Построение семантической сети из разнородных данных

Аспирант:

Александр Панченко

Научные руководители:

к.т.н. Юрий Николаевич Филиппович, МГТУ им.Н.Э.Баумана

Dr.Cédrick Fairon, Catholic University of Louvain

22 Сентября 2010, МГТУ им.Н.Э.Баумана, кафедра ИУ5

Содержание

- Основные понятия
- Цели, актуальность и план работы

Мотивация

- Характеристики семантической сети предметной области
- Критерии оценки качества семантической сети

Критерии оценки

- Создание семантической сети из коллекции текстовых документов
- Создание семантической сети из других источников информации
- Разработка обобщенного метода создания семантической сети

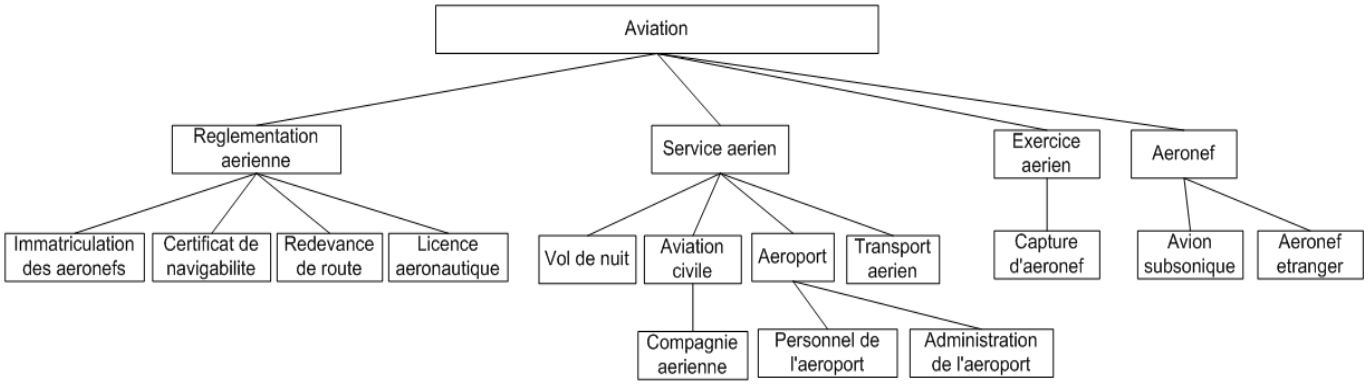
Методы

Основные понятия

- **Семантическая сеть** — модель **представления знаний** имеющая вид ориентированного графа, вершины которого соответствуют объектам, а дуги задают отношения между ними

- В данном исследовании мы работаем с (лексической) сематнической сетью представляющей **знания о какой-либо предметной области** (ПО)

- **Тезаурус** — разновидность семантической сети. Вершины – ключевые понятия ПО, дуги – семантические отношения между ними (синонимы, гипонимы, гиперонимы, ассоциативные связи)



Иерархия ключевых понятий тезауруса ПО “Авиация”

energy industry
- **NT1 energy conversion**
  - RT soft energy (6626)
- **NT1 energy technology**
  - RT bioconversion (6411)
  - RT energy policy
  - RT oil technology (6616)
  - RT soft energy (6626)
- **NT2 fuel cell**
- **NT1 energy-generating product**
- **NT1 fuel**
  - RT energy resources (5211)
- **NT2 fossil fuel**
  - RT coal (6611)
  - RT natural gas (6616)

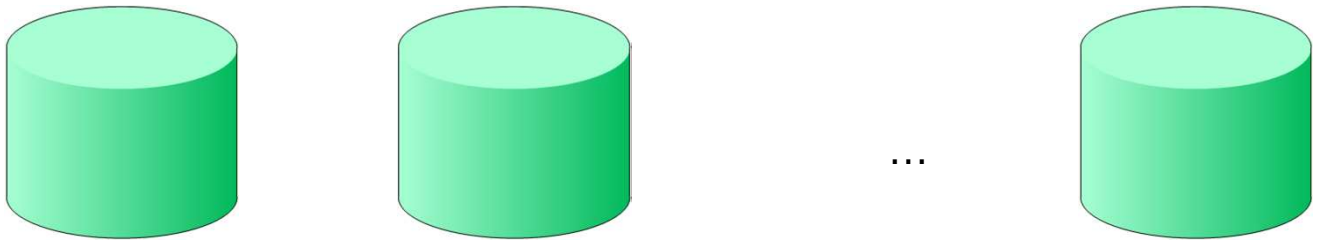
Концепт из тезауруса Eurovoc

Основные понятия

- **Разнородные источники информации** – данные в которых явно или неявно содержится информация (знания), необходимые для построения сематнической сети ПО

- **Разнородные источники информации =**

  Web: контент (текст), структура (ссылки), использование (логи, клики) **+** электронные словари**,** базы знаний, морфологическая информация **+** другие источники…

