



RCDL 2008

Автоматическое построение терминологической базы знаний

ОФИМ СО РАН
Чанышев О.Г.

fedorov22@yandex.ru



ОСНОВНЫЕ ЦЕЛИ

ИССЛЕДОВАТЕЛЬСКАЯ:

создание базы для исследований в области обработки естественно-языковых запросов на терминологической сети.

ПРАГМАТИЧЕСКАЯ:

раскрытие семантики сочетаний путем представления пользователю множества содержащих их предложений.



ОСНОВНЫЕ ПРОБЛЕМЫ

Критерий адекватности сочетаний предметной области?

Критерий группирования сочетаний в предметном указателе терминологической ИПС?

Мера ассоциативной близости сочетаний, которая может быть использована для поиска информации в терминологической сети?



ВЫДЕЛЕНИЕ ТЕРМИНОПОБНЫХ СЛОВСОЧЕТАНИЙ - 1

RCDI 2008

Известные условия, налагаемые на сочетания:

Устойчивость (повторение в тексте минимум дважды)

Контактность

Объектность (обязательное наличие существительного)

Семантическая завершенность

Наше дополнение (обеспечивающее адекватность предметной области):

ДОМИНАНТНОСТЬ



УСЛОВИЕ ДОМИНАНТНОСТИ

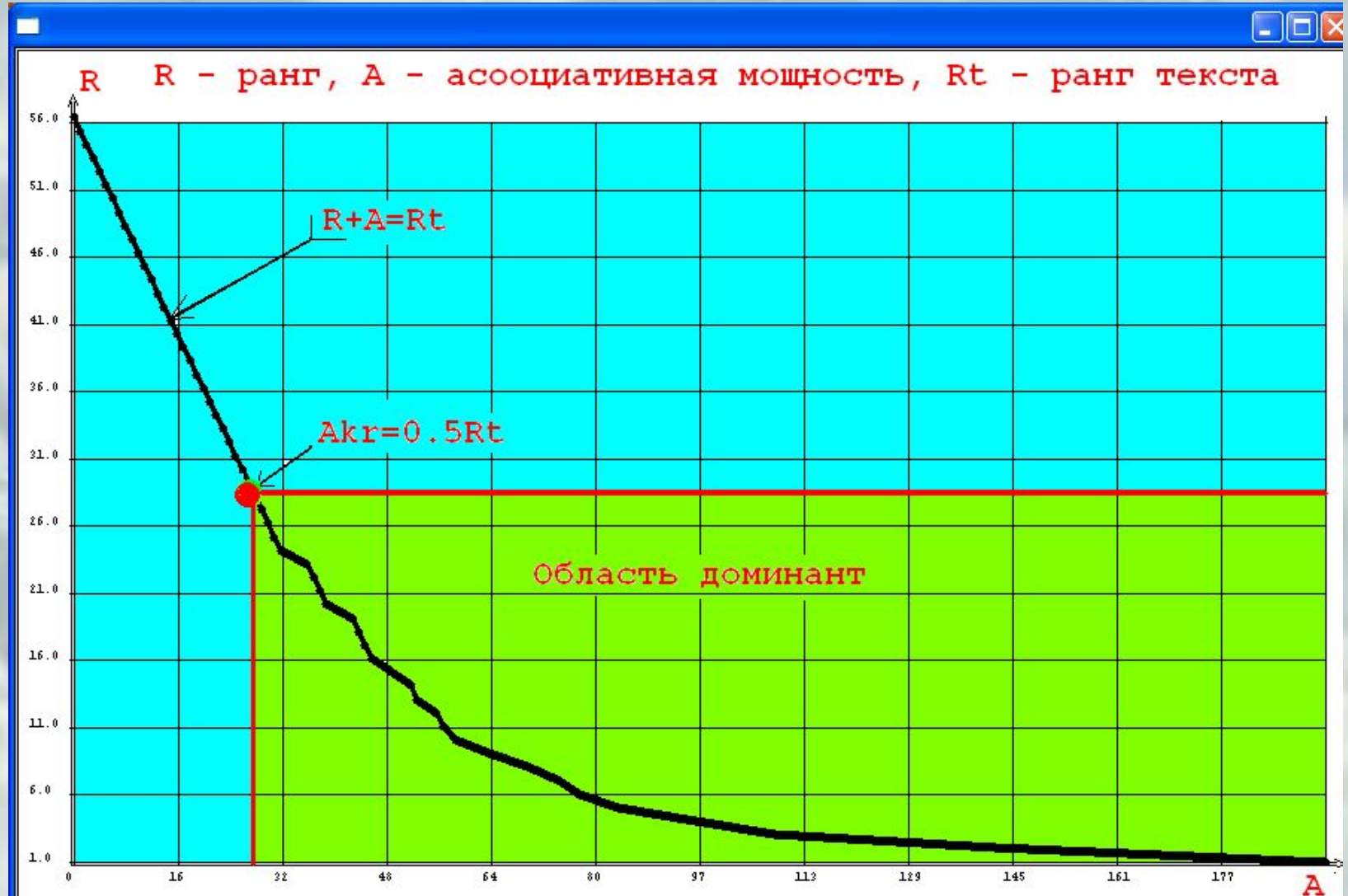
Терминоподобные
словосочетания должны
содержать слова, являющиеся
доминантами
хотя бы в одном из
анализируемых текстов



