Сиренко А.В.

2009.10.19

Текст выступления в научной школе SoftLine 2009

«Методы и алгоритмы автоматизации научных исследований психолингвистических моделей вербального сознания»

1. Введение

Актуальность. Психолингвистические эксперименты, как правило, масштабны. Методы обработки и анализа результатов являются предметом творческой работы экспериментаторов, и часто носят локальный характер. Вместе с тем, на сегодняшний день разработаны и находят себе применение методы и алгоритмы интеллектуального анализа данных, призванные работать с большими объемами данных, подчас плохо структурированными. Технологии извлечения знаний производят поиск зависимостей, скрытых большими объемами разнородной информации.

Перспективным также является построение формальных моделей психолингвистических экспериментов, с целью использования соответствующего математического аппарата.

Когнитивный и ассоциативный эксперименты. В работе языкового сознания человека условно можно выделить два режима работы – активный, осуществляющий переход от знака к смыслу и обратный ему пассивный [Караулов, Филиппович 2005, с.5-6]. Для фиксации результатов работы языкового сознания в этих режимах ранее были проведены два эксперимента: ассоциативный и когнитивный.

Ассоциативный эксперимент представлен ассоциативно-вербальной сетью. В основе проводимого исследования лежит ассоциативно-вербальная сеть, сформированная в ассоциативном эксперименте по формированию Русского ассоциативного словаря [Караулов,1994-1998].

Дополнительно используются словарь синонимов Абрамова, а также лемматизатор словоформ.

Интеграция экспериментов. В основе работы лежит гипотеза о возможности совмещения активного и пассивного режимов работы языкового сознания, путем интеграции результатов когнитивного и ассоциативного экспериментов. Метод интеграции представлен на слайде. Предлагается перевести интенсиональные элементы фигуры знания в форму узлов ассоциативной сети следующим образом: формула смысла преобразуется во множество стимульных узлов, а знак – реактивных. Затем должен осуществляться поиск путей в ассоциативной сети от формулы смысла к знаку, то есть в направлении, обратном зафиксированному в фигуре знания пассивному режиму работы языкового сознания.

Цель. Целью проводимого исследования является автоматизация моделирования языкового сознания носителя русского языка с использованием материалов когнитивного и ассоциативного

психолингвистического эксперимента. Фиксация результатов моделирования в лексикографической (словарной) форме.

Задачи. Задачи, подлежащие решению:

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

Научная новизна. В диссертационной работе подлежат разработке и обоснованию:

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

Практическая ценность. Практическая ценность работы заключается в разработанных в составе информационной технологии методиках и алгоритмах автоматизированной обработки результатов ассоциативного и когнитивного экспериментов. Данные методики и алгоритмы реализуются программным комплексом, спроектированным по клиент-серверной технологии. Программный комплекс создан с применением общедоступных программных компонент, с его помощью предполагается создание прототипов словарных статей печатного издания «Лингвокультурного тезауруса русского языка».

2. Аналоги и прототипы.

Основу работы составляют ассоциативный и когнитивный эксперименты. Ассоциативно-вербальная сеть в том смысле, какой в нее вкладывает ассоциативный эксперимент, представлена для русского языка в форме печатного издания [Караулов, 1994], а также ряда электронных ресурсов [АСНИ АЭ]. Существуют подобные англоязычные эксперименты. Способом изучения ассоциативно-вербальной сети, как правило, становятся фрагменты сети, списки реакций стимулов в разных языках, социальных группах.

Фигуры знания когнитивного эксперимента не привлекались в качестве носителя знаний о пассивном режиме работы при создании информационных систем тезаурусного типа.

Поэтому, информационная технология, в которую входят разрабатываемые методы и алгоритмы, не имеет прямых аналогов в силу оригинальности используемых психолингвистических экспериментов. Вместе с тем, предмет исследования – вербальное сознание человека – позволяет выделить класс систем, описывающих либо моделирующих его. Это определяет положение рассматриваемых психолингвистических экспериментов в области моделирования вербального сознания.

Классы систем, моделирующих вербальное сознание человека:

1. Ассоциативные тезаурусы.
    1.1. Ассоциативный эксперимент университета южной Флориды, США.
2. Идеографические словари.
    2.1. Тезаурус английских слов и фраз П.М. Роже.
    2.2. Идеографический словарь Баранова О.С.
3. Семантические сети;
    3.1. Шведский ассоциативный тезаурус.
    3.2. WordNet, EuroWordNet, VisualThesaurus.
    3.3. Проект lexfn.ru.

