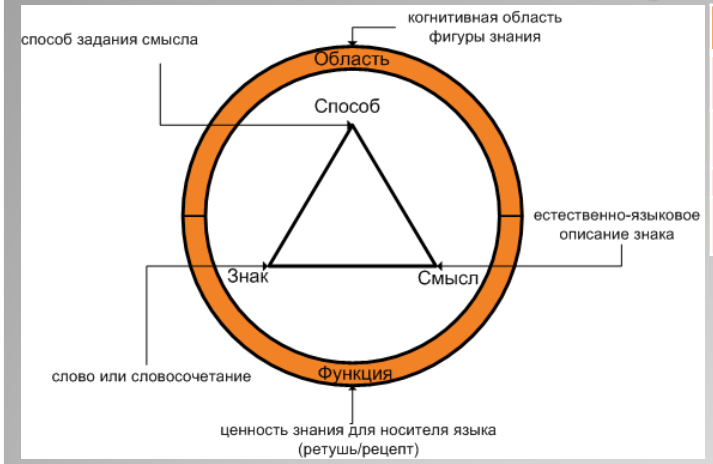
Методы и алгоритмы автоматизации научных исследований психолингвистических моделей вербального сознания

На материалах ассоциативного и когнитивного экспериментов

Сиренко Александр Викторович
2009

Когнитивный эксперимент

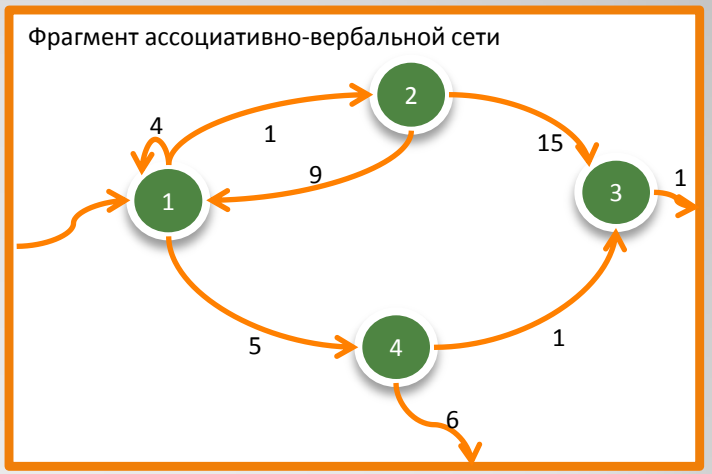


Знак	Формула смысла	Способ	Область	Функция
акцент	Особенность выговора человека, говорящего не на родном языке.	Дефиниция	Рецепт	язык
Абдулла	против него воевал красноармеец Сухов	Фрейм	Ретушь	история
абрикос	Фруктовый плод	Множество	Рецепт	ботаника
Або	Шведское название финского г. Турку	Смена кода	Ретушь	география

