МЕЖДУНАРОДНЫЙ СОЛОМОНОВ УНИВЕРСИТЕТ

Дмитрий Владимирович ЛАНДЭ

Лекция 11

“Основы концепции “глубинного анализа текстов” (Text Mining)”

Контент-анализ: определения

Один из истоков концепции Text Mining – контент-анализ. Понятие контент-анализа, корни которого в психологии и социологии, не имеет однозначного определения:

- *Контент-анализ - это методика объективного качественного и систематического изучения содержания средств коммуникации (Д.Джери, Дж. Джери).*
- *Контент-анализ - это систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника (Д. Мангейм, Р. Рич).*
- *Контент-анализ - это качественно-количественный метод изучения документов, которое характеризуется объективностью выводов и строгостью процедуры и состоит из квантификационной обработки текста с дальнейшей интерпретацией результатов (В. Иванов).*
- *Контент-анализ состоит из нахождения в тексте определенных содержательных понятий (единиц анализа), выявление частоты их встречаемости и соотношение с содержанием всего документа (Б. Краснов).*
- *Контент-анализ - это исследовательская техника для получения результатов путем анализа содержания текста о состоянии и свойствах социальной действительности (Э. Таршис).*

Контент-анализ и добыча данных

Контент-анализ в рамках исследования электронных информационных массивов - относительно новое направление, которое предусматривает анализ множеств текстовых документов.

Принято распределение методологий контент-анализа на две области: качественную и количественную. Основа количественного контент-анализа - частота появления в документах определенных характеристик содержания. Качественный контент-анализ основан на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания.

Технологии глубинного анализа текста Text Mining исторически предшествовала технология добычи данных, методология и подходы которой широко используются.

Основные задачи Text Mining

Как и большинство когнитивных технологий – Text Mining – это алгоритмическое выявление прежде не известных связей и корреляций в уже имеющихся текстовых данных.

Важная задача технологии Text Mining связана с извлечением из текста его характерных элементов или свойств, которые могут использоваться как метаданные документа, ключевых слов, аннотаций.

Другая важная задача состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации. Text Mining также обеспечивает новый уровень семантического поиска документов.

Возможности современных систем Text Mining могут применяться при управлении знаниями для выявления шаблонов в тексте, для автоматического «выталкивания» или размещения информации по интересующим пользователей профилям, создавать обзоры документов.

Основные элементы Text Mining

В соответствии с уже сформированной методологии к основным элементам Text Mining относятся:

- классификация (classification),
- кластеризация (clustering),
- построение семантических сетей,
- извлечение фактов, понятий (feature extraction),
- суммаризация (summarization),
- ответ на запросы (question answering),
- тематическое индексирование (thematic indexing),
- поиск по ключевым словам (keyword searching).

Также в некоторых случаях набор дополняют средства поддержки и создание таксономии (oftaxonomies) и тезаурусов (thesauri).

Классификация

При классификации текстов используются статистические корреляции для построения правил размещения документов в определенные категории. Задача классификации - это классическая задачу распознавания, где по некоторой контрольной выборке система относит новый объект к той или другой категории.

Особенность систем Text Mining заключается в том, что количество объектов и их атрибутов может быть очень большой, поэтому должны быть предусмотрены интеллектуальные механизмы оптимизации процесса классификации.

В существующих сегодня системах классификация применяется, например, в таких задачах: группировка документов в intranet-сетях и на Web-сайтах, размещение документов в определенные папки, сортировка сообщений электронной почты, избирательное распространение новостей подписчикам.

Кластеризация

Кластеризация базируется на признаках документов, которые использует лингвистические и математические методы без использования определенных категорий. Результат - таксономия или визуальная карта, которая обеспечивает эффективный охват больших объемов данных.

Кластеризация в Text Mining рассматривается как процесс выделения компактных подгрупп объектов с близкими свойствами. Система должна самостоятельно найти признаки и разделить объекты по подгруппам. Кластеризация, как правило, передует классификации, поскольку разрешает определить группы объектов. Различают два основных типа кластеризации - иерархическую и бинарную.

Кластеризация применяется при реферировании больших документальных массивов, определение взаимосвязанных групп документов, упрощения процесса просмотра при поиске необходимой информации, нахождения уникальных документов из коллекции, выявления дубликатов или очень близких по содержанию документов.