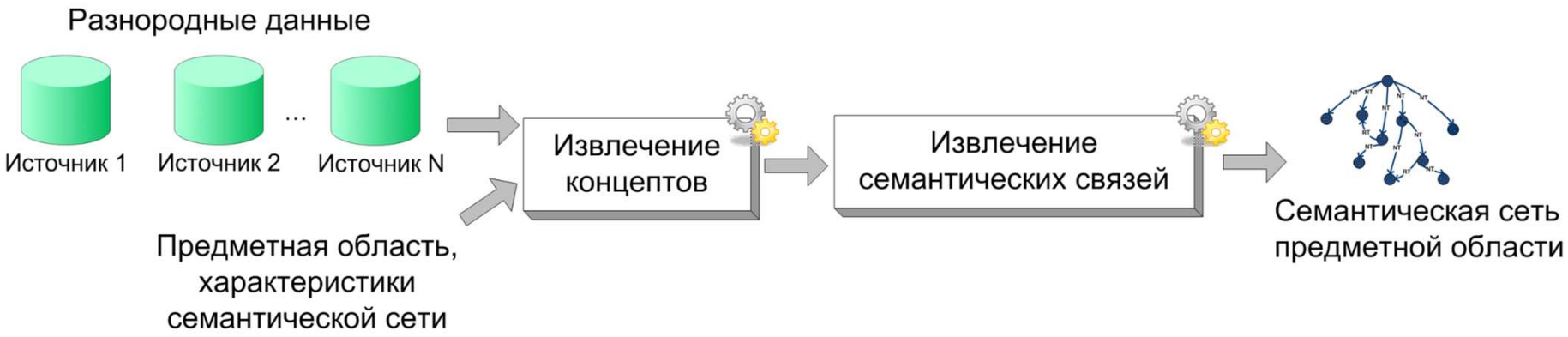


Цели работы

- **Разработка технологии извлечения семантических знаний :**
  - о заданной предметной области, с заданными характеристиками
  - из разнородных источников информации
  - для построения сематнической сети предметной области

- Разработка **критериев качества** построенной семантической сети и ее соответствия заданным характеристикам

- Реализация **программного средства** на основе предлагаемой технологии



Актуальность работы

Технологию построения семантической сети ПО можно применить для:

- **Автоматизации** создания семантических ресурсов, таких как информационно-поисковый тезаурус (рис.1)
- **Визуализации** и улучшенной **навигации** по коллекции документов (рис.2)*
- **Пополнения** существующей семантической сети, тезауруса или онтологии
- **Улучшения** производительности приложений требующих **знания о ПО**

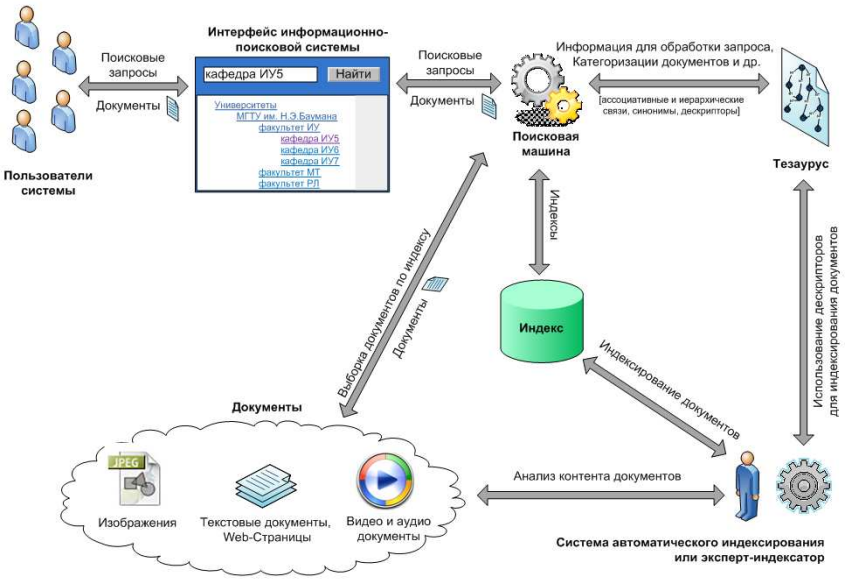


Рис.1 Использование тезаруса в инфорационно-поисковой системе

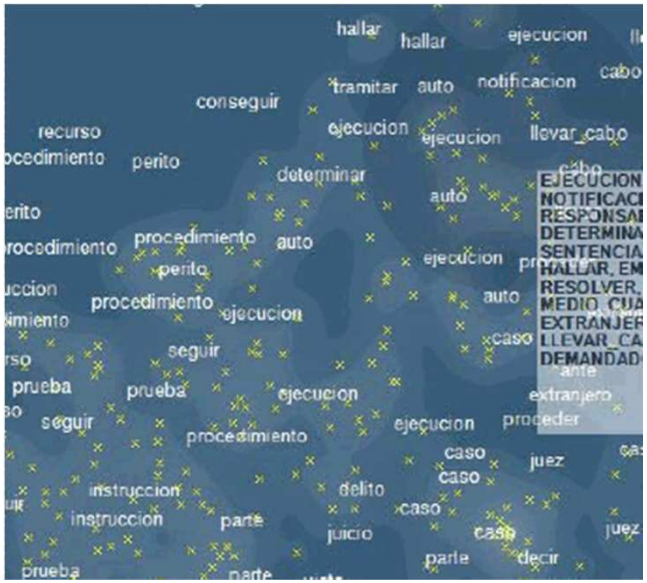


Рис.2 Визуализация текстов

*Иллюстрация из B. Fortuna, M. Grobelnik and D.Mladenić, Visualization of Text Document Corpus , Informatica 29 (2005) 497–502

План работы

1. **Изучение** существующих работ: научные статьи, книги и главы из книг, диссертации, созданные системы, активные научные группы, активные и завершенные научные проекты, стандартные наборы данных и др.

2. **Экспериментирование** с существующими методами:
   - Создание семантической сети из коллекции текстовых документов
   - Создание семантической сети из электронных словарей
   - …
   - Создание семантической сети из ***источника i*** с помощью **метода *j***

3. **Разработка** обобщенного метода создания семантической сети

4. **Реализация** програмного средства создания семантической сети, на основании разработанных методов

Характеристики семантической сети предметной области

1. **Предметная область**

   Мы работаем со следующими предметными областями (приоритетные направления плана Marshall): agro-food industry, transport and logistics, life sciences, aeronautics–aerospace, mechanical engineering.