ВЫДЕЛЕНИЕ ТЕРМИНОПОБНЫХ СЛОВСОЧЕТАНИЙ - 2

Отбор доминант

RCDL 2008



ВЕСА ДОМИНАНТ И СЛОВСОЧЕТАНИЙ

Вес доминанты в фиксированном тексте равен ее обратному рангу в убывающей по значению ассоциативной мощности последовательности доминант.

Вес нормы доминанты во множестве файлов равен сумме весов ее доминантных грамматических форм.

Вес словосочетания равен сумме весов входящих доминант.

Вес нормы словосочетания равен сумме весов элементов его парадигмы.



ВЫДЕЛЕНИЕ ТЕРМИНОПОДОБНЫХ СЛОВСОЧЕТАНИЙ - 5

РСДИ 2008

Вход программы выделения терминоподобных словосочетаний
список полных имен файлов, содержащих тексты из фиксированной предметной области;

файлы с текстами.

Выход

Множество фактов (в синтаксисе Пролога), представляющие:

дерево вхождений отфильтрованных словосочетаний в тексты и предложения текстов,

предметный указатель.

Файлы с текстами, в которых отмечены начала предложений



Файлы и факты

Наименования словарей	
Наименование словаря	Имя файла со словарем

1 ○

Файл со словарем

Файл, содержащий факты предметного указателя

Группа		
Кардинальное слово	Норма сочетания	
2 ○		
Ссылка_на_группу		
Кардинальное слово 1	Кардинальное слово 2	Норма сочетания

имена_баз	
наименование ПО	имя файла фактов

1 ○

Файл фактов, представляющий входения сочетаний

Норма_сочетания			
Номер нормы	Норма сочетания		
2 ○			
Элемент_парадигмы			
Номер нормы	Элемент парадигмы		
Ссылка_на_текст_вхождения			
Номер нормы	Имя файла	Наимен. текста	Список предложений
3 ○			

Файл с текстом

Номер предложения Предложение ...Номер предложения Предложение ...



Предметный указатель -1

RCDL 2008

Главные (кардинальные) слова терминоподобных словосочетаний.

Для организации **предметного указателя**

в каждом словосочетании выделяется доминанта с наибольшим весом – **кардинальное слово**.

Словосочетания группируются по признаку общего кардинального слова.

В группах могут выделяться подгруппы с общими повторяющимися сочетаниями слов с кардинальным.



Пример групп и подгрупп

система

система искусственный интеллект

совершенствование

система искусственный интеллект

современный

система искусственный интеллект

система ии

современный система ии

построение система ии

история развитие система ии



Предметный указатель -3

RCDI 2008

Ссылки на включения

В результате группирования часть кардинальных слов, выбираемых последовательно из их множества, частично упорядоченного по убыванию веса, может остаться без своих включающих словосочетаний.

В таком случае для них организуются ссылки на соответствующие группы.

Пример:

понимание->система->система понимание естественный язык



Контекстная мера ассоциативной близости

$$A(K_i, K_j) = aN / (1 + L \times L_{min}),$$

где

K_i, K_j – группы сочетаний, идентифицированные i -ым и j -ым кардинальными словами ,

N – число общих текстов (в которые входят хотя бы по одному элементу парадигмы из различных групп),

L, L_{min} – среднее и минимальное расстояния между предложениями, включающими элементы парадигм различных групп,

a – нормировочный коэффициент



ЭКСПЕРИМЕНТ.

Группы анализируемых текстов

1. Философия (12 текстов, 33 файла),
2. Психология (19 текстов, 19 файлов)
3. СУБД (13 файлов).
4. Искусственный интеллект (13 текстов, 18 файлов)
5. Политология (3 текста, 32 файла).
6. Монография Н.А. Олифер, В.Г. Олифер
"Сетевые операционные системы" (10 файлов).
7. Карамзин "История государства Российского" (12 файлов)
8. Бунин (52 файла),
9. Чехов (11 файлов),
10. Борис Акунин (5 романов, 57 файлов).