Другие элементы

Построение семантических сетей

Построение семантических сетей или анализ связей, которые определяют появление дескрипторов (ключевых фраз) в документе для обеспечения навигации.

Извлечение фактов

Извлечение фактов, предназначенное для получения некоторых фактов из текста с целью улучшения классификации, поиска и кластеризации.

Прогнозирование

Состоит в том, чтобы предсказать по значениям одних признаков объекта значения остальных.

Нахождение исключений

Поиск объектов, которые своими характеристиками сильно выделяются из общей массы.

Визуализация.

Визуализация используется как средство представления контента текстовых массивов, а также для реализации навигационных механизмов.

Автоматическое реферирование

Автоматическое реферирование (Automatic Text Summarization) - это составление коротких изложений материалов, аннотаций или дайджестов, т. е. извлечения наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных и информационно-насыщенных отчетов. Существует два направления автоматического реферирования - квазиреферирование и краткое изложение содержания.

Квазиреферирование основано на экстрагировании фрагментов документов - выделении наиболее информативных фраз и формировании из них квазирефератов.

Краткое изложение исходного материала основывается на выделении из текстов с помощью методов искусственного интеллекта и специальных информационных языков наиболее важной информации и порождении новых текстов, содержательно обобщающих первичные документы.

Семантические методы формирования рефератов-изложений допускают два основных подхода: метод синтаксического разбора предложений, и методы, базирующиеся на понимании естественного языка, методах искусственного интеллекта.

3 направления квазиреферирования

В рамках квазиреферирования выделяют три основных направления, зачастую применяемых совместно:

- статистические методы, основанные на оценке информативности разных элементов текста по частоте встречаемости, которая служит основным критерием информативности слов, предложений или фраз;
- позиционные методы, которые опираются на предположение о том, что информативность элемента текста есть зависимым от его позиции в документе;
- индикаторные методы, основанные на оценке элементов текста, исходя из наличия в них специальных слов и словосочетаний - маркеров важности, что характеризуют их содержательную значимость.

Определение веса фрагментов при квазиреферирования

Определение веса фрагментов (предложений или абзацев) исходного текста выполняется по алгоритмам, которые стали уже традиционными. Общий вес текстового блока на этом этапе определяется по формуле:

$$Weight = Location + KeyPhrase + StatTerm$$

Коэффициент ***Location*** определяется расположением блока в исходном тексте и зависит от того, где появляется данный фрагмент - в начале, в середине или в конце, а также используется ли он в ключевых разделах текста, например, в выводе.

Ключевые фразы (***KeyPhrase***) представляют собой конструкции-маркеры, которые резюмируют, типа "в заключение", "в данной статье", "в результате анализа" и т.п. Весовой коэффициент ключевой фразы может зависеть также от оценочного термина, например, "отличный".

Статистический вес текстового блока (***StatTerm***) вычисляется как нормированная по длине блока сумма весов входящих в него строк - слов и словосочетаний.

Поисковые образы документов

На основе методов автоматического реферирования возможно формирование поисковых образов документов. По автоматически построенным аннотациям больших текстов (поисковым образам документов) проводится поиск, который характеризуется высокой точностью (естественно, за счет полноты).

В этом случае аннотированные тексты рассматриваются как поисковые образы документов (ПОД). Хотя ПОД часто для больших документов оказывается образованием, лишь отдаленно напоминающим исходный текст и не всегда оказывается воспринимаемым человеком, но за счет содержания наиболее весомых ключевых слов и фраз, он может приводить к вполне адекватным результатам при полнотекстовом поиске.

Особенности реализации систем

Рассматриваются системы:

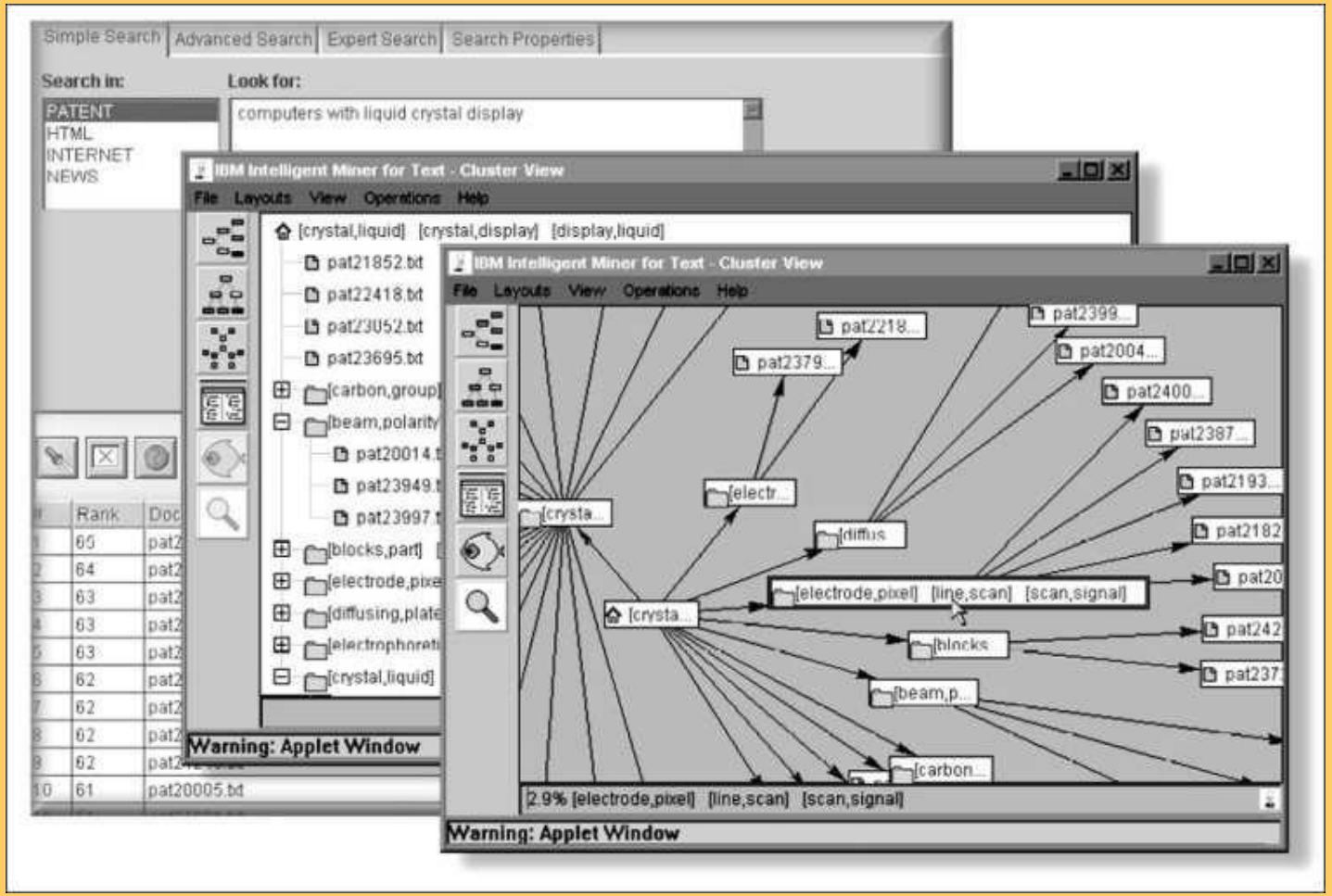
- Intelligent Miner for Text (IBM)
- PolyAnalyst (Мегапьютер Інтеллидженс)
- Text Miner (SAS)
- SemioMap (Semio Corp.)
- Oracle Text (Oracle)
- Knowledge Server (Autonomy)
- RetrievalWare (Convera)
- Galaktika-ZOOM (корпорация "Галактика")
- InfoStream (ИЦ "ЭЛВИСТИ")

Intelligent Miner for Text (IBM)

(http://www-3.ibm.com/software/data/iminer/fortext/)
Система является одним из лучших инструментов глубинного анализа текстов. Содержит утилиты :
Language Identification Tool - утилита определения языка - для автоматического определения языка, на котором составлен документ.
Categorisation Tool - утилита классификации - автоматического отнесения текста к некоторой категории (входной информацией на обучающей фазе работы этого инструмента может служить результат работы следующей утилиты - Clusterisation Tool).
Clusterisation Tool - утилита кластеризации - разбиения большого множества документов на группы по близости стиля, формы, различных частотных характеристик выявляемых ключевых слов.
Feature Extraction Tool - утилита определения нового - выявление в документе новых ключевых слов (собственные имена, названия, сокращения) на основе анализа заданного заранее словаря.
Annotation Tool - утилита "выявления смысла" текстов и составления рефератов - аннотаций к исходным текстам.