2. **Уровень детализации** (размер): количество концептов и связей

3. Требуемый тип отношений между концептами: синонимы, гиперонимы и т.п.

4. **Архитектура** сети: дерево, лес, сеть, сбалансированное дерево, отсутвие ограничений и т.п.

5. Приемлимая **достоверность** извлеченных знаний: высокая, средняя, слабая

Критерии оценки качества семантической сети: ground truth

1. **Предметная область:** Требуется *образцовая семантическая сеть* $T=(C,R)$

   A) Количество концептов и связей не относящихся к предмеой области

   $$\text{Precision}_R = \frac{\left|R \cap \hat{R}\right|}{\left|\hat{R}\right|},\ \text{Precision}_C = \frac{\left|C \cap \hat{C}\right|}{\left|\hat{C}\right|},\underbrace{\text{Recall}_R = \frac{\left|R \cap \hat{R}\right|}{\left|R\right|},\ \text{Recall}_C = \frac{\left|C \cap \hat{C}\right|}{\left|C\right|}}_{\text{сложно точно оценить}}$$

   R, C могут быть заменены *расширенными версиями* $R \subset R_E, C \subset C_E$

   B) Подобие графов: изоморфизм графов, максимальный общий подграф, анализ распределиния вершин и т.п.

   $$sim((C,R),(\hat{C},\hat{R}))$$

   C) Привязка вершин графов с помощью “приближенных совпадений”

2. **Уровень детализации**: $\left(C_{\max} \geq \left|\hat{C}\right| \geq C_{\min}\right) \wedge \left(R_{\max} \geq \left|\hat{R}\right| \geq R_{\min}\right)$
3. **Типы отношений между концептами:** сравнение с образцовой сетью (см. 1).
4. **Архитектура сети**: численная мера несоответствия заданной архитектуре.
5. **Приемлимая достоверность извлеченных знаний**: чем выше требуется достоверность – тем меньше $\left|R_E\right|, \left|C_E\right|$

Критерии оценки качества семантической сети: ground truth

Baseline – производительность (качество) стандартного метода (методов) построения семантической сети, **оцененного по той же методике**

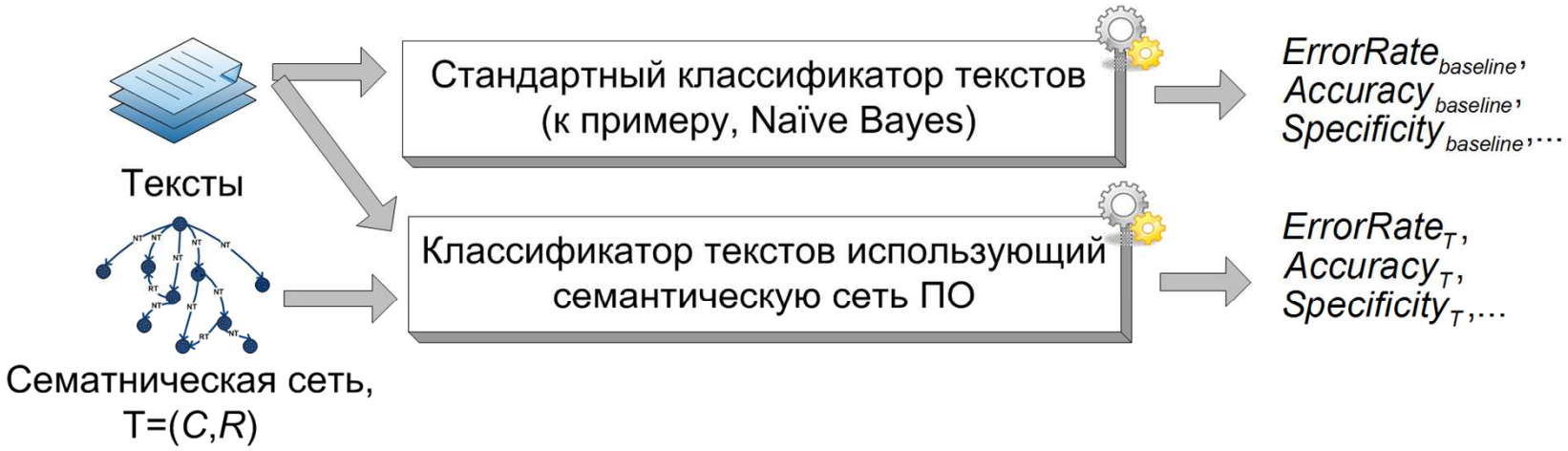
- Объективная оценка качества семантической сети
- Позволяет сравнить предлагаемый метод с множеством других
- “Дешевый и быстрый” метод, можно повротять многократно

- Адекватность оценки практически полностью определяется качеством образцовой семантической сети (сетей)
- Ресурсы содержатся в различных форматах, их совмещение – сложная задача
- Для многих ПО трудно или невозможно найти адекватный образцовый ресурс:

Предметная область	**Образцовая семантическая сеть**
–	WordNet, EuroWordNet, Cyc/OpenCyc, DOLCE, SUMO, Roget’s Thesaurus, GermaNet, …
politics	EuroVoc, Stratego, РусТез (УИС Россия), Library of Congress Thesauri (LIV)
life sciences, medicine	MeSH, SNOMED CT, UMLS
agro-food industry	AgroVoc
transport and logistics	~TOVE
aeronautics–aerospace, mechanical engineering	?
all	**Google?, Yandex? Yahoo?, Wikipedia?,…**

Критерии оценки качества семантической сети: end-to-end

- Оцениваем качество построенной семантической сети по улучшению качества работы приложения которое ее использует, к примеру:
  - Информационно-поисковой системы
  - Системы категоризации текстовых документов
  - Системы “вопрос-ответ” и т.п.