Ассоциативные тезаурусы

Ассоциативный эксперимент университета южной Флориды, США.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms.

Условия ассоциативного эксперимента [Nelson, McEvoy, Schreiber, 1998] практически соответствуют аналогичному русскоязычному ассоциативному эксперименту, описанному в [Черкасова, 2004]. Ассоциативная сеть лемматизирована, эксперимент проводился в большом масштабе (более 5000 стимулов, более 72000 реакций).

Идеографические словари

Суть идеографического словаря заключается в расположении слов (соответствующих им словарных статей), в порядке, определяемом их смысловой близостью. Этим идеографический словарь отличается от традиционного, группирующего элементы согласно их форме, а не содержанию [Морковкин, 1970].

Тезаурус английских слов и фраз П.М. Роже

С этого тезауруса начинаются идеографические словари в их классическом понимании. Тезаурус Роже был издан в 1852 году [Guttenberg]. Тезаурус Роже множество раз перерабатывался и переиздавался [Roget, 1988].

Группы, на которые разбиваются словарные статьи, содержат три уровня иерархии. Общее число статей: 1000.

Идеографический словарь Баранова О.С.

Идеографический словарь Баранова является самым крупным русскоязычным идеографическим словарем, идейно продолжающим тезаурус Роже.

Тезаурус Баранова: 7 уровней иерархии понятий, 7500 статей, группирующих 100 000 лексем [Баранов, 2002]

Семантические сети

Шведский ассоциативный тезаурус (Swedish Associative Thesaurus)

Охватывает полный словник шведского языка [Lönngren, 1998]. Группой составителей сформирована иерархическая структура (от частного к общему), выражающая семантические отношения между понятиями окружающей действительности.

Группа проектов на основе технологии WordNet: WordNet, EuroWordNet, VisualThesaurus

В основе словаря лежит понятие синсета (synset) – группы общих по смыслу понятий. Вокруг них обозначены гипонимы, гиперонимы, меронимы (имеет части…) и т.д. Представлено естественно-языковое описание и примеры употребления. Синсеты образуют семантическую сеть.

WordNet – семантическая сеть английского языка [WordNet].

EuroWordNet – семантическая сеть испанского, итальянского, датского и английского языков [EuroWordNet].

Lexical FreeNet www.lexfn.com

Виды связей:

1. Присутствующие в WordNet 1.6
2. Рифмы, созвучные слова
3. Библиографические данные персоналий проекта S9.com
4. Анаграммы
5. Связи, обусловленные высокой частотой совместной встречаемости в естественно-языковых текстах. Алгоритм извлечения связей разработан Адамом Бергером [Бергер, 1997].

4. Лемматизация АВС с устранением частичной омонимии.

Лемматизация сети.

Так как на ответы респондентов не накладывалось ограничений, одно и то же понятие может быть представлено в ассоциативной сети несколькими узлами. Лемматизация, наряду с усилением связей между узлами, может скрывать семантические отличия между ассоциациями. Информационная технология позволять работать как с лемматизированной, так и с исходной сетью.

Лемматизация проведена с помощью орфографического словаря iSpell, используемого для проверки орфографии в среде Unix. Он включает в себя материалы множества изданий.

Суммарное число лемм: 127 000, словоформ: 1 300 000.

Результаты лемматизации можно видеть в таблице

Число	Исходная сеть	Лемматизированная сеть
Узлов	103 000	63 700
Связей	457 000	394 000
Стимулов	6665	3833

Число узлов сократилось на 40%, связей на 16%.

Частичная омонимия.

Под частичной омонимией подразумевается совпадение отдельных словоформ у разных по написанию и смыслу лемм. Примерами частичных омонимов могут быть: чеки (чек, чека), белок (белок, белка), шерсти (шерсть, шерстить). Анализ ассоциативной сети показал, что она содержит 1030 частичных омонимов, что делает неопределенными 25 880 связей.

