Автоматизированная система построения тезауруса AUTHECO

Цель разработки

Автоматизация процесса составления информационно-поискового тезауруса предметной области на базе корпуса текстов предметной области.

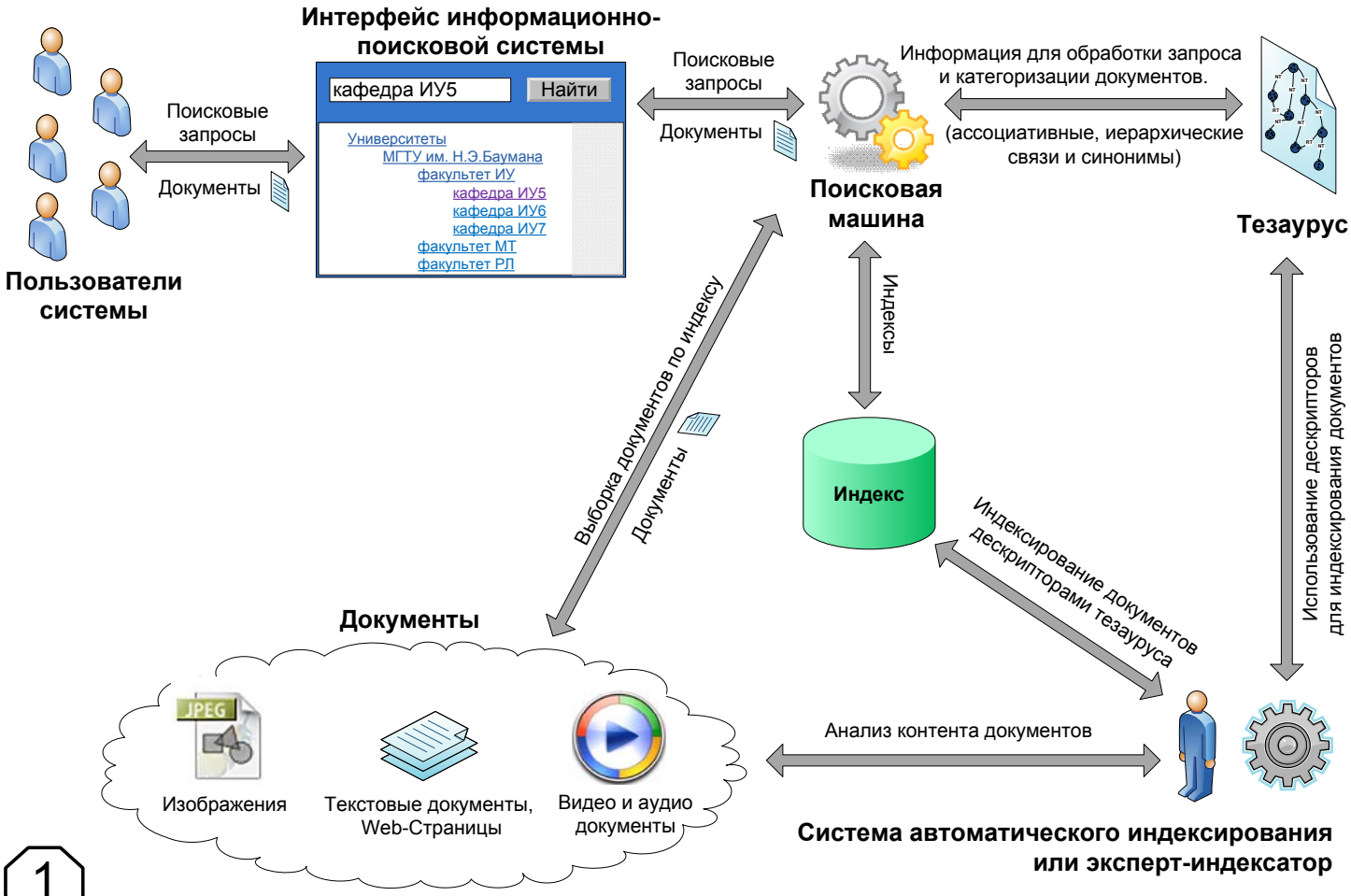
Назначение системы

Система предназначена для использования экспертами в области построения информационно-поисковых тезаурусов. Программный комплекс облегчает процесс составления семантического ресурса производя интеллектуальный анализ текстов и определяя наиболее вероятные кандидатуры слов и связей между лексическими единицами для включения в тезаурус.

Задачи, решаемые в рамках дипломного проекта

1. Анализ предметной области обработки естественных языков (NLP) и информационного поиска (IR) с использованием тезаурусов
2. Выбор методов и алгоритмов, подходящих для автоматизации процесса составления тезауруса.
3. Формализация предметной области и разработка технологии автоматизированного составления тезауруса на базе корпуса текстов данной предметной области.
4. Проектирование архитектуры системы.
5. Конструирование программной системы: кодирование алгоритмов, создание лингвистической базы данных и т.п.

Роль тезауруса в информационно-поисковой системе



Структура и назначение тезауруса

Определение

Информационно-поисковый тезаурус – контролируемый словарь терминов (нормализованной лексики) предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области. Используется для индексирования документов и при поиске информации.

Назначение

Для индексирования документов:

- Предоставление стандартного набора терминов для индексирования документов
- Решение проблемы неоднозначности естественного языка и многозначности слов
- Снижение “шума” в наборе ключевых слов
- Обеспечение последовательности в присваивании индексных терминов

Для поиска информации:

- Оптимизация пользовательских запросов используя семантическую информацию
- Использование иерархических связей для обобщения и уточнения запросов
- Переформирование запроса используя синонимы

Структура

Дескрипторы – основные понятия предметной области удовлетворяющие специальным требованиям:

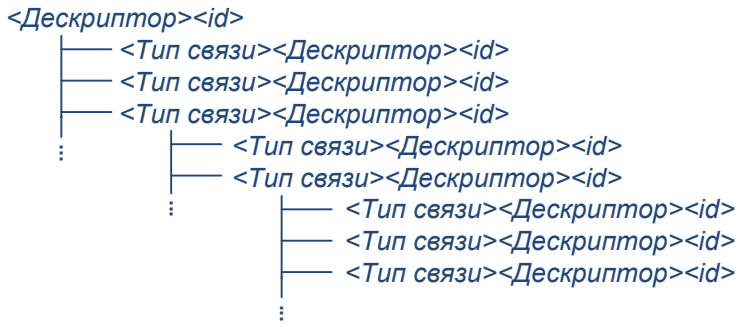
- Обозначает отдельное понятие
- Может быть однословным или многословным
- Должны быть однозначными
- Должны быть реально использоваться в текстах
- Для уточнений значений – комментарии и другие (точное описание можно найти в стандартах ГОСТ, ISO и ANSI/NISO).