Переход от формулы смысла к знаку: пассивный режим языкового сознания

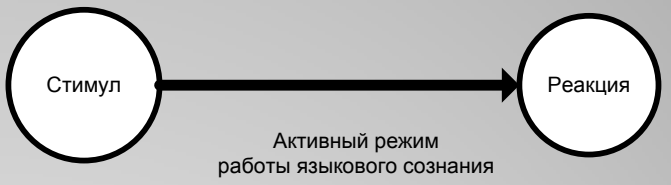
Ассоциативный эксперимент

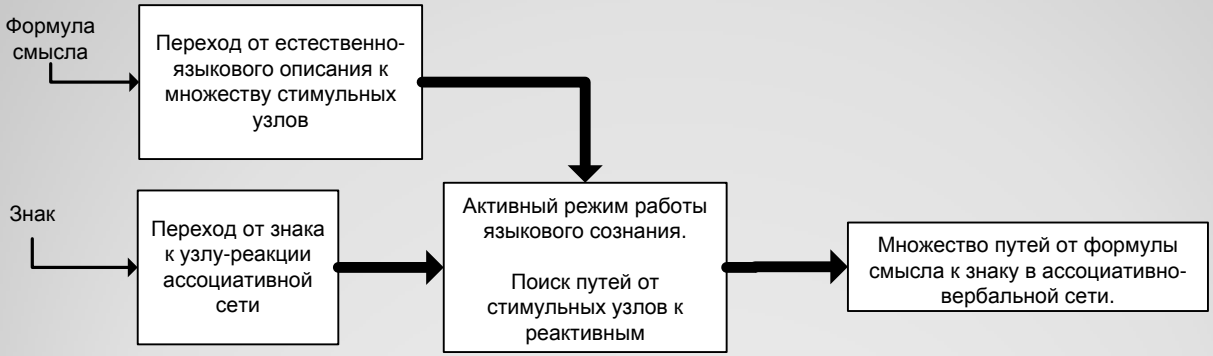
Стимул	Реакция	Число повторов ассоциации
ворота	открыты	6
ворота	в ад	4
ворота	дверь	4
ворота	открывать	4
ворота	рая	3
ворота	закрыты	3
ворота	резные	2
ворота	к знанию	1
ворота	квадратные	1



Переход от стимула к реакции: активный режим языкового сознания

Когнитивный и ассоциативный эксперименты





Интеграция экспериментов

Цель

- Автоматизация моделирования языкового сознания носителя русского языка с использованием материалов когнитивного и ассоциативного психолингвистических экспериментов;
- Фиксация результатов моделирования в лексикографической форме;

Задачи

- Анализ существующих подходов к построению когнитивных лингвистических систем.
- Организация хранения и обработки лингвистических данных.
- Лемматизация ассоциативно-вербальной сети с устранением частичной омонимии.
- Поиск путей в ассоциативно-вербальной сети между интенсиональными элементами фигур знания.
- Анализ структуры ассоциативно-вербальной сети методом кластеризации.
- Создание программного комплекса, реализующего информационную технологию.
- Автоматизированное построение словарных статей лингвокультурного тезауруса русского языка.

Общие сведения

Актуальность

- Востребованность тезаурусного подхода описания языкового сознания;
- Готовностью значимых результатов психолингвистических экспериментов, составляющих информационную основу работы.

Научная новизна

- Методика интеграции когнитивных фигур знания и ассоциативно-вербальной сети для моделирования работы языкового сознания;
- Алгоритм поиска в ассоциативно-вербальной сети;
- Методика кластеризации ассоциативно-вербальной сети.

Практическая ценность

- Разработанные методики и алгоритмы обработки результатов ассоциативного и когнитивного экспериментов;
- Реализация в виде программного комплекса;
- Прототипы словарных статей Лингвокультурного тезауруса русского языка;

Компоненты работы

Ассоциативные тезаурусы

- Ассоциативный эксперимент университета южной Флориды, США.

5 000 стимулов, 72 000 реакций.

Идеографические словари

- Тезаурус английских слов и фраз П.М.Роже

3 уровня иерархии, 1000 словарных статей

- Идеографический словарь О.С.Баранова

7 уровней иерархии, 100 000 лексем в 7500 статьях

Семантические сети

- Шведский ассоциативный тезаурус;

Понятия образуют иерархическую структуру. Максимально полный словник шведского языка.

- WordNet, EuroWordNet, VisualThesaurus;

WordNet -английский язык,
EuroWordNet - испанский, итальянский, датский, английский,
VisualThesaurus - визуализация WordNet в виде сетевой структуры.

- Проект lexfn.com

Связи WordNet 1.6, рифмы, созвучные слова, библиографические данные s9.com,
анаграммы, частоты совместной встречаемости (ассоциации).