Обработка частичных омонимов была проведена с помощью программных средств. Связи ассоциативной сети, распределялись среди «конкурирующих» лемм экспертным мнением. Выбор эксперта сохранялся в виде текстовых файлов вида:

```
СЛОВОФОРМА

# ЛЕММА1->РЕЛЯТОР_ЛЕММЫ1

СЛОВОФОРМА_СТИМУЛ->СЛОВОФОРМА_РЕАКЦИЯ->ЧИСЛО_СВЯЗЕЙ

...

# ЛЕММА2->РЕЛЯТОР_ЛЕММЫ2

СЛОВОФОРМА_СТИМУЛ->СЛОВОФОРМА_РЕАКЦИЯ->ЧИСЛО_СВЯЗЕЙ

...
```

Релятор – символ или слово, используемое для различения значений многозначного слова (ГОСТ 7.74-96 Информационно-поисковые языки). Релятор введен как атрибут леммы для различения полных омонимов. Таким образом, информационная технология позволяет различать полные омонимы, но, в силу трудоемкости разделения (десятки тысяч полных омонимов в русском языке), производить его в рамках диссертационной работы не планируется.

```
шерсти

# шерстить->

# шерсть->

комок->шерсти->1

клубок->шерсти->7
```

Поиск путей

Марковский процесс.

При проведении ассоциативных экспериментов было замечено, что последовательность предъявления стимулов имеет значение. То есть предыдущая пара «стимул-реакция» оказывает влияние на реакцию предъявляемого следом стимула. Это основано на инерции сознания и механизмах работы памяти. В теории активации распространения Коллинза и Лофтус этот эффект обозначен эффектом предварительной подготовки [Солсо, 2006, гл.8]. Поэтому в ассоциативных экспериментах стараются сделать взаимовлияние ассоциативных пар менее значимым (не систематическим), предъявляя материал испытуемым в разном порядке, с некоторыми временными интервалами между вопросами. При в масштабном ассоциативном эксперименте ассоциации можно считать независимыми.

Обозначим как S – систему состояний ассоциативного эксперимента, состоящую из A_1, A_2, …A_n несовместимых состояний, являющихся узлами ассоциативной сети. Переход осуществляется в дискретные моменты времени и определяется по стохастическому закону.

Тогда смена состояний системы S является простой однородной цепью Маркова с конечным числом состояний и дискретным временем [Романовский, 1949, с. 9-11].

Вероятность перехода из состояния A_i в состояние A_j за один шаг дискретного времени P_{ij} зависит от числа зафиксированных ассоциаций A_i->A_j, а также от числа предъявления стимула A_i.

P_{ij} формируют стохастическую матрицу переходов M_{p1}, где 1 в индексе означает переход за один интервал дискретного времени.

$$M_{p1} = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n0} & p_{n1} & \cdots & p_{nn} \end{pmatrix} \quad (2)$$

Уравнение Колмогорова-Чепмена применительно к матрице переходов принимает вид:

$M_{pn} = (M_{p1})^n$ (3) [Портенко, 1989],

таким образом определение расстояний между узлами сети может проводиться с помощью возведение в степень матрицы переходов M_{p1}.

Вероятность перехода по пути соответствует произведению вероятностей перехода входящих в путь ассоциаций. Данный метод расчета путей используем для определения количественной характеристики пути. Поиск путей будет осуществляться с позиций равной значимости ассоциаций, таким образом, поиск выполняется с фиксированными входными данными, за определенное число шагов возвращая результат.

Двунаправленный поиск.

Ассоциативно-вербальная сеть содержит циклы. Эмпирически определено, что в случае, если два узла сети достижимы, они достижимы за число шагов, не превышающее число 7. Несомненно,

при заполнении АВС могут возникнуть ассоциативные цепочки, нарушающие данное правило, но важно его выполнение в подавляющем большинстве запросов, так как с точки зрения практических задач больший интерес представляют пути менее 7.

Была предложена модификация волнового алгоритма поиска, состоящая в поиске как от стимула к реакции, так и в обратном направлении, одновременно. Алгоритм балансируется в процессе выполнения согласно соотношению узлов в множествах волн в прямом и обратном направлениях.

Приведем последовательность поиска при поиске путей от стимулов «оружие», «лук» к реакции «арбалет» путей длиной до 6, для сравнения с первоначальным алгоритмом (**Ошибка! Источник ссылки не найден.**).

Модификация алгоритма позволила определить пути, совершив обход 2,2% числа связей, рассмотренных базовым волновых алгоритмом.

Кластеризация

Основой любого метода кластеризации является определение расстояния между узлами. Расстояния, разумеется, умозрительного, в качестве меры похожести, общности объектов. В работе «Ассоциации информационных технологий» [Филиппович, Черкасова, Дельфт, 2002] в числе прочих задач, проводилась кластеризация компактной ассоциативной сети, ограниченной областью информационных технологий. Авторами была предложена мера близости стимулов через общность множеств их реакций.

В АВС ассоциативного эксперимента мы можем расширить метрику, учитывая также непосредственные связи между сравниваемыми узлами, одновременно сделав эту меру общей для всех узлов сети, не только стимульных.