Intelligent Miner for Text (IBM)

Визуализация кластеров в IBM Intelligent Miner for Text:



PolyAnalyst (Мегапьютер Інтеллидженс)

(http://www.megaputer.com/)

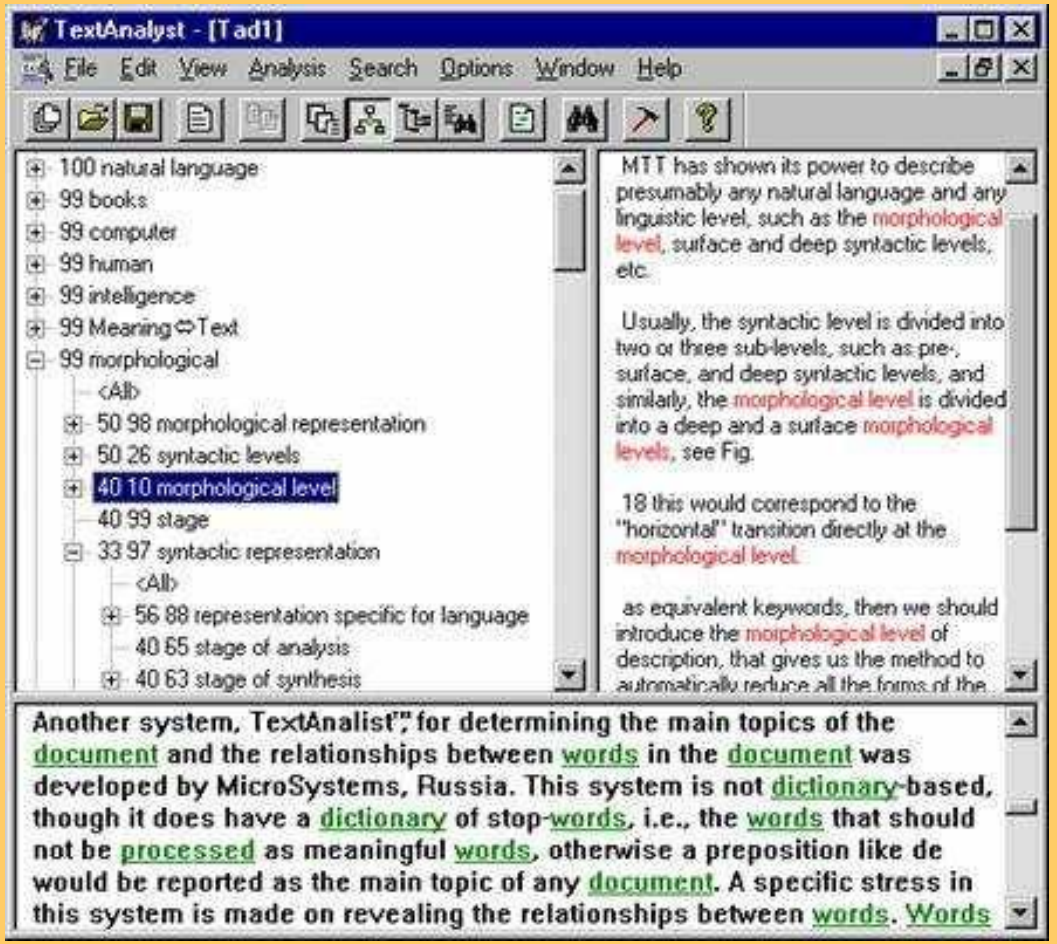
PolyAnalyst может применяться для автоматизированного анализа числовых и текстовых баз данных с целью обнаружения ранее неизвестных, нетривиальных, полезных и доступных пониманию закономерностей.

PolyAnalyst является клиент-серверным приложением. При этом пользователь работает с программой PolyAnalyst Workplace. Математические же модули выделены в серверную часть - PolyAnalyst Knowledge Server.