Аналоги и прототипы

Проект	Представление	Элементы	Отношения	Фиксируемая картина мира	Моделируемый режим работы вербального сознания
Ассоциативный эксперимент университета южной Флориды	Лемматизированная ассоциативно-вербальная сеть	Узлы сети-леммы	Ассоциации	Наивная	Активный
Идеографические словари Баранова, Роже	Словарные статьи распределены по иерархической структуре семантических групп.	Словарная статья представлена ЕЯ описанием понятия	Образованы иерархией семантических групп	Научная	Пассивный
Шведский ассоциативный тезаурус	Словарные статьи образуют иерархическую структуру.	Словарная статья представлена ЕЯ описанием понятия	Отношения частное-общее	Научная	Пассивный
WordNet, EuroWordNet, VisualThesaurus	Основу составляют синсеты – объединенные синонимичными отношениями понятия	Словарная статья представлена ЕЯ описанием понятия	Синонимы, антонимы, гиперонимы, гипонимы.	Научная	Пассивный
Lexical Free Net	Основу составляют синсеты – объединенные синонимичными отношениями понятия	Словарная статья представлена ЕЯ описанием понятия	Связи WordNet, рифмы, созвучные слова, библиографические данные s9, анаграммы, частотные ассоциации	Научная	Пассивный
Лингвокультурный тезаурус русского языка	Лемматизированная ассоциативно-вербальная сеть, фигуры знания, дополнительные словари	Пятикомпонентная фигура знания, леммы ассоциативно-вербальной сети	Ассоциации, синонимы, отношения, образованные когнитивными областями фигур знания	Научная – фигуры знания. Наивная – ассоциативно-вербальная сеть.	Активный – ассоциативно-вербальная сеть. Пассивный – фигуры знания.

Сводная таблица аналогов и прототипов

Орфографический словарь iSpell

- Электронный орфографический словарь «Корректор» (120 тыс. словоформ);
- Грамматический словарь русского языка: словоизменение. Зализняк А.А. (110-тыс. словоформ);
- Сводный словарь современной русской лексики 1991г. (170 тыс. словоформ);
- Русский орфографический словарь. Под ред. В. Лопатина (180 тыс. словоформ);
- Прочие источники.

Всего лемм: 127 000, словоформ: 1 300 000.

Число	Исходная сеть	Лематизированная сеть
Узлов	103 000	63 700
Связей	457 000	394 000
Стимулов	6 670	3 830

Число омонимичных словоформ: 1030 (частичная омонимия)

Связей для обработки: 25 000.

СЛОВОФОРМА # ЛЕММА1->РЕЛЯТОР_ЛЕММЫ1 СЛОВОФОРМА_СТИМУЛ->СЛОВОФОРМА_РЕАКЦИЯ->ЧИСЛО_СВЯЗЕЙ … # ЛЕММА2->РЕЛЯТОР_ЛЕММЫ2 СЛОВОФОРМА_СТИМУЛ->СЛОВОФОРМА_РЕАКЦИЯ->ЧИСЛО_СВЯЗЕЙ …	Пример: шерсти # шерстить->* # шерсть->* комок->шерсти->1 клубок->шерсти->7

Файл обработки омонимичной словоформы.

Лематизация ассоциативно-вербальной сети

Множество состояний ассоциативной сети $S = \{A_1, A_2, \dots A_n\}$.

1. $P(A_i|T_0) = p_{0i}$;
2. $P_{ij} = P(A_j, T_{k+1}| A_i, T_k)$, $k = 1, 2, \dots \infty$;

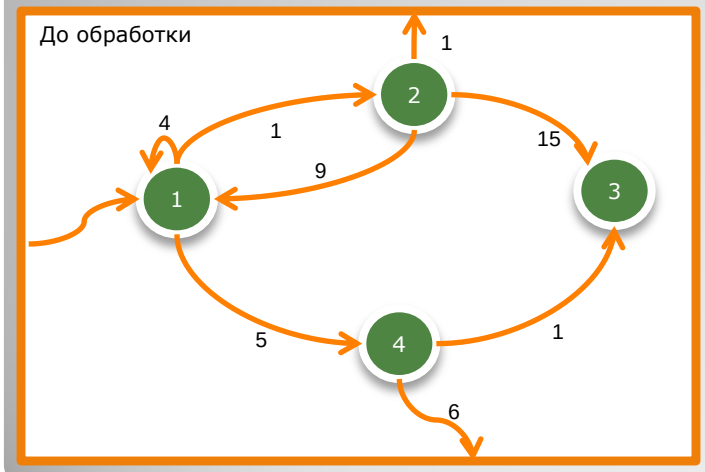
Вероятность перехода системы из Ai в Aj в следующий момент времени:

$$P_{ij} = \frac{C_{ij}}{\sum_{j=0}^{N} C_{ij}}$$

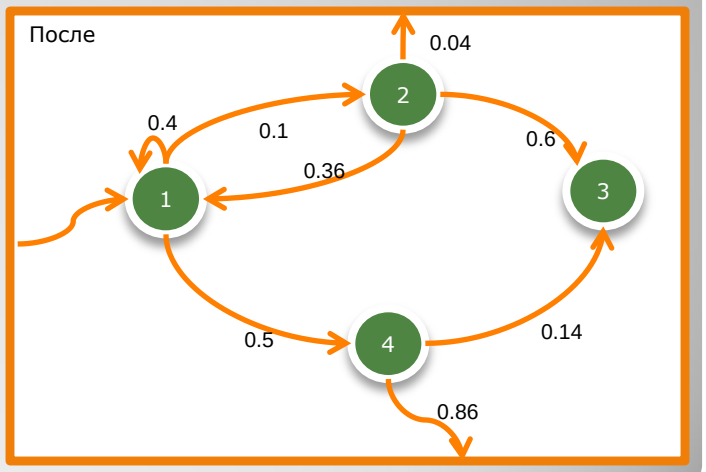
Стохастическая матрица переходов M_{p1}

$$M_{p1} = \begin{pmatrix} p_{00} & p_{01} & \dots & p_{0n} \\ p_{10} & p_{11} & \dots & p_{1n} \\ \dots & \dots & \dots & \dots \\ p_{n0} & p_{n1} & \dots & p_{nn} \end{pmatrix}$$

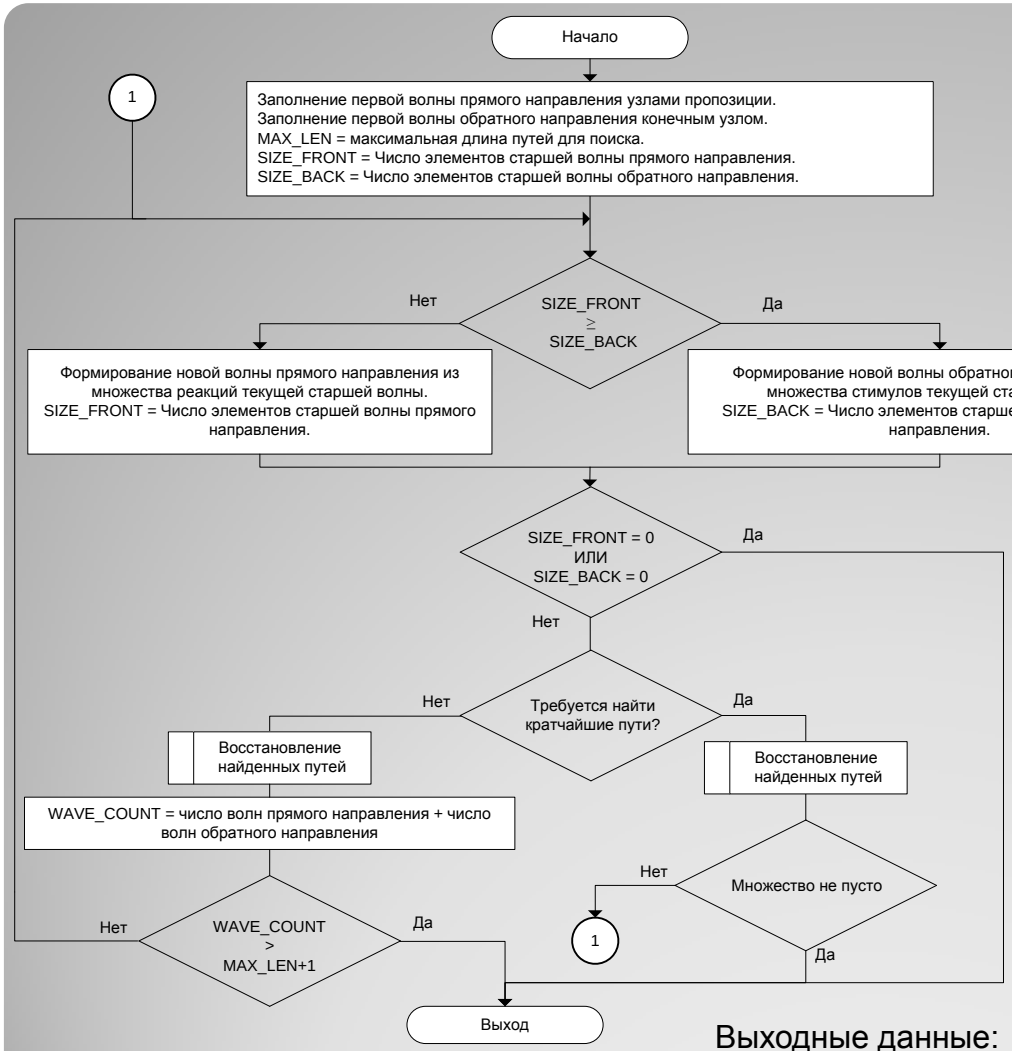
Уравнение Колмогорова-Чепмена:
$M_{pn} = (M_{p1})^n$