Пример:
автомобилестроительный завод, рабочий, сверлильный станок

Отношения между дескрипторами:

NT Родо-видовые (иерархические):
завод NT цементный завод, кирпичный завод
RT Ассоциативные:
завод RT рабочий, станок, промышленность, Цех
USE Синонимии:
завод, фабрика USE промышленное предприятие.

Мини-тезаурус – подмножество тезауруса состоящее из дескрипторов и отношений относящихся к какой-либо предметной области (экономика, радиотехника, информатика и т.п.).



<Тип связи> ::= RT | NT | USE
<Дескриптор> ::= <Слово> | <Выражение>
<id> – число

Примеры

Фрагмент технического тезауруса

АНТЕННЫ
NT ВЫСОКОЧАСТОТНЫЕ АНТЕННЫ
NT НИЗКОЧАСТОТНЫЕ АНТЕННЫ
NT ШИРОКОПОЛОСНЫЕ АНТЕННЫ
NT ПАРАБОЛИЧЕСКИЕ АНТЕННЫ
NT ТЕЛЕСКОПИЧЕСКИЕ АНТЕННЫ
NT ШТЫРЕВЫЕ АНТЕННЫ
NT ЩЕЛЕВЫЕ АНТЕННЫ

Фрагмент тезауруса EuroVoc (мини-тезаурус “Industry”)

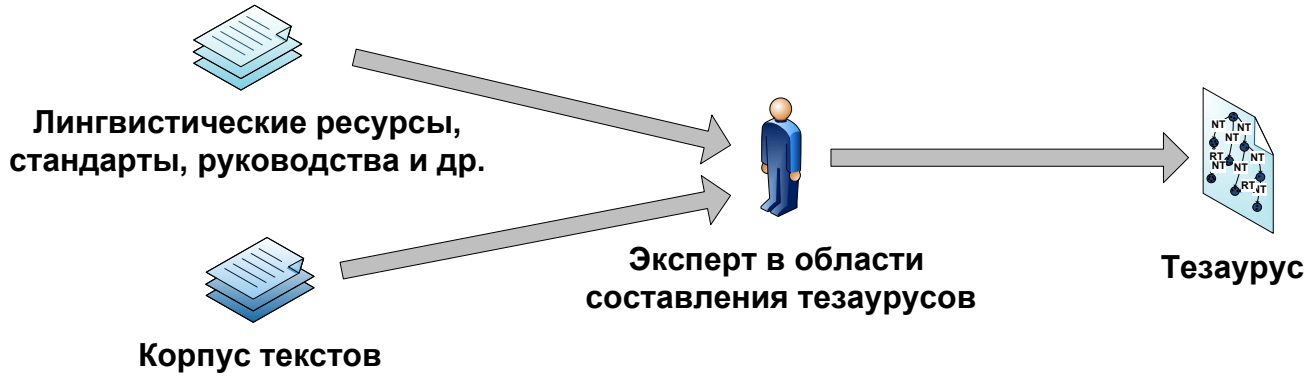
chemical industry
RT biochemistry (3606)
RT chemical pollution (5216)
RT chemistry (3606)
RT industrial chemistry (3606)
NT1 chemical accident
NT1 chemical product
RT chemical compound
RT chemical element
NT1 glass industry
NT2 glass

Фрагмент тезауруса русского языка “РуТез”

алгоритмические языки
RT языки алгоритмические
программирование
программа
теория алгоритмов
BT программное обеспечение
формальные языки
NT си
паскаль
кобол