RCDL 2008





RCDL 2008

ЭКСПЕРИМЕНТ.

Контроль адекватности

Эталонные множества словосочетаний (нормированные наименования статей):

а) «Новейший философский словарь под редакцией Грицанова А.А.», 1390 наименований, («Философия-эталон»);

б) «Психологический словарь», 2172 наименования, («Психология-эталон»).

в) «Словарь компьютерной лексики», 1213 наименований, («КомпЛекс-эталон»).

Контрольные множества словосочетаний: «СУБД», «СетОпСист», «Иск. Инт.», «Философия», «Психология»

Для контроля качества подборок был проанализирован Краткий справочник «Психологические теории и концепции личности..») и нормированные двухсловные словосочетания включили в контрольную подборку («ПсихТеор»).



ЭКСПЕРИМЕНТ. Контроль адекватности

RCDL 2008

	Комп-Лекс эталон	Философия- эталон	Психология- эталон
СУБД	359.05 39		
Сет. Оп. Сист.	82.05 56		1.79 1
Иск. Интелл.	49.61 10		5.73 4
Философия	0.18 1	184.09 28	7.16 6
Психология	2.27 10	2.47 2	47.83 44
ПсихТеор	2.35 3		28.57 13



ЭКСПЕРИМЕНТ. Пример. Первые 10 словосочетаний. «Сетевые операционные системы»

Упорядоченность: а) по убыванию веса, б) по убыванию числа повторений в различных текстах, б.2) по литературным данным

а) сетевая ос, операционная система, сервер netware, база данных, файловая система, менеджер памяти, сетевая операционная система, функции операционной системы, сервер сети, драйвер файловой системы;

б) операционная система, программное обеспечение, файловая система, рабочая станция, структура данных, получение доступа, передача сообщений, виртуальная память, оперативная память, реальное время;

б.2) операционная система, файловая система, адресное пространство, ввод-вывод, оперативная память, рабочая станция, системный вызов, база данных, право доступа, программное обеспечение.

RCDI 2008



ЭКСПЕРИМЕНТ. Пример. Первые 10 словосочетаний. «СУБД»

Упорядоченность: а) по убыванию веса, б) по убыванию числа повторений в различных текстах

а) база данных, распределенная база данных, страница данных, сервер базы данных, объект базы данных, состояние базы данных, локальная база данных, модель данных, система баз данных, тип данных;

б) база данных, ограничение целостности, внешняя память, язык sql, реляционная СУБД, прикладная программа, оперативная память, кортеж отношения, информационная система, управление базами данных;

RCDL 2008



ЭКСПЕРИМЕНТ.

Первые тройки правил (по частоте использования) лексико-морфологического фильтра

Компьютерная лингвистика

- 21 Последнее слово не существительное и не прилагательное
- 9 Первое слово начинается не с кириллицы и второе слово не в именительном падеже
- 8 Нет существительного в составе

Искусственный интеллект

- 38 Первое слово - элемент парадигмы "какой-либо"
- 32 Последнее слово не существительное и не прилагательное
- 23 Первое слово "система"|"system", второе - латинская буква

СУБД

- 46 Последнее слово не существительное и не прилагательное
- 30 Первое слово - элемент парадигмы "какой-либо"
- 20 Первое слово начинается не с кириллицы и второе слово не в именительном падеже

Философия

- 90 Последнее слово не существительное и не прилагательное
- 37 Нет существительного в составе
- 32 Первое слово есть глагол в несовершенной форме

Психология

- 55 Последнее слово не существительное и не прилагательное
- 40 Нет существительного в составе
- 26 Первое слово - элемент парадигмы "какой-либо"



ИПС. Меню выбора сочетания из группы

RCDL 2008

The screenshot displays the DBS_Termin2 application window. The title bar reads "DBS_Termin2". The menu bar includes "ВЫБОР ПРЕДМЕТНОЙ ОБЛАСТИ", "ПАРАДИГМА ТЕРМИНА", "ВСЕ ТЕКСТЫ ПО", "ВЫБОР ТЕРМИНА", "ВЫБОР СЛОВАРЯ", "ВХОЖДЕНИЯ", "АССОЦИАЦИИ и БЛИЗОСТИ", "Window", and "Help". A toolbar with various icons is located below the menu bar. A dialog box titled "Выберите кардинальное слово" is open, showing a list of words: "память", "отношение", "система", "СУБД", "транзакция", "бд", "выполнение", "операция", and "память". The "память" option is selected. Below the list are "OK" and "Cancel" buttons. In the background, a "Messages" window is visible with the text "S*** Активная ПО >> СУБД". The taskbar at the bottom shows the Start button, several icons, and open applications: "Microsoft PowerPoint - [...]" and "DBS_Termin2". The system tray on the right shows the date and time as "11:23".