Переход от числа ассоциаций к вероятности перехода



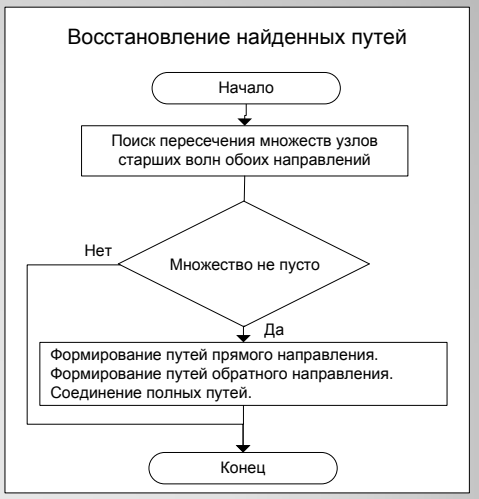
Обработка ассоциативно-вербальной сети



Входные данные:

1. Узлы формулы смысла.
2. Узел знака.
3. Максимальная длина путей.
4. Условие поиска кратчайших путей.
5. Списки смежных вершин

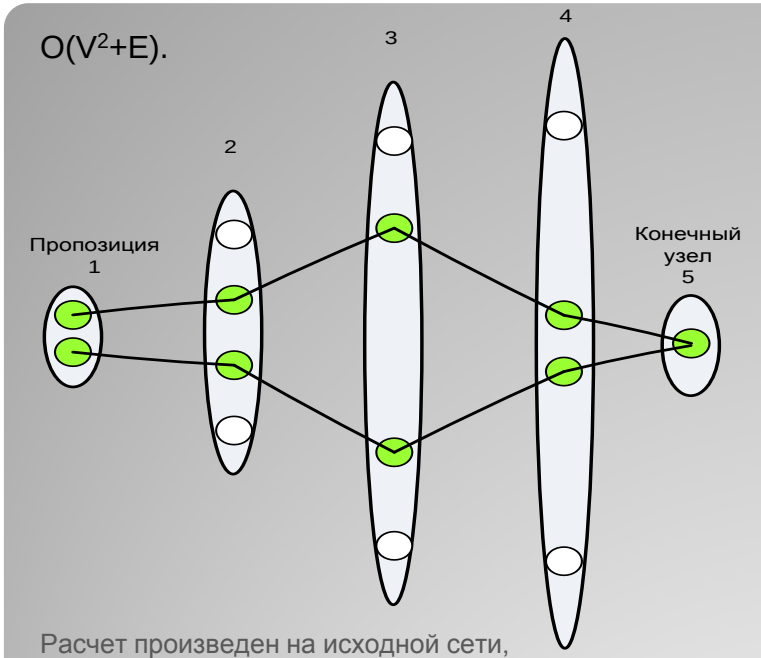
узлов ассоциативной сети.