PolyAnalyst работает с разными типами данных. Это - числа, логические переменные, текстовые строки, даты, а также свободный текст. PolyAnalyst может обрабатывать исходные данные из различных источников, к примеру, файлы Microsoft Excel 97/2000, ODBC- совместимая СУБД, SAS data files, Oracle Express, IBM Visual Warehouse.

TextAnalyst

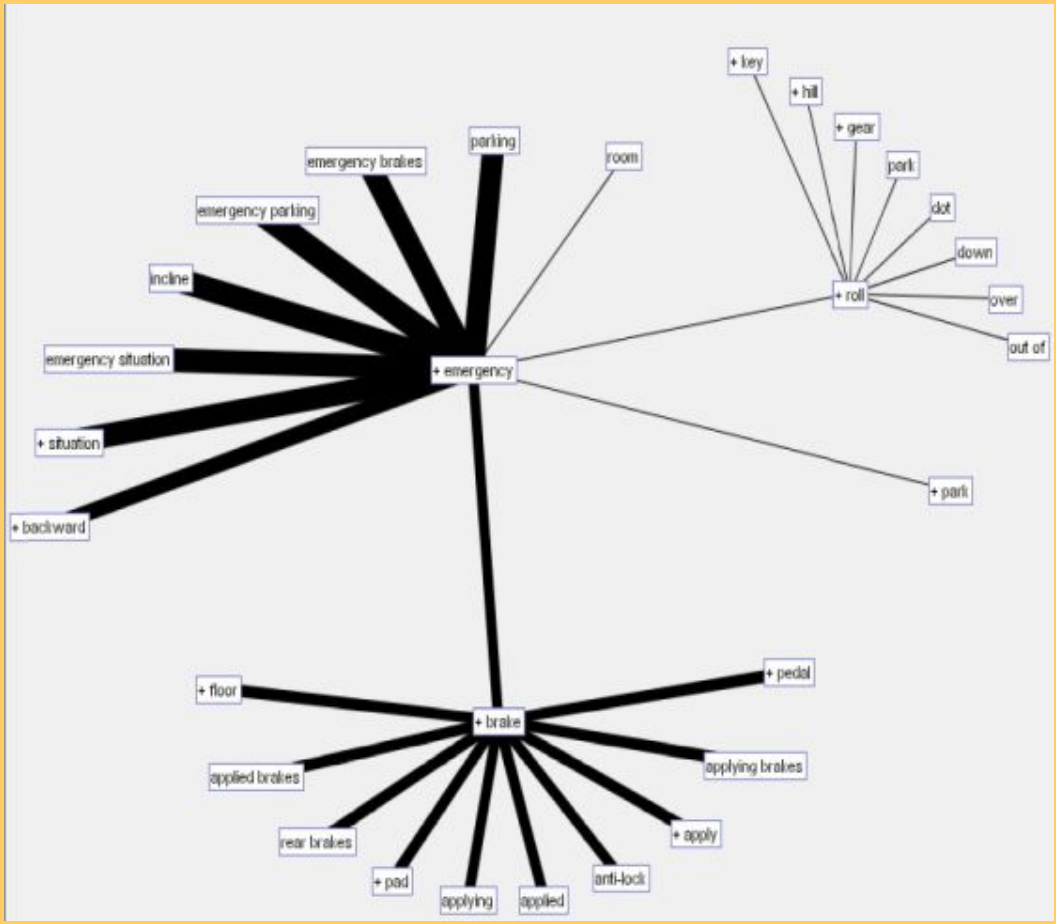
В состав PolyAnalyst входит система TextAnalyst, которая решает такие задачи Text Mining: создание семантической сети большого текста, подготовка резюме текста, поиск по тексту и автоматическая классификация и кластеризация текстов. Построение семантической сети - это поиск ключевых понятий текста и установление взаимоотношений между ними.