Лучшая семантическая сеть предметной области T – такая, что

$$T^* = \text{argmax}_{T_i} \left(ErrorRate_{baseline} - ErrorRate_{T_i} \right)$$

Критерии оценки качества семантической сети: end-to-end

- **Baseline** – производительность системы до использования семантической сети предметной области

- “Технический” подход: дает оценку насколько извлеченные знания полезны для конкретного приложения
- Даже если сгенерированная семантическая сеть неполна / неточна она может быть полезна для конкретного приложения

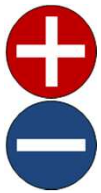
- Необохомость реализации baseline систем
- Реализация метода, который бы эффективно использовал сгенерированную семантическую сеть для конкретного приложения – отдельная сложная задача

Критерии оценки качества семантической сети: “вручную”

1. Выбираем случайные подмножества из автоматически построенной сети
2. Просим респондентов оценить численно релевантность концептов заданной предментной области и корректность отношений
3. Вычисляем коэффициент корреляции между вычисленной мерой семантической связанности и соотв.оценкой респондента → Даем оценку сети

Three best hypernyms	Hypernym 1		Any hypernym	
	majority	any	majority	any
worker/craftsmen/personnel	13	13	13	13
cost/expense/area	7	10	9	10
cost/operation/problem	6	8	11	17
legislation/measure/proposal	3	5	9	18
benefit/business/factor	2	2	2	5
factor	2	7	2	7
lawyer	14	14	14	14
firm/investor/analyst	13	13	14	14
bank/firm/station	0	0	15	17
company	6	6	6	6
AVERAGE	6.6 / 33.0%	7.8 / 39.0%	9.5 / 47.5%	12.1 / 60.5%

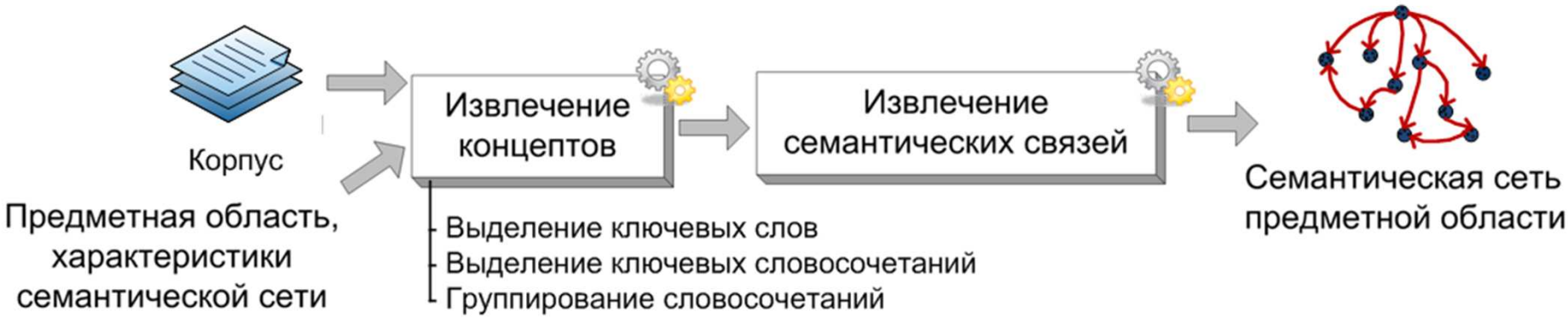
Table 2.2: The results of the judges' evaluation for the preliminary experiment.



Точный, но дорогостоящий метод. Сложно повторить. Имеет смысл применять только на конечной стадии разработки технологии.

* Иллюстрация из Sharon A.Caraballo, “Automaic Construction of a Hyperym-Labeled Noun Hierarchy from Text” , PhD thesis, Brown University, 2001, p.24.

Создание семантической сети из коллекции текстовых документов



Набор данных:

Корпус:

•Размер: 20 миллионов слов, 11.386 документов

•Язык: Французский

•Стиль: Политические тексты: Запросы депутатов к министрам и т.п.

•Предметная область: 12 подобластей, таких как, законодательство, экономика, международные отношения и др.

Образцовая семантическая сеть (тезаурус):

•Размер: 2514 концептов, 4771 дескрипторов (WordNet 3.0 > 200.000 дескрипторов)

•Отношения: иерархические (2456), ассоциативные (1530) и синонимы

Создание семантической сети из коллекции текстовых документов

Выделение ключевых слов (baseline):

1. Нормализация, токенизация, лемматизация: ~149.500 лемм $D_j = \{d_{1j}, d_{2j}, ...\}$
2. Фильтрация на основе **лингвистической информации**: удаление стоп слов, имен собственных, чисел, дат, всего остального кроме существительных и прилагательных. ~78.250 лемм (-50%)
3. Ранжирование слов-кандидатов с использованием **статистической информации**:

   А) По **частоте** $rank_i = n_i$, ключевые слова – первые *x%*

   Б) **“Глобальный” TF-IDF**, ключевые слова – первые *x%*

$$rank_i = \frac{n_i}{\sum_{j=0}^{N} n_j} \log\left(\frac{|DOC|}{|\{doc : d_i \in doc\}|}\right), \qquad D_{key} = \left\{ d_i : \frac{x}{100}|D| \le rank_i \right\}$$