ИПС. Предложения вхождения

RCDL 2008

The screenshot displays the DBS_Termin2 application window. The title bar reads "DBS_Termin2". The menu bar includes "ВЫБОР ПРЕДМЕТНОЙ ОБЛАСТИ", "ПАРАДИГМА ТЕРМИНА", "ВСЕ ТЕКСТЫ ПО ВЫБОР ТЕРМИНА", "ВЫБОР СЛОВАРЯ", "ВХОЖДЕНИЯ", "АССОЦИИИ И БЛИЗОСТИ", "Window", and "Help". The toolbar contains icons for file operations and help.

Two "Окно_Редактора" (Editor Window) windows are open:

- The top-left editor window shows text: "ТЕКСТ >> С. Д. КУЗНЕЦОВ. ВВЕДЕНИЕ В СУБД: ЧАСТЬ 2." followed by two paragraphs of text starting with "<14>" and "<15>".
- The bottom-right editor window shows text: "ТЕКСТ >> Г.М.ЛАДЫЖЕНСКИЙ. РАЗДЕЛ 4. ОБРАБОТКА ТРАНЗАКЦИЙ." followed by several paragraphs of text starting with "<59>", "<60>", "<65>", "<98>", "<101>", and "<108>".

A "Messages" window is open at the bottom, displaying two messages:

- "\$*** Выбран термин: | организация внешней памяти | из группы: память"
- "\$*** Выбран термин: | процесс выполнения транзакция | из группы: транзакция"

The Windows taskbar at the bottom shows the Start button, several application icons, and the system tray with the date and time "12:06".

RCDL 2008

ИПС. Результаты поиска ассоциаций с кардинальными словами «система» и «данный»

The screenshot shows a Windows XP desktop environment. At the top, a menu bar for 'DBS_Termin2' includes options like 'ВЫБОР ПРЕДМЕТНОЙ ОБЛАСТИ', 'ПАРАДИГМА ТЕРМИНА', 'ВСЕ ТЕКСТЫ ПО', 'ВЫБОР ТЕРМИНА', 'ВЫБОР СЛОВАРЯ', 'ВХОЖДЕНИЯ', 'АССОЦИАЦИИ и БЛИЗОСТИ', 'Window', and 'Help'. Below the menu is a toolbar with various icons. The main area contains three windows:

- Окно_Редактора (Left):** Displays search results for the term 'система'. It lists various text snippets and provides summary statistics:
 - ИТОГ: АССОЦИАЦИЯ система-транзакция
 - ИТОГ: Среднее расстояние по текстам = 22
 - ИТОГ: Число общих текстов = 8
 - ИТОГ: Минимальное среднее расстояние = 2
 - Критерий близости = 1.77777778
- Окно_Редактора (Right):** Displays search results for the term 'данный'. It lists various text snippets and provides summary statistics:
 - ИТОГ: АССОЦИАЦИЯ данный-система
 - ИТОГ: Среднее расстояние по текстам = 4
 - ИТОГ: Число общих текстов = 12
 - ИТОГ: Минимальное среднее расстояние = 1
 - Критерий близости = 24
- Messages (Bottom):** Shows a notification: '\$*** Активная ПО >> СУБД' and '\$*** Выбран термин: | число блоковый внешний память | из группы: память'.

The taskbar at the bottom shows the Start button, several application icons, and the system tray with the date and time '11:55'.

ЭКСПЕРИМЕНТ. Кардинальное слово «Память». Ассоциации с другими кардинальными словами

СУБД	Сетевые операционные системы	Психология
Журнал	Адрес	АСФС
Функция	Страница	Психика
Файл	Сетевой	Семантический
Управление	Использование	Уровень
Число	Область	Информация
Страница	Пространство	Мозг
Объект	Управление	Состояние
Организация	Таблица	Расстройство
Значение	Сервер	Исследование
Кортеж	Сообщение	Реакция



ЗАКЛЮЧЕНИЕ

RCDL 2008

Представленный метод выделения терминоподобных словосочетаний, основанный на предварительном определении доминант, как наиболее тематически значимых слов текста, гарантирует адекватность выделенных словосочетаний предметным областям и пригоден для автоматической генерации терминологических баз знаний.

Предложенная мера ассоциативной близости кардинальных слов может быть использована при интерпретации запросов, как запросов на поиск наиболее нагруженных путей между предложениями, включающими выделенные из запросов кардинальные слова.

RCDL 2008

Благодарю
за внимание!



Олег Чаньышев