К метрике расстояний предъявляются требования:

1. Расстояние между элементами должно зависеть от списка общих реакций элементов и соответствующих вероятностей ассоциативной сети.
2. Расстояние между элементами должно зависеть от связей между ними в ассоциативной сети.
3. При отсутствии общих реакций и связей, расстояние должно равняться некоторому Lmax, при увеличении связности элементов стремиться к Lmin. Промежуточные значения должны располагаться от Lmin до Lmax.

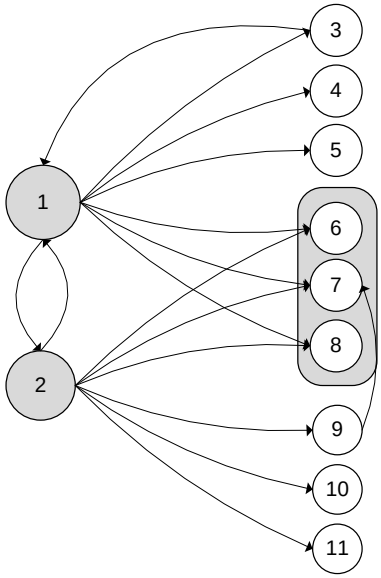

Рис. 1 Определение расстояния для кластеризации

Рис. 1 иллюстрирует определение расстояния между узлами 1 и 2. В общем случае узлы имеют совпадающие реакции (узлы 6,7,8), и непосредственные связи между собой.

Обозначим P_{ij} величину, соответствующую вероятности перехода в ассоциативной сети от узла i к узлу j по соответствующей связи (параметр Prob кортежа свойств ассоциации).

Пусть L_{ij} – искомое расстояние между узлами i и j. Тогда связность узлов i и j

$$F_{ij} = P_{ij} + P_{ji} + \sum_{k=1}^{N} \mathrm{Min}(P_{ik}, P_{jk}) \qquad (1)$$

где N соответствует числу узлов сети.

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

Коэффициент 2 в числителе формулы (2) установлен из соображений, что F_{ij} принимает значения от 0 до 2;

Сумма в формуле (1) на практике не производит просмотр всех узлов сети, поскольку для большинства узлов k связей P_{ik}, P_{jk} не существует. Для слабосвязных графов большой размерности эффективнее хранить связи в виде списков смежных узлов [Сэджвик, 2002].

Применительно к Рис. 1:

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

Методика построения словарных статей ЛКТ.

Словарная статья должна состоять из фигуры знания и путей ассоциативной сети, ведущих от стимульных узлов формулы смысла к реактивному узлу знака.

Процесс создания статьи включает этапы:

1. Определение множества стимульных узлов, соответствующих формуле смысла. Определение реактивного узла, соответствующего знаку;
2. Вычисление путей в ассоциативной сети. Предварительная фильтрация результатов;
3. Формирование документа установленного образца;
4. Ручной контроль результатов.

В этапах 1 и 4 велико участие человека, остальные операции должны быть автоматизированы (выполнение автоматически согласно настройкам пользователя).

Формирование словарных статей должно производиться с помощью xml-файлов, созданных на предыдущем этапе.

Перед созданием документа производится выбор необходимых параметров: метод сортировки/группировки словарных статей, максимальное число путей для отображения.

Затем производится автоматическое создание документа в формате html установленного образца, который должен корректно отображаться программой Microsoft Word 2003.

Реализация

Методы и алгоритмы обработки психолингвистических экспериментов реализуются программным комплексом «Лингвокультурный тезаурус русского языка». Будучи предназначенным для научных целей, комплекс должен быть доступен исследователям без необходимости покупки лицензий, желательно использование компонент, находящихся в свободном доступе.

К компонентам и средствам разработки предъявляются требования:

1) Доступность интегрированной среды разработки, справочной информации, распространенность технологии;
2) Реляционная СУБД, с поддержкой хранимых процедур;
3) Наличие справочной информации.
4) Отсутствие необходимости покупать лицензии конечным потребителем.

Желательными качествами являются:

1) Большое число доступных компонент, совместимых с программной технологией (визуализации, работы с XML, сетевым доступом).
2) Доступность на рынке хостинга с поддерживаемой СУБД.

В результате были выбраны и используются:

Среда разработки: Java SE Development Kit (JDK) 6;

Средство проектирования: Sun Microsystems NetBeans;

СУБД: PostgreSQL

Заключение.

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.