   В) **“Локальный” TF-IDF**, ключевые слова – объединение первых *x%*, но не более чем для *y* слов

$$rank_{ij} = \frac{n_{ij}}{\sum_{j=0}^{N} n_{ij}} \log\left(\frac{|DOC|}{|\{doc : d_i \in doc\}|}\right), \qquad D_{key} = \bigcup_{j=1}^{|DOC|} \left\{ d_i : \min\left\{ y, \frac{x}{100}|D_j| \right\} \le rank_{ij} \right\}$$

Создание семантической сети из коллекции текстовых документов

Выделение ключевых (baseline):

Метод ранжирования	**Нет**	**По частоте**			**“Глобальный” TF-IDF**			**“Локальный” TF-IDF**		
x	100%	10%	15%	33%	10%	15%	33%	10%	15%	33%
Полнота, %	92%	62	74	87	62	73	87	24	24	24
Точность, %	2%	13	10	5	13	10	5	17	17	17
Количество ключевых слов	73.513	7.351	11.027	24.259	7.351	11.027	24.259	1.196	2.198	2.198

Как улучшить этот метод?

- Использование информации о распределении слов или модели языка (language model) в корпусе общей лексики
- Использование информации о подобии неизвестного слова известным дескрипторам для данной предметной области
- ...

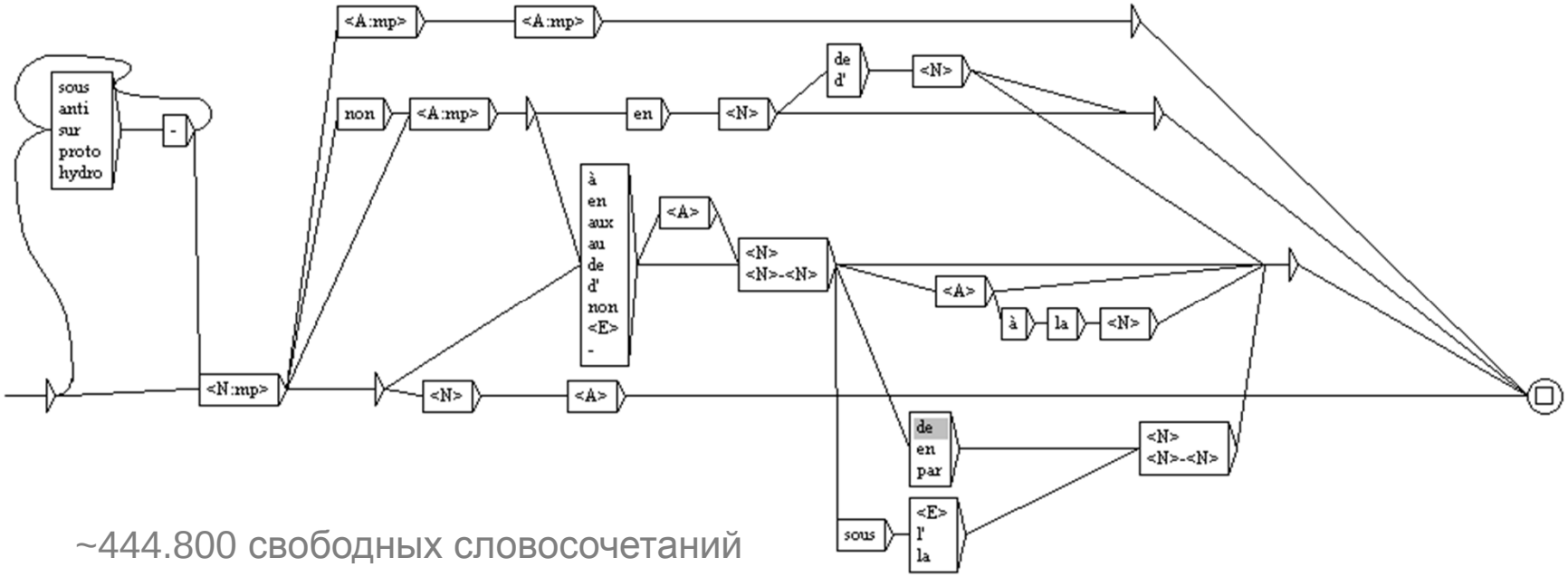
Создание семантической сети из коллекции текстовых документов

Выделение ключевых словосочетаний:

1. Извлечение **свободных словосочетаний**

   Прим.: “**Прозрачный воздух** был теплым и нежным”

   Использование **символьного метода**: набор конечных автоматов, фиксирующих лингвистический феномен и словарей (Unitex).



~444.800 свободных словосочетаний

Создание семантической сети из коллекции текстовых документов

2. Группирование **словосочетаний-кандидатов** путем поиска наибольших общих подстрок.

   Для каждого строки (словосочетания) d_i найти все строки, которые содержали бы содержали d_i : $\{d : d_i \subseteq d\}$

2. Ранжирование **словосочетаний** с помощью следующей формулы:

$$rank_i = g_i \frac{n_i}{\sum_{j=0}^{N} n_j} \log\left(\frac{|DOC|}{|\{doc : d_i \in doc\}|} \right),$$