Text Miner (SAS)

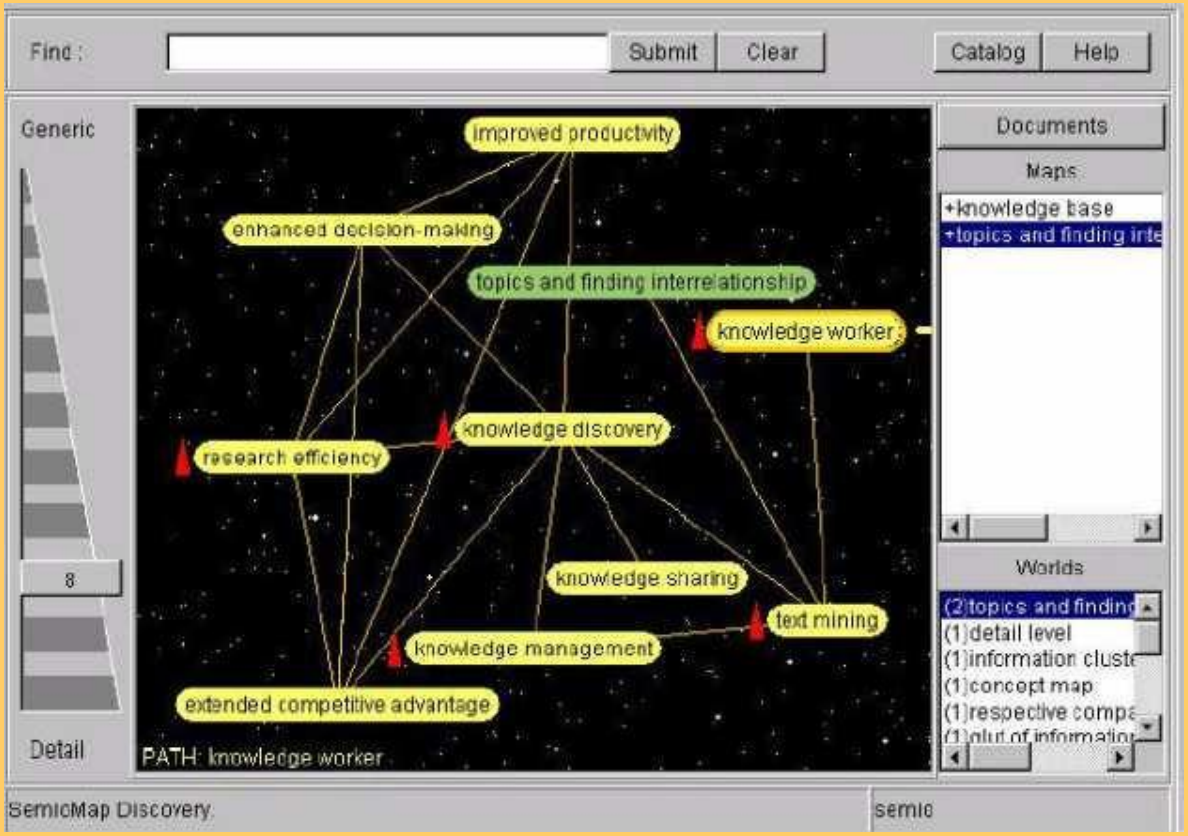
http://www.sas.com/technologies/analytics/datamining/textminer/

Система SAS Text Miner может работать с текстовыми документами различных форматов из баз данных, файловых систем и Web. Text Miner обеспечивает логическую обработку текста в среде мощного пакета SAS Enterprise Miner. Это позволяет интегрировать текстовую информацию со структурированными данными.



SemioMap (Semio Corp.)

http://www.entrieva.com/entrieva/products/semiomap.asp?Hdr=semiomap
SemioMap - это продукт компании Entrieva, созданный в 1996 г. ученым-семиотиком Клодом Фогелем (Claude Vogel). В мае 1998 г. продукт был выпущен как промышленный комплекс SemioMap 2.0 - первая система Text Mining, работающая в архитектуре клиент-сервер.



SemioMap (Semio Corp.)

Система SemioMap состоит из двух основных компонент - сервера SemioMap и клиента SemioMap. Работа системы протекает в три фазы:

Индексирование: сервер SemioMap автоматически читает массивы неструктурированного текста, извлекает ключевые фразы (понятия) и создает из них индекс;