Ручная и автоматическая технология построения тезауруса

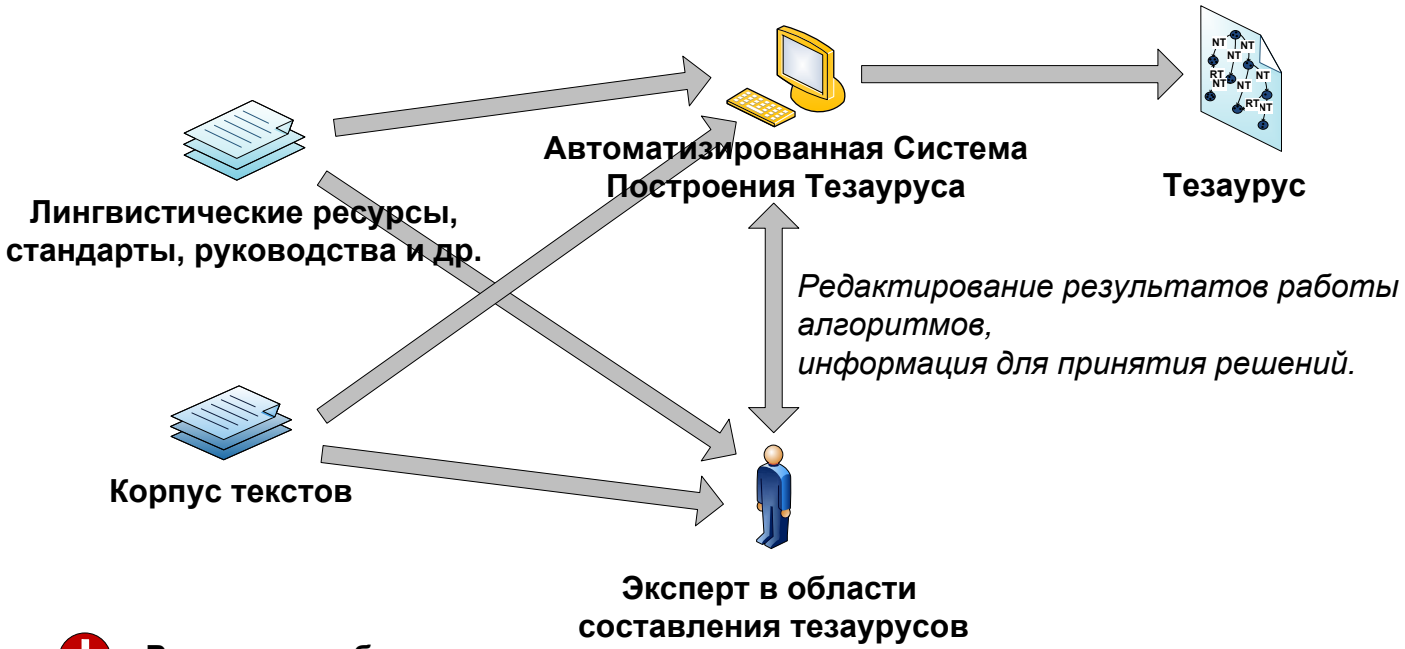
Ручное создание тезауруса



Проблемы

1. Высокая стоимость создания тезауруса
2. Длительность процесса создания тезауруса (до 1 года)
3. Зависимость результата от квалификации и опыта эксперта
4. Невозможность проанализировать весь корпус текстов
5. Малая гибкость процесса построения тезауруса

Автоматизированное создание тезауруса



Решения проблем

1, 2: Уменьшение стоимости и сроков разработки тезауруса за счет снижения трудоемкости процесса составления
3: Снижение влияния квалификации эксперта на качество тезауруса за счет использования статистических данных при построении семантического ресурса
4: Автоматическая обработка всего корпуса текстов, предоставление статистической информации о словах и словосочетаниях
5: Возможность перезапуска процедур автоматического анализа корпуса текстов

Сравнительный анализ аналогов и прототипов

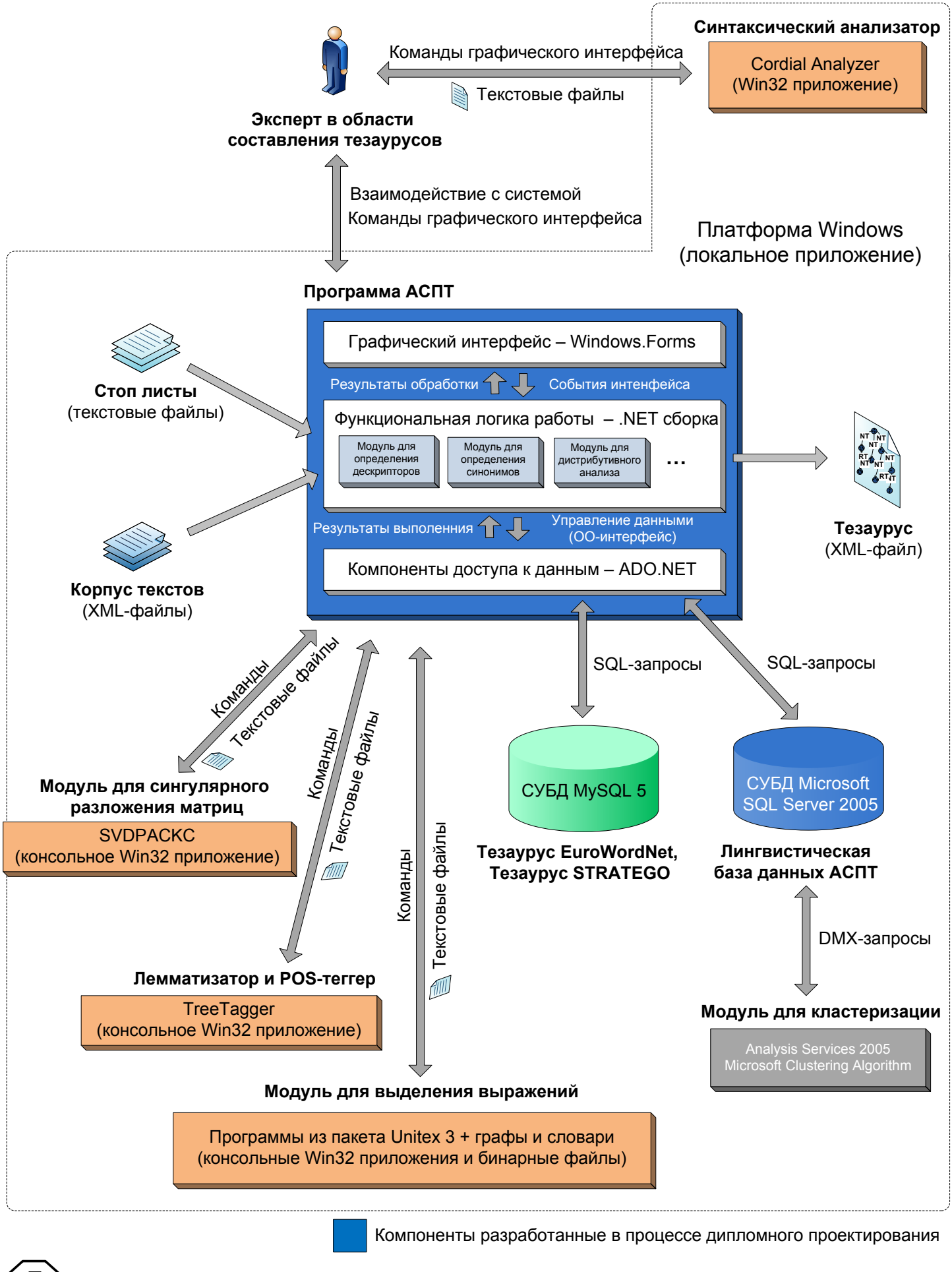
Сравнительный анализ функциональности систем