   g_i – коэффициент группирования

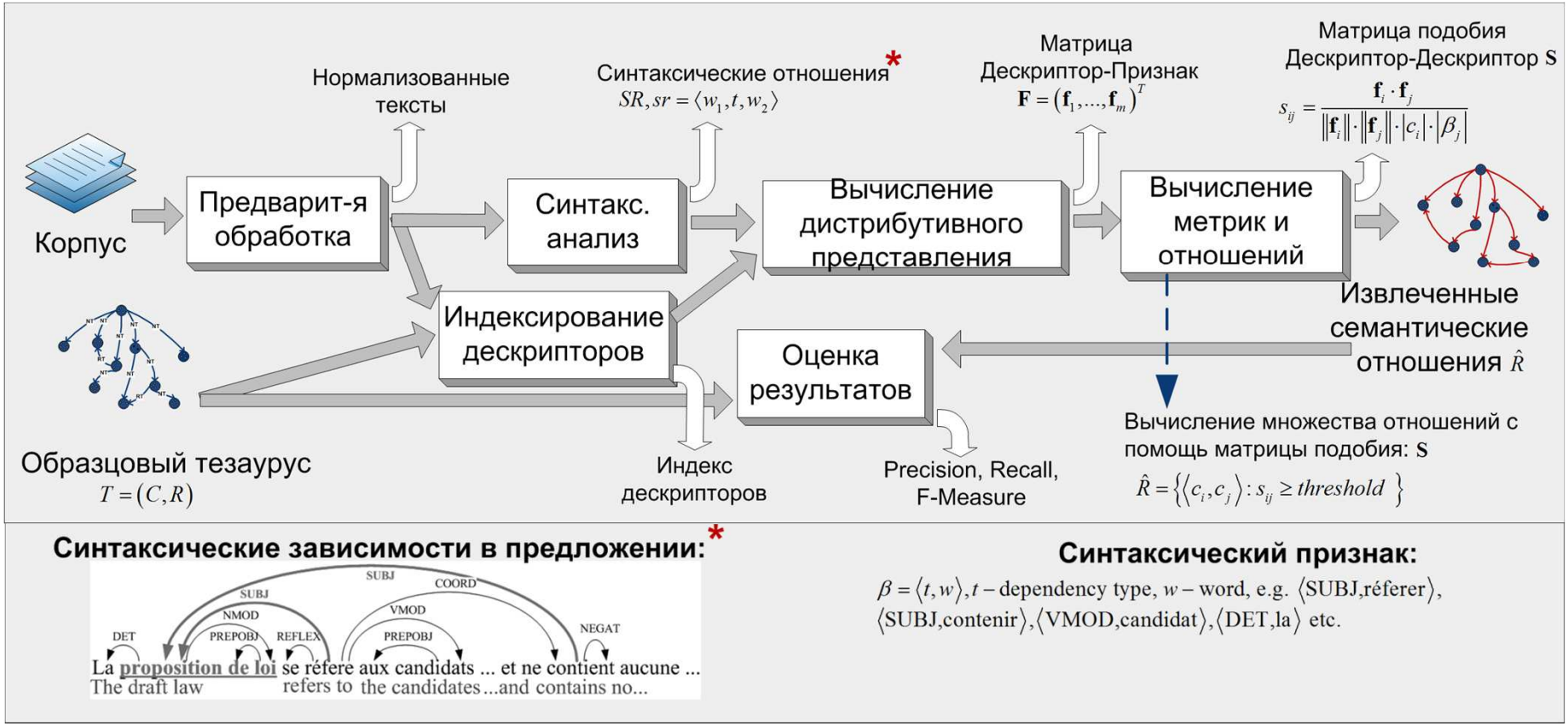
3. Ключевые словосочетания – первые x%

ANCIEN CHEF
- ancien chef
- ancien chef de la centrale
- ancien chef d'etat
- ancien chef d'exploitation,
- anciens chefs de gouvernement
- ancien chef du laboratoire
- ancien chef politique
- ancien chef du service
- ancien chef des services secrets

Создание семантической сети из коллекции текстовых документов

Извлечение семантических связей: найти множество бинарных отношений между концептами $\hat{R} = \{r_1,..,r_n\}, r_i = \langle d_1, d_2 \rangle, d_i \in D_{key}$

Дистрибутивно-статистический анализ: слова встречающиеся в *похожих* синтаксических **контекстах** семантически связаны.



Создание семантической сети из коллекции текстовых документов

Некоторые автоматически извлеченные отношения:

Дескриптор	Связанный дескриптор	
	из образцового тезауруса	из автоматически построенного тезауруса
administration of taxes	administration of the state	administration of the cadastre and the topography (2), state socio-educational center (8), public education (4), cultural institution (8), institute of hygiene and public health (7), state vineyard station (6)
admission to studies	school organization, education, admission to employment	archives of the state (9), certificate of teacher (6), program of studies (2)
medical assistance	medical organization	emergency medical services (1), medical analysis (6), medically assisted procreation (6) hygiene (6), wine institute (9), medical organization (1) medical profession (3), vaccination (5)
european election	election, political life, european parliament	legislative election (2)
education grants	school life, education	youth movement (11)

Оценка качества извлеченных отношений:

Ground truth – отношения из образцового тезауруса R + их расширенная версия R_E

Вычисление R_E:

1. Взвешиваем каждую семантическую связь из образцового тезауруса
2. Вычисляем кратчайшие пути между дескрипторами тезауруса
3. Генерируем дополнительные отношения между дескрипторами связанными кратчайшими путями

Результаты:

Exact Precision = 7%
Precision Fuzzy3 = 35%
Precision Fuzzy4 = 46%

Создание семантической сети из других источников информации

- Использование корпуса текстов как источника информации и дистрибутивного анализа – только **один из возможных подходов**

- Другие методы, **основанные на анализе текста**:
  - Лексико-синтаксические шаблоны
  - Анализ совместной встречаемости слов (коллокации)
  - Латентно-семантический анализ (LSA) и его разновидности (PLSA, LDA, Latent Class Model)
  - Анализ морфологии и формы слов

- Кроме этого, можно извлекать (находить) информацию о семантической связности концептов **из других источников** …