Выходные данные:

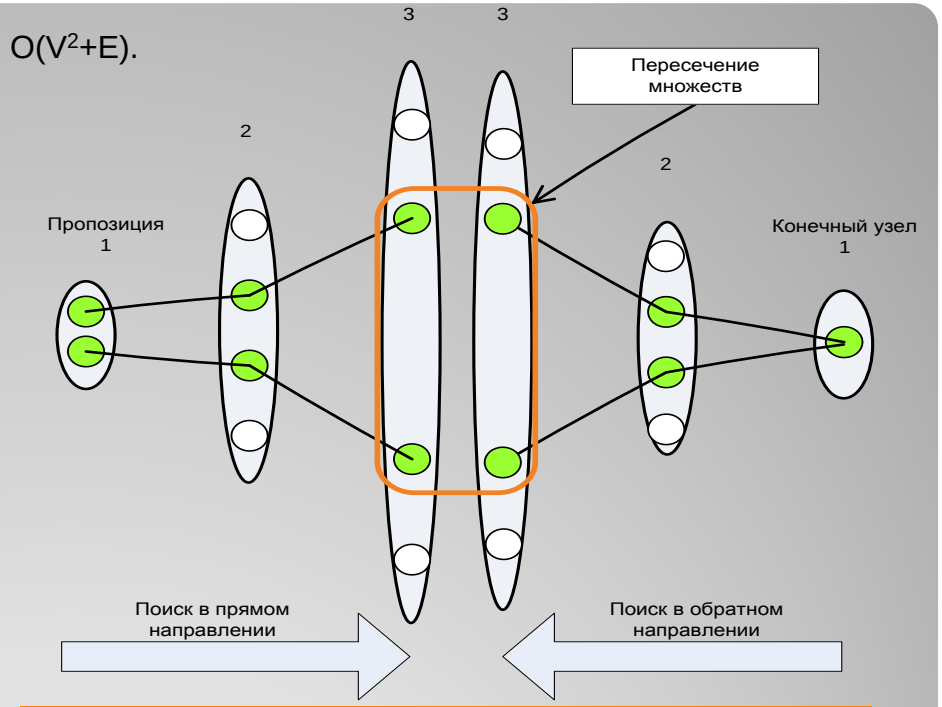
1. Множество путей требуемой длины.

Двунаправленный алгоритм поиска



Расчет произведен на исходной сети, без лемматизации

Номер шага	Число достижимых узлов
1	2
2	105
3	3607
4	42252
5	92526
6	99850



Номер шага	Направление	Число достижимых узлов
1	Прямое	2
	Обратное	1
2	Обратное	2
3	Обратное	34
4	Прямое	104
5	Обратное	1419
6	Прямое	3606

Рассмотрено 2,2% от числа узлов, обработанных волновым поиском

Сравнение с волновым алгоритмом

Требования к метрике сети:

- Зависимость от списка общих узлов;
- Симметричность;
- Зависимость от непосредственной связи узлов;
- При отсутствии общих реакций и связей, расстояние должно равняться некоторому L_{max}, при увеличении связности элементов стремиться к L_{min}. Промежуточные значения должны располагаться от L_{min} до L_{max};

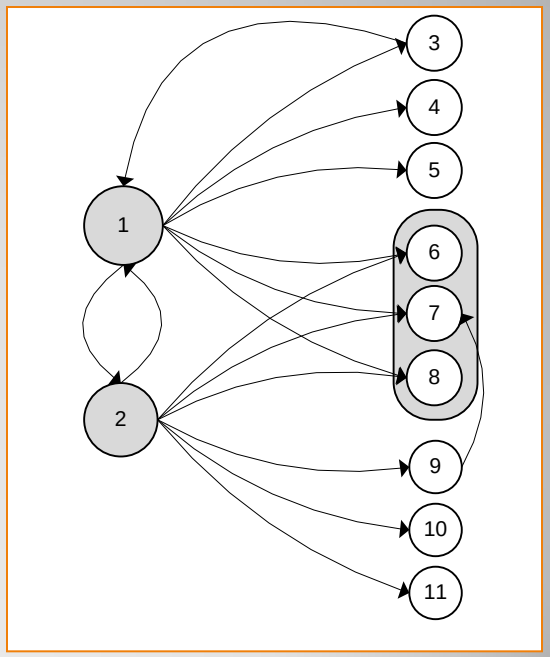
Сила связи узлов: $F_{ij} = P_{ij} + P_{ji} + \sum_{k=1}^{N} Min(P_{ik}, P_{jk})$