	$К_{зн}$	XIP	Oracle Text	GALILEI	a.k.a.	AUTHECO
Готов к использованию	0,1	—	—	+	+	+
Препроцессинг	0,15	+	+	+	—	+
Извлечение выражений из текста	0,15	+	—	+	—	+
Построение множества предпочтительных дескрипторов	0,15	—	—	—	—	+
Нахождение отношений синонимии	0,05	—	—	—	—	+
Определение отношений между дескрипторами	0,2	+	+	—	—	+
Кластеризация документов	0,1	—	+	—	—	+
Средства хранения и редактирования тезауруса	0,1	—	+	—	+	+
Сумма =		**0,5**	**0,55**	**0,4**	**0,2**	**1**

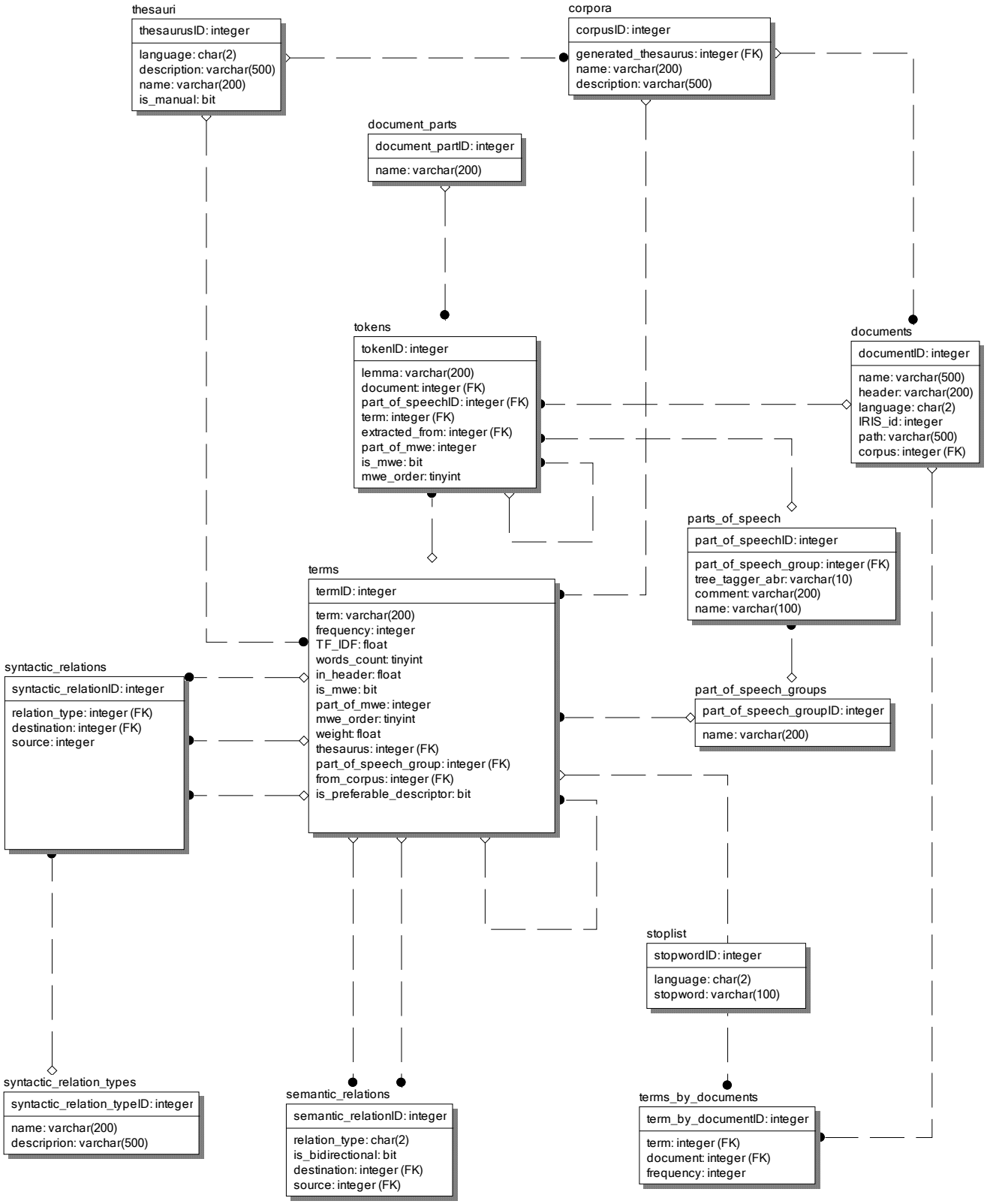
Сравнительный анализ по методам обработки данных

	Препроцессинг	Извлечение выражений из текста	Построение множества предпочтительных дескрипторов	Нахождение отношений синонимии	Определение отношений между дескрипторами	Кластеризация документов
XIP	Лемматиция, Части речи, Синтаксические зависимости	N-граммы, специальные грамматики и др. лингвистические ресурсы	—	—	Дистрибутивно-статистический анализ, Методы Data Mining, Алгоритм Apriori	—
Oracle Text	Лемматизация, Исправление опечаток	—	—	—	Методы Data Mining: алгоритм Apriori, Association Rules и др.	Enhanced K-Means, Orthogonal Portioning Clustering
GALILEI	Лемматизация	—	—	—	—	—
a.k.a.	—	—	—	—	—	—
AUTHECO	Лемматизация, Части речи, Синтаксические зависимости	Лингвистические графы и словари	Функция отбора, основанная на статист. и лингвист. данных	Использование базы синонимов	Дистрибутивно-статистический анализ	Латентно-семантический анализ + K-Means