Создание семантической сети из других источников информации

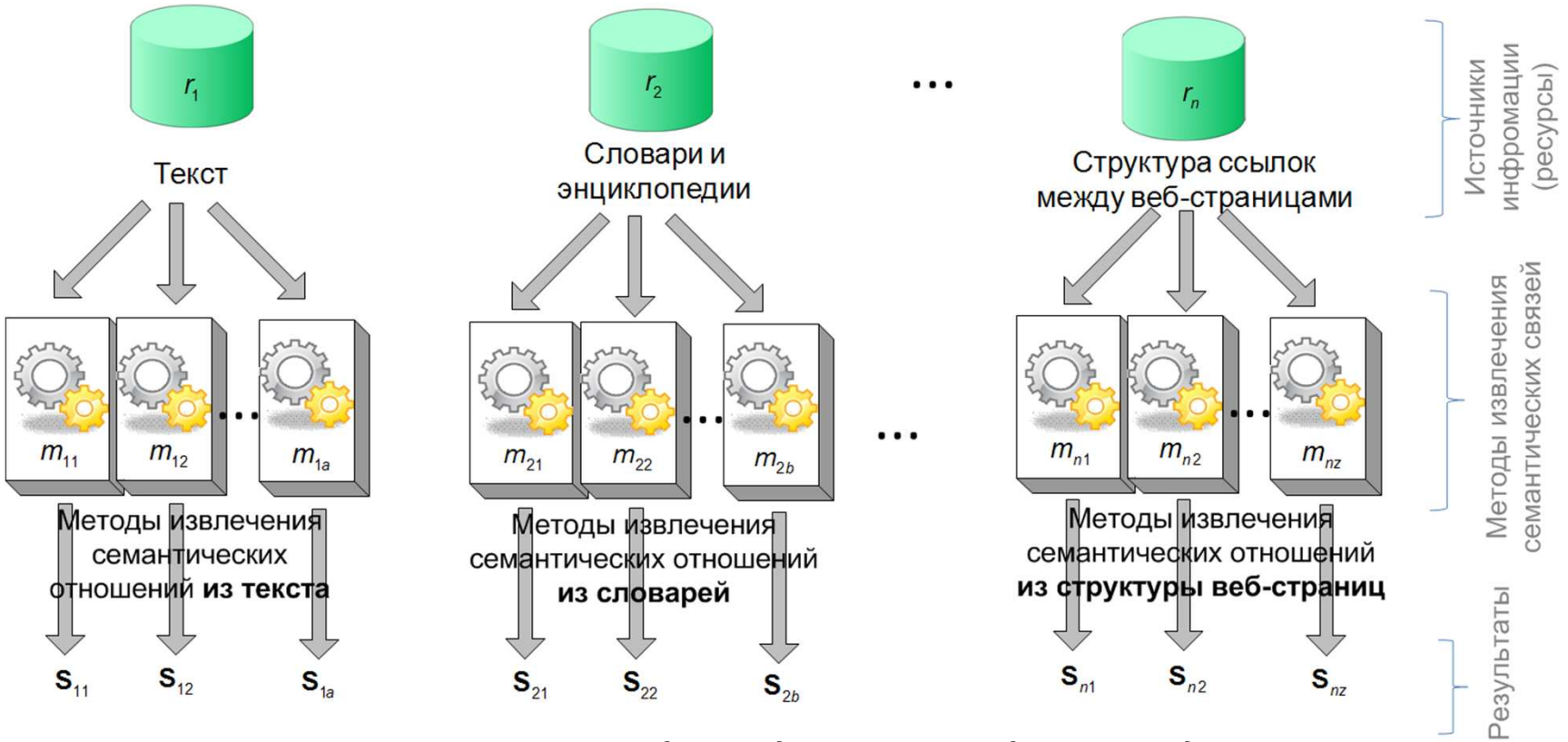
Источник информации	**Методы извлечения семантических связей:**
Текст	Дистрибутивный анализ, Лексико-синтаксические шаблоны, Анализ совместной встречаемости слов, Латентно-семантический анализ (LSA) и его разновидности (PLSA, LDA, Latent Class Model)
Словари и энциклопедии	Компонентный анализ, Extended gloss overlap, Extended Lesk, Graph Mining методы на словаре
Семантические сети, тезаурусы, фолксономии (wikipedia)	Меры подобия основанные на длине кратчайшего пути и его разновидности (Resnik similarity, Jiang-Conrath distance, …)
Морфология и форма слова	Расстояние Левенштейна, Меры подобия основанные на количестве общих морфем
Использование: логи, клики	Методы расширения / переформулирования поисковых запросов с помощью анализа логов и кликов.
Структура ссылок между веб-страницами	Методы анализа ссылок (Link Analysis)
Количество проиндекси-рованных страниц	Google correlation coefficient, Google Normalized Distance

Создание семантической сети из других источников информации

Разные методы представляют концепты (дескрипторы) по-разному:

Method name	**Source of information**	**Type of observation**
Distributional analysis	Text	Term is characterized by: (1) context words (paragraph, document ,…) surrounding the term, (2) context consisting of syntactic relations in which term is participating in, (3) context …
Surface analysis methods	Morphology and surface of a term	Term is characterized by: (1) sequence of its letters, (2) bag of its letters, (3) its root, (4) set of its morphemes, (5) its lemma or stem, (6)…
Dictionary-based approaches	Dictionaries and encyclopedias	Term is characterized by: (1) its definition (2) its context terms (neighbors in the dictionary) (3) …
Lexico-syntactic patterns	Text	Term is characterized by set of patterns it matches (set of patterns matching piece of text where the term occurred).
Latent Structure Approaches (LSA, PLSA, Latent Class Model)	Text	Different for different methods.
…	…	…

Разработка обобщенного метода создания семантической сети



- Разные источники информации $\{r_1,...,r_n\}$ и методы $\{m_{11},...,m_{nz}\}$ приводят к **различным оценкам** $\{\mathbf{S}_{11},...,\mathbf{S}_{nz}\}$ семантической связанности между концептами
- **Гипотеза 1**: различные оценки раскрывают **различные аспекты семантической связанности** концептов

  Один метод находит подобие формы и морфологии, другой подобие контекстов, третий подобие дефиниций и т.п.