Расстояние между узлами:

$$L_{ij} = L_{max} - \frac{F_{ij}}{2}(L_{max} - L_{min})$$

Для узлов 1 и 2:

$F_{12} = P_{12} + P_{21} + Min(P_{16},P_{26}) + Min(P_{16},P_{26}) + Min(P_{16},P_{26})$;

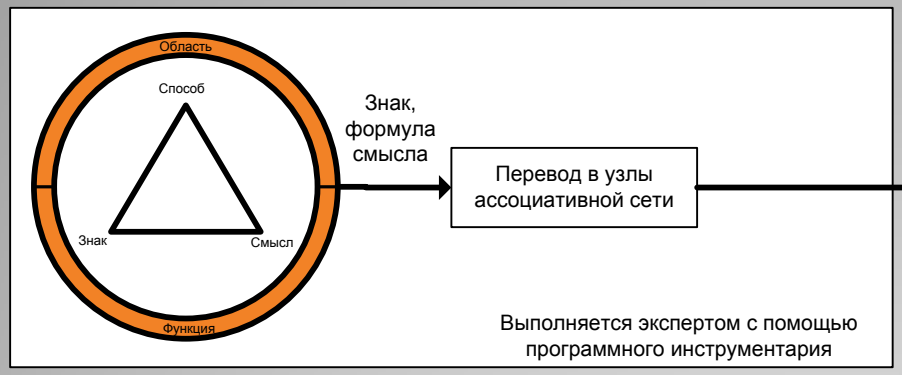


Кластеризация ассоциативно-вербальной сети

Этапы формирования словарных статей:

№	Описание	Входные данные	Выходные данные
1	Перевод формулы смысла в множество стимульных узлов ассоциативной сети	БД, включающая фигуры знания, лемматизированную ассоциативную сеть, словарь стимулов и лемм. Идентификаторы целевых фигур знания.	XML-документы, соответствующие идентификаторам обработанных фигур знания. Документы содержат знак, значимые элементы формулы смысла соответствующие и соответствующие им узлы ассоциативной сети. Далее обозначаются Нормализованной фигурой знания.
2	Вычисление путей в ассоциативной сети. Предварительная фильтрация результатов.	XML-документы нормализованных фигур знания. Параметры поиска.	XML-документы результатов поиска. Требования изложены в описании соответствующего этапа.
3	Формирование документа установленного образца.	XML-документы результатов поиска. Параметры формирования словарных статей, установленных пользователем.	HTML-документ установленного образца, содержащий словарные статьи.
4	Ручной контроль результатов	HTML-документ словарных статей.	HTML-документ словарных статей для публикации.

Формирование словарных статей



Знак, формула смысла

Перевод в узлы ассоциативной сети

XML документы фигур знания, приведенных к узлам ассоциативной сети

Выполняется экспертом с помощью программного инструментария

Поиск путей в ассоциативной сети

Параметры поиска: Ограничение длины путей, допустимость циклов.

XML документы результатов поиска

Формирование словарных статей

Параметры формирования словарных статей: Шаблон статьи, фильтры путей.