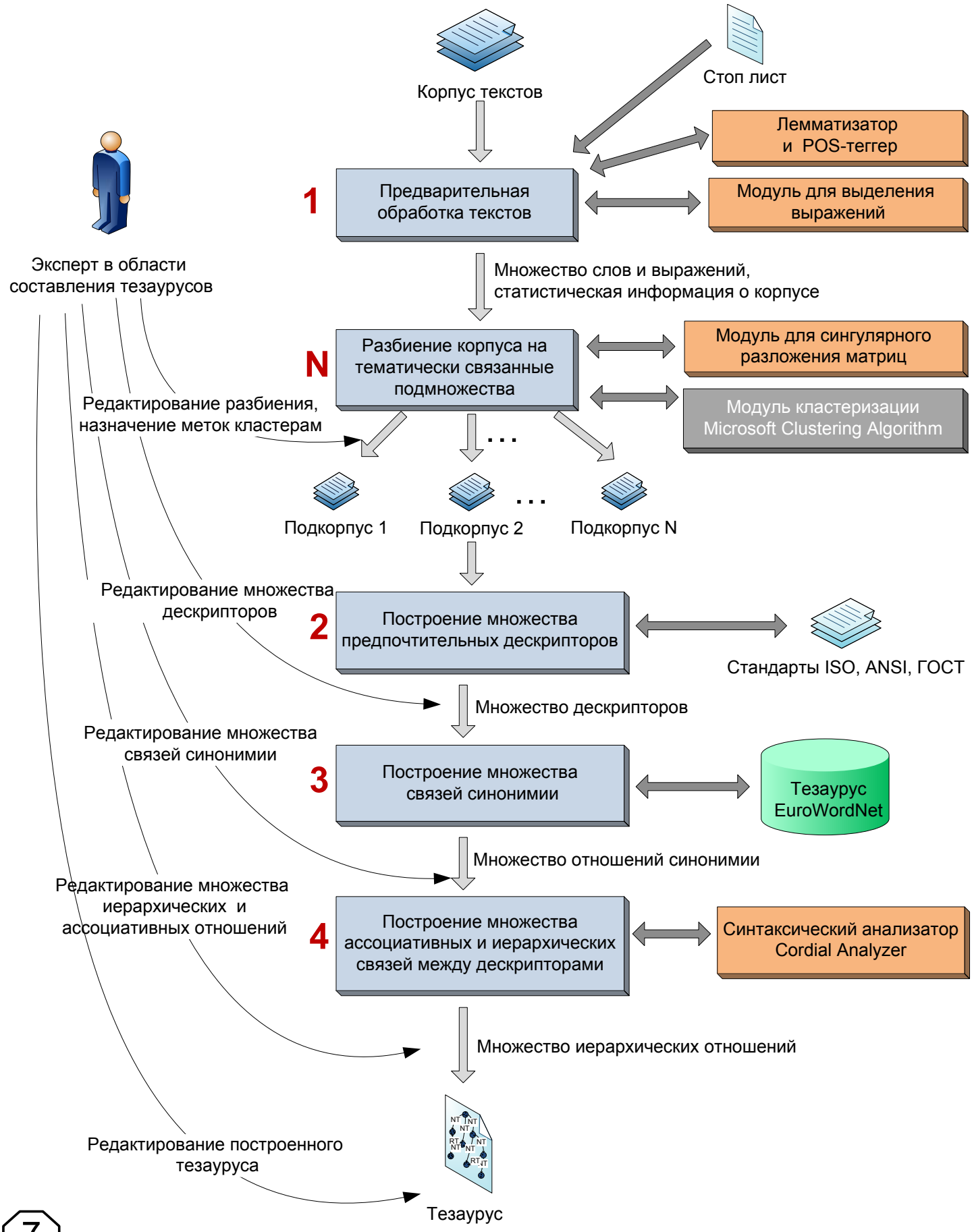
Архитектура системы AUTHECO (АСПТ)



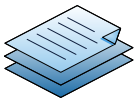
Даталогическая схема лингвистической базы данных АСПТ



Технология автоматизированного построения тезауруса



Предварительная обработка текстов (1)



Корпус текстов

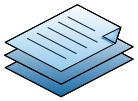
Пример документа:

По сути дела, пищевая промышленность - пример того, на что способна российская экономика при ...

Модуль для выделения выражений

Программы из лингвистического пакета Unitex 3, специальные графы и словари

Разметка документов XML-тегами с выражениями



Помеченный корпус текстов

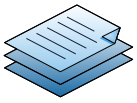
Пример документа:

По сути дела, **<MWE>**пищевая промышленность**</MWE>** - пример того, на что способна **<MWE>**российская экономика**</MWE>** при ...

Токенизатор, Лемматизатор и POS-теггер

TreeTagger

Извлечение всех слов и выражений, определение части речи и леммы



Лемматизированный, токетизированный и POS-размеченный корпус текстов

Статический коэффициент TF-IDF:

$$TF \times IDF = \frac{n_i}{\sum_k n_k} \cdot \log \frac{|D|}{|(d_i \supset t_i)|},$$

где n_i есть число рассматриваемых употреблений слова, а в знаменателе — общее число словоупотреблений, $|D|$ — количество документов в корпусе, $|(d_i \supset t_i)|$— количество документов, в которых встречается t_i (когда $n_i \neq 0$).

Модуль вычисления статистики

Вычисление частотных словников, коэффициентов TF*IDF и др.

Список слов и выражений

Lemma	Frequency	POS	TF_IDF	D
в	745	предлог	0.23	
пищевая индустрия	33	сс	0.74	2
важная	51	прилаг	0.25	1
для	23	предлог	0.26	2
качество	102	существств	0.27	3
продукты питания	329	существств	0.58	3
покупатель	330	существств	0.49	2
из	331	предлог	0.30	1
когда	332	наречье	0.31	1
сырье	333	существств	0.32	4
...	...	...	...	...

Модуль работы со стоп-листом

Удаление слов из стоп листа

Стоп лист

Слова и выражения без слов из стоп-листа

Lemma	Frequency	POS	TF_IDF	D
в	745	предлог	0.23	
пищевая индустрия	33	сс	0.74	2
важная	51	прилаг	0.25	1
для	23	предлог	0.26	2
качество	102	существств	0.27	3
продукты питания	329	существств	0.58	3
покупатель	330	существств	0.49	2
из	331	предлог	0.30	1
когда	332	наречье	0.31	1
сырье	333	существств	0.32	4
...	...	...	...	...

