4	5	*Беззаботно*	2	Безпе-\чально. *Живу беззаботно.*
4	6	**ОБЕЗ-ЗАБО-ЧИ-ВАЮ**	1	Свобождаю кого отъ заботы, отъ\хлопотъ. *Ежели ты дѣло ето воз-\мешь на себя, такъ совершенно меня\обезпзаботишь.*
4	7	*Обезпзаботи-ваюсь*	1	Освобождаюсь отъ хло-\потъ. *Пристроивъ дѣтей къ мѣсту\обезпзаботился.*

Пример 5.

Таблица 3. Грамматические характеристики заголовочных слов.

№ столбца	№ словарной статьи	заголовочное слово	Варианты	Ч. речи	Род	Число	Гл. харр-ки	Помепы
2	1	ЗАБОБОНЫ	бонъ.	с.	ж.	мно ж.		
2	2	*Забабонщи-на*	щины.	с.	ж.			проспонар.
2	3	ЗАБОТА	ты.	с.	ж.			

3	1	*Заботливый*	вая, вое. *Забот-ливъ*,|ва, во.	прил.				
3	2	*Заботливо*		нар.				
3	3	*Заботли-вость*	спи.	с.	ж.			
3	4	*Заботный*	ная, ное. *Забо-тенъ*, тна, тно.	прил.				
3	5	*Заботно*		нар.				
3	6	*Забогу*	тишь, забо-тилъ, тить.	гл.			д.	просп.
3	7	*Забогусь*	*ся*, тишься, забо-тился, забо-|титься.	гл.			возв.	
4	1	*Беззаботли-вый*	вая, вое. *Безза-бот-ливъ*,	прил				

			ва, во.					
4	2	*Беззаботливо*		нар.				
4	3	*Беззаботливость*	спи.	с.	ж.			просп.
4	4	*Беззаботный*	ная, ное. *Беззаботенъ*, пна, пно.	прил.				
4	5	*Беззаботно*		нар.				
4	6	**ОБЕЗЗАБОЧИВАЮ**	ешь, забопилъ, бочу, бочивапь, забопить.	гл.			д.	проспо-нар.
4	7	*Обеззаботиваюсь*	ешься, бочуся, забопился, обеззабопить-ся.	гл.			возвр.	проспонар.

Пример 6.
Таблица 4. Характеристики значений.

№ столбца	№ словарной статьи	заголовочное слово	№ значения	Толкование	*Примеры*	*Цитата*	Источник
2	1	ЗАБОБОНЫ	ı	Бред-|ни, сказки по большей часпи основан-|ныя на суевѣрїи, и не доспойныя |вниманїя умнаго человѣка.		*Сей ли |лутъ сласенї я, яко ломра ченъ забо- |обнами не знаешь, что ты должен з |еси Богу, госуда рю, отечес тву?*	Ѳеоф. Проко п.
2	1	ЗАБОБОНЫ	ı	Бред-|ни, сказки по большей часпи основан-|ныя на суевѣрїи, и не доспойныя		*Ты жилъ ло тѣмъ законамъ,|Ко торые писалъ*	М. Лом.

				|вниманїя умнаго человѣка.		;|*Смѣялся забабонамъ.*	
2	2	*Забабонщина*	1	Рѣчи и разсказы пустыя, основанныя |на суевѣрїи.	*Какой забабонщинѣ ты |вѣришь!*		
2	3	ЗАБОТА	1	Попеченїе о чемЪ,|пщанїе, спараніе соединенное сЪ без-|покойствомЪ и печалїю.	*Домашнїя|заботы сокрушили его здоровье. Весь|вѣкъ живутъ въ заботахъ.*	*Межъ тѣмъ на Горизонтѣ заря восходитъ.| И смертныхъ дѣнїя. и тьмы заботъ возводитъ.*	В. Непр. Ен.
3	1	*Заботливый*	1	Спарапель-ный, попе-чи-|тельный; копорый о чемЪ, или	*Заботливая хо-|зяйка. Заботливъ о*		

				о комъ|печетс я, забо- тится.	*лодчи- ненныхъ.*		
3	2	*Заботливо*	1	Спарапельно, попечи- |тельно.	*Этотъ хозяинъ заботливо|за всемъ смотритъ.*		
3	3	*Заботливость*	1	Свойство по- |печительнаго, спара- тельнаго, пща- |тельнаго человѣка; попечи- тель- |ность.	*Съ его заботливостїю всякой|трудъ преодолѣть можно.*		
3	4	*Заботный*	1	Съ хлопотами соединенный.	*Заботная должность. Заботной день*		.
3	5	*Заботно*	1	Хлопотно, многодѣльно, |суетливо.	*Всякому крестьянину въ|лѣтнюю ло-*		

					ру заботно.		
3	6	Забочу	1	Причиняю кому заботу, хлопоты.	Напрасно ты меня такою безлѣпицею заботишь.		
3	7	Забочусь	1	Спараюся, пекуся о комъ или о чемъ; заботу имѣю.	Заботчусь о воспитанїи дѣтей. Добрый человѣкъ о другихъ заботится.		
4	1	Беззаботливый	1	Безпечный, заботливости неимѣющїй; небрегущїй о томъ, что надобно.	Такъ беззаботливъ, что ни за чемъ не смотритъ.		
4	2	Беззаботливо	1	Безъ хлопотъ, безпечно.	Беззаботливо вѣкъ провож-		

					|даетъ.		
4	3	*Беззаботливость*	1	Безпечность, нерадивость.	*Ты своею|беззаботливостїю лроронилъ ето|мѣсто.*		
4	4	*Беззаботный*	1	Безпечный, нера-|дивый.	*Беззаботный человѣкъ себѣ|вреденъ, другимъ безлолезенъ.*		
4	4	*Беззаботный*	2	Безпечальный, свободный отъ вся-|кихъ хлопотъ, попеченїй.		*Ты (кузнечикъ) скачешь и лоешь.	свободенъ, беззаботенъ.*	М. Л.
4	5	*Беззаботно*	1	Безпечно.			
4	5	*Беззаботно*	2	Безпе-|чально.	*Живу беззаботно.*		
4	6	О Б Е З - З А Б О - Ч И В А Ю	1	Свобождаю кого отъ	*Ежели ты*		

				заботы, опѢ|хло-потѢ.	*дѣло его возъ|мешь на себя, такъ совершенно меня|обезз аботишь.*		
4	7	*Обеззаботива-ясь*	1	Освобождаюсь опѢ хло-|потѢ.	*Пристроивъ дѣтей къ мѣсту|обезза-ботился.*		

Научно-образовательный кластер
«Computer Linguistics–Artificial Intelligence–Multimedia»
(CLAIM)

Научное издание

Интеллектуальные технологии и системы
Сборник учебно-методических работ и статей аспирантов и студентов

Выпуск 8

Составитель: *Филиппович Юрий Николаевич*

Редактирование и корректура выполнена авторами статей

Формат 60 × 84/16. Бумага офсетная. Гарнитура «Таймс». Печать на ризографе.
Усл.п.л. 20.5. Тираж 100 экз.