Словарные статьи

Шаблон

Знак – Формула смысла.
Способ, Область, Функция.
Стимул$_1$, стимул$_2$,…стимул$_n$. Реакция.
Стимул$_1$ -(P)> Узел-(P)>… -(P)> Реакция;
Стимул$_1$ -(P)> Узел-(P)>… -(P)> Реакция;
…
Стимул$_2$ -(P)> Узел-(P)>… -(P)> Реакция;
..
Стимул$_n$ -(P)> Узел-(P)>… -(P)> Реакция;

Образец

абрикос – Фруктовый плод.
Множество, Рецепт, Ботаника.
плод, фруктовый. абрикос.
плод->(0.011) фрукт ->(0.071) спелый->(0.02) абрикос;
плод->(0.0050) кислый ->(0.0020) персик->(0.019) абрикос;
…
фруктовый->(0.049) компот ->(0.0090) персик->(0.019) абрикос;

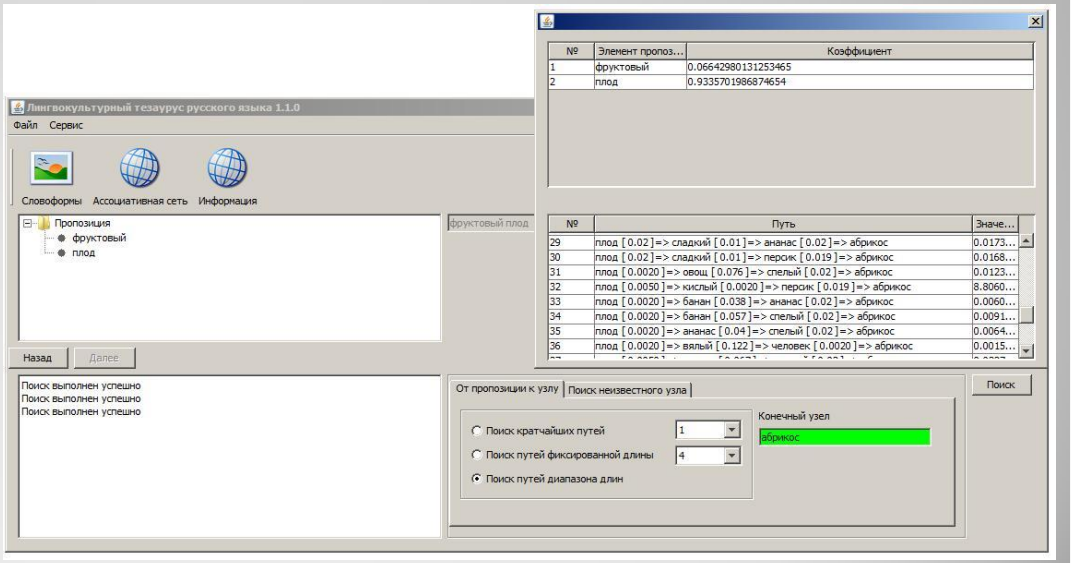
Образец словарной статьи

Требования к средствам разработки:

- Доступность интегрированной среды разработки;
- Реляционная СУБД, с поддержкой хранимых процедур;
- Отсутствие необходимости покупать лицензии конечным потребителем;
- Доступность на рынке web-хостинга с применяемой СУБД.
- Наличие совместимых компонент визуализации, сетевого доступа, работы с XML.

В результате используются:

- Программная платформа: Java SE 6.0;
- Интегрированная среда разработки: Sun Microsystems NetBeans 6.7;
- СУБД PostgreSQL 8.4.



Реализация информационной технологии

Произведено

- Проанализированы существующие подходы моделирования вербального сознания и их характерные представители.
- Осуществлено проектирование базы данных и ее реализация в реляционной СУБД.
- Проведена лемматизация ассоциативно-вербальной сети с устранением частичной омонимии.
- Предложен алгоритм поиска путей в ассоциативно-вербальной сети между интенсиональными элементами фигур знания.
- Предложен метод расчета метрики ассоциативно-вербальной сети для кластеризации.
- Программный комплекс реализует разработанные методы и алгоритмы.

Необходимо

- Разработать методику кластеризации ассоциативно-вербальной сети на основе предложенной метрики.
- Провести анализ структуры ассоциативно-вербальной сети.
- Разработать методику построения словарных статей.
- Реализовать алгоритмы и методики в программном комплексе исследования.

Заключение