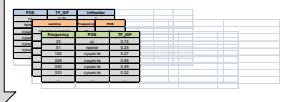
ВЫХОД

Сохранение результатов в лингвистической базе данных АСПТ

Построение множества предпочтительных дескрипторов (2)

Стандарты ISO, ANSI, ГОСТ

Извлеченные слова и выражения, статистическая и лингвистическая информация о корпусе



Модуль взвешивания дескрипторов

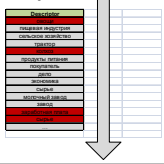
Множество взвешенных дескрипторов

Модуль построения множества дескрипторов

Отбор дескрипторов:

Во множество предпочтительных дескрипторов D_{pref} включаются те дескрипторы, d_i вес которых не менее заданного минимума:
$d_i \in D_{pref} : \omega_i \geq \omega_{min}$.

Множество предпочтительных дескрипторов



Графический интерфейс редактирования множества дескрипторов

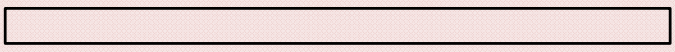
Редактирование множества дескрипторов

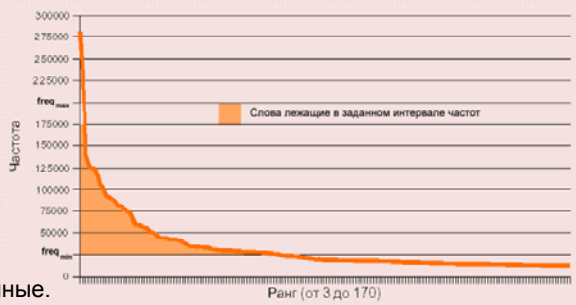
Эксперт в области составления тезаурусов

Окончательное множество предпочтительных дескрипторов

ВЫХОД Сохранение результатов в лингвистической базе данных АСПТ

Взвешивание дескрипторов функцией отбора:





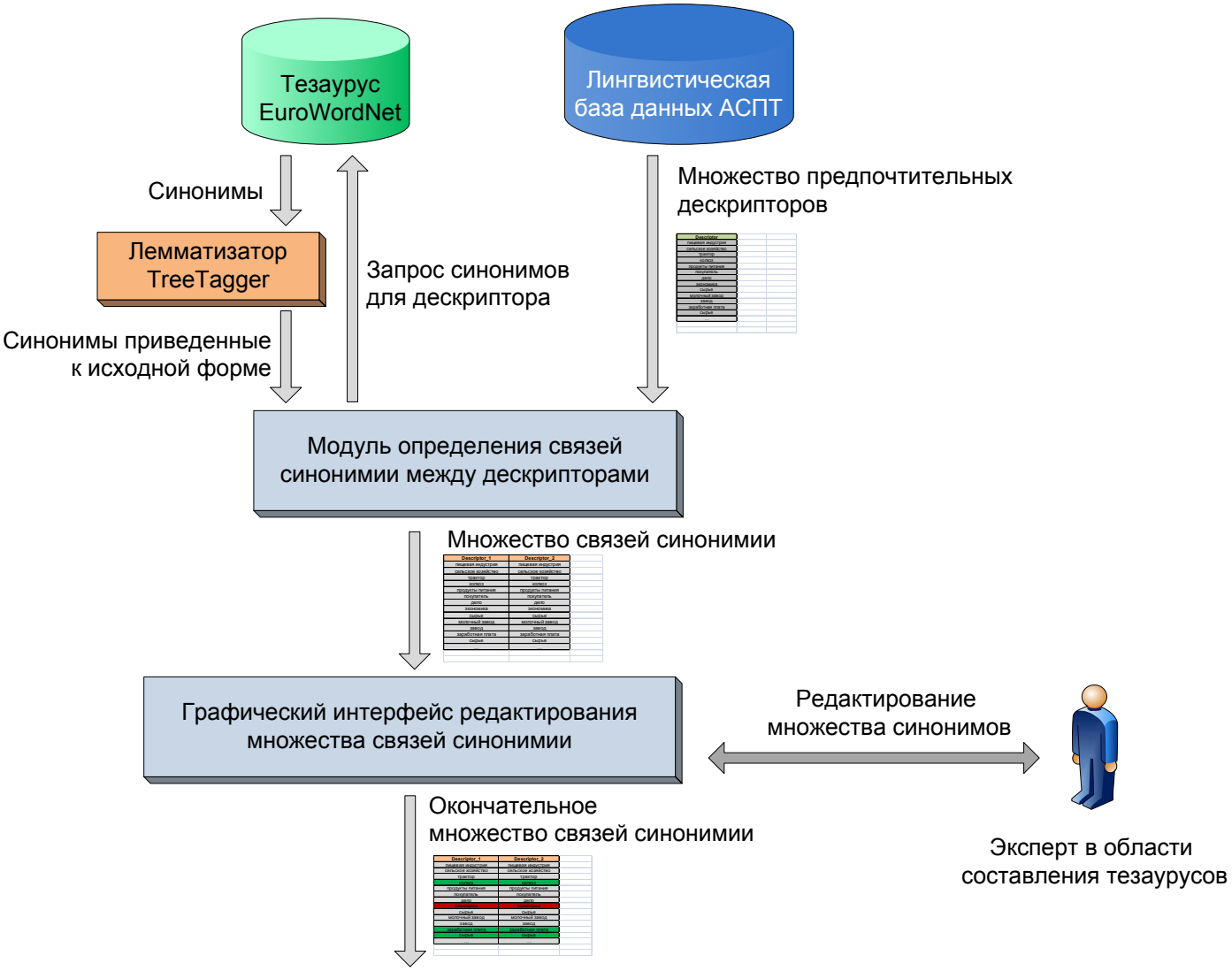
Закон Ципфа

Функция представляет собой комбинацию ограничений на данные переменные.
К примеру, $freq_i$ должен лежать в определенном интервале $freq_i \in [freq_{min}; freq_{max}]$ в соответствии с законом Ципфа (см. рисунок).

Выполнил: Панченко А.И

Утвердил: Филиппович А.Ю.

Построение множества связей синонимии (3)



Алгоритм определения связей синонимии между дескрипторами

Множество дескрипторов
Пусть множество *предпочтительных дескрипторов* – $D_{pref} = \{d_1, d_2, \dots, d_n\}$,
где n – мощность множества дескрипторов;
причем d_i – дескриптор, являющийся упорядоченным кортежем, таким что $d_i = (w_1, w_2, \dots, w_k),\ k > 0$,
где w_i – слово.