Разработка обобщенного метода создания семантической сети

Гипотеза 2: информация о семантической связанности концептов, содержащаяся во множестве оценок превышает информацию содаржащуюся в какой-либо отдельной оценке

$$\sum_{i,j} I(\mathbf{S}_{ij}) > \text{argmax}_{\mathbf{S}_{ij}} I(\mathbf{S}_{ij})$$

Делаем вывод о необходимости разработки обобщенного метода, который бы комбинировал информацию из различных источников и методов

Цель объединения методов: Использование нескольких методов, работающих недостаточно хорошо для получения модели которая работает лучше каждого отдельного метода.

Как объединить методы?

- Объединить результаты методов с помощью линейной комбинации

$$\mathbf{S} = \sum_{i,j} \alpha \mathbf{S}_{ij}, \qquad \alpha = (n \cdot z)^{-1}$$

- Ensemble Methods: Объединить результаты с помощью взвешенной комбинации

$$\mathbf{S} = \sum_{i,j} \alpha_{ij} \mathbf{S}_{ij}, \qquad \alpha_{ij} -$$

вычисляется с помощью алгоритма машинного обучения

Создание обобщенного метода создания семантической сети

- **Объединить признаки или результаты** используемые в различных методах **в модели обобщенного представления концепта**
  - Обобщенную модель условно называем “**семантический тензор**”
  - **Тензор** это многомерный массив, объект линейной алгебры, обобщение векторов и матриц

Вектор

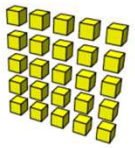
$\mathbb{R}^n$

$$\begin{matrix} 4 \\ 15 \\ -45.3 \\ 12.345 \end{matrix} \qquad \begin{matrix} 1.3 \\ 1.7 \\ 2.1 \\ 1.4 \\ 0.2 \end{matrix} \qquad \begin{matrix} 65 \\ -15 \\ -23.44 \end{matrix}$$

Матрица

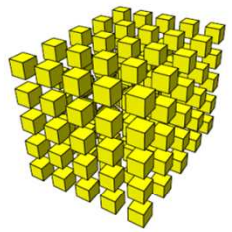
$\mathbb{R}^{m \times n}$

$$\begin{matrix} 1.3 & 0.6 & 1.1 & 2.0 & 1.9 \\ 1.7 & 1.0 & 1.3 & 2.8 & 2.6 \\ 2.1 & 2.1 & 1.6 & 3.7 & 2.7 \\ 1.4 & 1.4 & 1.4 & 2.2 & 2.0 \\ 0.2 & 1.3 & 0.9 & 1.9 & 1.7 \end{matrix}$$



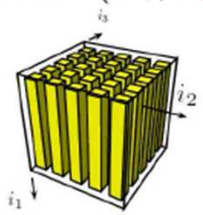
Тензор

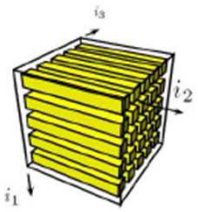
$\mathbb{R}^{m \times n \times p}$

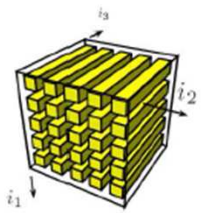


Ранг тензора

Rank-(R_1, R_2, R_3)







* Иллюстрации из M. Ishteva, P.-A. Absil, S. Van Huffel, L. De Lathauwer, “Numerical methods for the best low multilinear rank approximation of higher-order tensors”, CESAME seminar, Catholic University of Louvain, 2010.

СПАСИБО ЗА ВНИМАНИЕ

Источники

1. B. Fortuna, M. Grobelnik and D.Mladenić, Visualization of Text Document Corpus , Informatica 29 (2005) 497–502
2. Manning C., Schütze H.(1999). «Foundations of Statistical Natural Language Processing». MIT Press. Cambridge, MA.
3. Филиппович Ю., Прохоров А. (2002) «Семантика информационных технологий: опыты словарно-тезаурусного описания».
4. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. (2003). «Формирование базы терминологических словосочетаний по текстам предметной области».
5. Б.В. Добров, Н.В. Лукашевич (2001). «Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ».
6. Лукашевич Н.В. «Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России».

Источники

5. ГОСТ 7.25 - 2001. Тезаурус информационно-поисковый одноязычный
6. ISO 2788: Documentation - Guidelines for the establishment and development of monolingual thesauri. Second edition.
7. ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.
8. Gregory G. (1994) «Explorations in Automatic Thesaurus Discovery».
9. Cimiano P.(2006). «Ontology Learning and Population from Text». Algorithms, Evaluation and Applications.

11. Curran J., Moens M. (2002). «Improvements in automatic thesaurus extraction». Proceedings of the Workshop on Unsupervised Lexical Acquisition
12. Schutze H. (1998). «Automatic word sense discrimination». Computational Linguistics
13. Van der Plas L., Bouma G (2005). «Syntactic contexts for finding semantically related words». Proceedings of Computational Linguistics in the Netherlands 15.
14. Peirsman Y., Heylen K., Speelman D. (2006). «Putting things in order. First and second order context models for the calculation of semantic similarity»
15. Gregory G.(1993). «Automatic thesaurus generation from raw text using knowledge-poor techniques»