Множество связей синонимии
Пусть S_{ewn} – непустое множество всех связей синонимии содержащихся в тезаурусе EuroWordNet, такое что $S_{ewn} = \{s_1, s_2, \dots, s_m\}$,
где s_i – отношение синонимии которое можно представить в виде неупорядоченной пары вида $(d_i; d_j)$ или записывая по-другому $d_i \rightleftharpoons d_j$.

Алгоритм нахождения синонимов
Алгоритм нахождения синонимов в заданном множестве дескрипторов заключается в построении множества отношений
синонимии $S_D = \{s_1, s_2, \dots, s_l\};\ S_D \subseteq S_{ewn}$.

1. Для каждого дескриптора d_i из множества D_{pref} находим такое подмножество $S_i \subseteq S_{ewn}$ что $S_i = \{s_1^i, s_2^i, \dots, s_t^i\}$: $\forall\, s_j^i = (d_i; d_k); j \in [1; t],\ d_k \in D_{pref}$.
   Иначе говоря, отбираем те связи синонимии, которые содержат дескрипторы только из D_{pref}.
2. Добавляем множество S_i к искомому: $S_D = S_D \cap S_i$.
3. Переходим к следующему дескриптору. Если все дескрипторы перебраны тогда выход.

Выполнил: Панченко А.И
Утвердил: Филиппович А.Ю.

Построение множества ассоциативных и иерархических связей между дескрипторами (4-a)

Лингвистическая база данных АСПТ

Лемматизированный, токетизированный и POS-размеченный корпус текстов

Множество предпочтительных дескрипторов

Корпус текстов

4-b

Модуль построения векторной модели текста. Алгоритмы дистрибутивного анализа first order bag-of-words и синтаксический метод.

Текстовые файлы

Синтаксический анализатор Cordial Analyzer

** только для синтаксического метода*

Множество векторов признаков

Нормализация (метода контекстного окна):

Модуль нормализации множества вектора признаков

PMI – Поточечное полное количество информации

$$PMI(x,y) = log\frac{P(x,y)}{P(x)P(y)} = log\frac{freq_{(x;y)}}{freq_x \cdot freq_y},$$

где $freq_i$ – частота встречаемости слова.

Множество нормализованных векторов признаков

Вычисление матрицы подобия:

Модуль вычисления матрицы подобия

$$F'_{D_{pref}} = \begin{bmatrix} \overrightarrow{f_{d_1}} \\ \overrightarrow{f_{d_2}} \\ \overrightarrow{f_{d_3}} \\ \dots \\ \overrightarrow{f_{d_n}} \end{bmatrix} = \begin{bmatrix} e_{11} & e_{12} & e_{13} & \dots & e_{1m} \\ e_{21} & e_{22} & e_{23} & \dots & e_{2m} \\ e_{31} & e_{32} & e_{33} & \dots & e_{3m} \\ \dots & \dots & \dots & \dots & \dots \\ e_{m1} & e_{m2} & e_{m3} & \dots & e_{mn} \end{bmatrix} = \begin{bmatrix} 6 & 43 & 31 & \dots & 32 \\ 52 & 142 & 17 & \dots & 26 \\ 46 & 62 & 22 & \dots & 75 \\ \dots & \dots & \dots & \dots & \dots \\ 312 & 12 & 8 & \dots & 85 \end{bmatrix} \begin{matrix} \cos\propto_{12} \\ \cos\propto_{23} \\ \\ \cos\propto_{(n-1)n} \end{matrix}$$

где $\overrightarrow{f'_{d_i}}$ – вектор признаков для дескриптора d_i .

Для вычисления матрицы подобия S_F используется $cos\alpha$ между векторами признаков.

Так $S_F(d_\alpha, d_\beta) = \cos\left(\overrightarrow{f'_{d_\alpha}}, \overrightarrow{f'_{d_\beta}}\right) = \frac{\overrightarrow{f'_{d_\alpha}} \cdot \overrightarrow{f'_{d_\beta}}}{|f'_{d_\alpha}| \cdot |f'_{d_\beta}|} = \frac{\sum_{i=1}^{m}(e_{\alpha i} \cdot e_{\beta i})}{\sqrt{\sum_{i=1}^{m} e_{\alpha i}^2 \cdot \sum_{i=1}^{m} e_{\beta i}^2}}$

Матрица подобия

Модуль построения множества отношений

Отбор отношений:

Во множество отношений R_{pref} включаются те отношения, r_i сила связи которых w_i не менее заданного минимума:

$r_i \in R_{pref} : w_i \geq w_{min}$.

Множество отношений для включения в тезаурус (без указания типа)

Descriptor_1	Descriptor_2	Relation	Type
пищевая индустрия	индустрия	0.54	
сельское хозяйство	фермер	0.55	
трактор	техника	0.56	
молоко	молочная	0.77	
продукты питания	молоко	0.48	
покупатель	продавец	0.49	
дело	время	0.50	
молочный завод	завод	0.61	
сырье	зерно	0.62	
молочный завод	завод	0.63	
завод	труба	0.64	
заработная плата	премия	0.65	
...	...	...	...

Графический интерфейс редактирования множества связей синонимии

Редактирование множества отношений и **указание типа отношений**

Эксперт в области составления тезаурусов

Окончательное множество связей с указанием типа

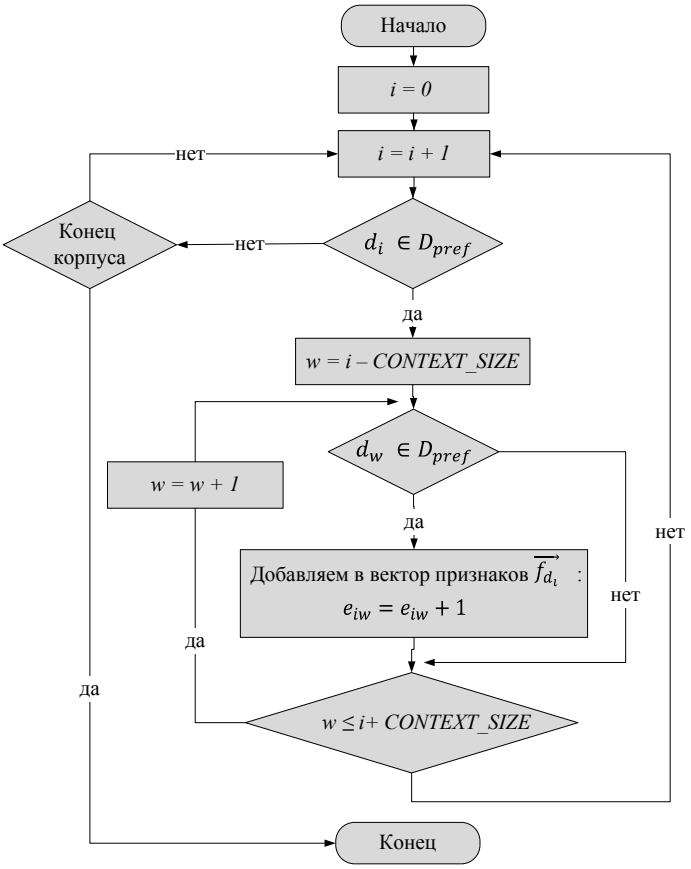
Descriptor_1	Descriptor_2	Type
пищевая индустрия	индустрия	NT
сельское хозяйство	фермер	RT
трактор	техника	NT
молоко	молочная	RT
продукты питания	молоко	NT
покупатель	продавец	RT
дело	время	RT
молочный завод	завод	NT
сырье	зерно	NT
молочный завод	завод	NT
завод	труба	RT
заработная плата	премия	RT
...	...	...

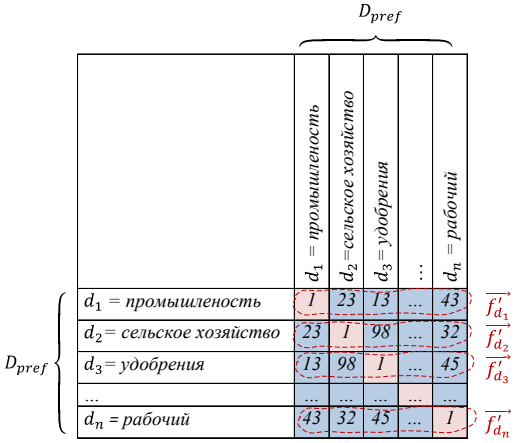
ВЫХОД Сохранение результатов в лингвистической базе данных АСПТ

Построение векторной модели текста в алгоритмах дистрибутивно-статистического анализа (4-b)

Метод контекстного окна (first order bag-of-words)

Дистрибутивная гипотеза – слова появляющиеся в одинаковых контекстах связаны.



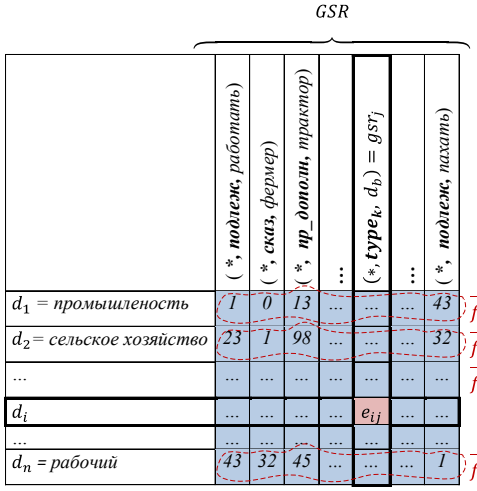


Формат матрицы признаков, где $\overrightarrow{f_{d_i}}$ – вектор признаков дескриптора d_i

Параметр алгоритма – размер контекстного окна (CONTEXT_SIZE).
1 слово ≤ CONTEXT_SIZE ≤ 1 параграф
i – указатель на обрабатываемый дескриптор
w – указатель на слово обрабатываемое внутри контекстного окна

Пример для контекстного окна = 1 слово:

```
[Снятие  омонимии] полезно во  многих приложениях...
[Снятие  омонимии  полезно] во  многих приложениях...
 Снятие [омонимии  полезно  во] многих приложениях...
 Снятие  омонимии [полезно  во  многих] приложениях...
 Снятие  омонимии  полезно [во  многих  приложениях]...
```



Формат матрицы признаков,

где $\overrightarrow{f_{d_i}}$ – вектор признаков дескриптора d_i

Синтаксические зависимости в предложении:

«Лемматизация играет важную роль в задачах компьютерной лингвистики.»
(лемматизация, **подлеж_гл**, играть); (играть, **сказ**, лемматизация);
(играть, **прям_дополн_гл**, важная роль)…$(d_i, type_k, d_j)$.

d_i, d_j – дескрипторы, причем $d_i \in D_{pref}$

$type_k$ – тип синтаксической связи $type_k \in T$

Синтаксический метод

$SR_{all} = \{sr_1, sr_2, \dots, sr_n\}$ – множество всех извлеченных синтаксических зависимостей, причем sr_i – упорядоченная тройка вида $(d_a, type, d_b)$.

$SR_{pref} \subseteq SR_{all}$; $SR_{pref} = \{sr \mid (sr \in SR) \wedge (type \in T_{pref}) \wedge (d_a \in D_{pref})\}$, где
$T_{pref} = \{type_1, type_2, \dots, type_k\} = \{подлеж\_гл\_v,\ прям\_дополн\_глаг\_v,\ обстоят\_глаг\_v,\ \dots, имеет\_предл\_p\}$
– множество рассматриваемых в методе синтаксических зависимостей.
$T_{pref} \subseteq T$ – множество всех извлеченных типов отношений с использованием синтаксического анализатора.

$GSR = \{gsr_1, gsr_2, \dots, gsr_m\} = \{gsr_i \mid (\exists\, sr = (d_a, type, d_b) \in SR_{pref})\}$, таким образом GSR содержит множество элементов типа $(*, подлеж\_гл,\ работать)$