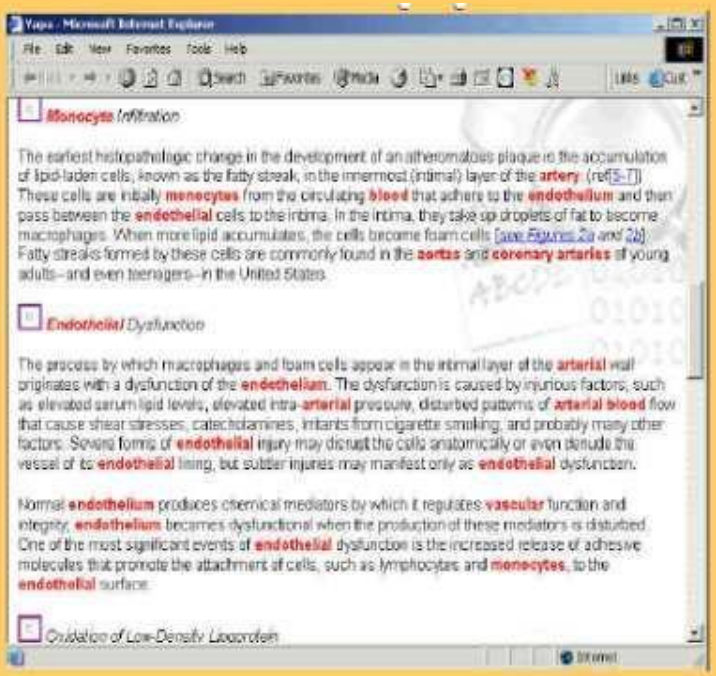
Кластеризация понятий: сервер SemioMap выявляет связи между извлеченными фразами и строит из них, на основе совместной встречаемости, лексическую сеть ("понятийную карту");

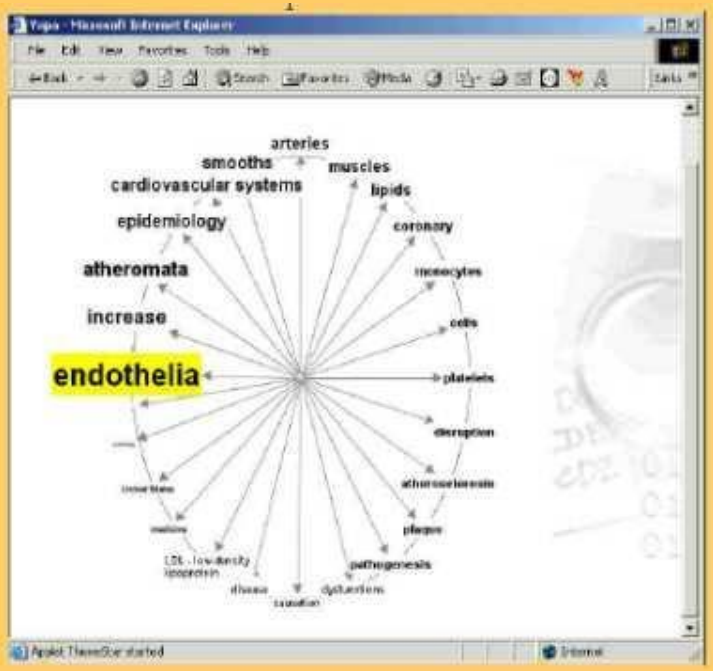
Графическое отображение и навигация: визуализация карт связей, которая обеспечивает быструю навигацию по ключевым фразам и связям между ними, а также возможность быстрого обращения к конкретным документам.

Oracle Text (Oracle)

(www.oracle.com/technology/products/text/)

Средства Text Mining, начиная с Text Server в составе СУБД Oracle 7.3.3 и картриджа interMedia Text в Oracle8i, являются неотъемлемой частью продуктов Oracle. В Oracle9i эти средства развились и получили новое название - Oracle Text.





Oracle Text (Oracle)

Основной задачей, на решение которой нацелены средства Oracle Text, является задача поиска документов по их содержанию - словам или фразам, которые при необходимости комбинируются с использованием булевых операций. Результаты поиска ранжируются по релевантности, с учетом частоты встречаемости слов запроса в найденных документах. Для повышения полноты поиска Oracle Text предоставляет ряд средств расширения поискового запроса, среди которых можно выделить: расширение слов запроса всеми морфологическими формами, расширение слов запроса близкими по смыслу словами за счет подключения тезауруса, а также расширение запроса словами, близкими по написанию и по звучанию - нечеткий поиск и поиск созвучных слов.

Система Oracle Text обеспечивает проведение тематического анализа текстов на английском языке. В ходе обработки текст каждого документа подвергается процедурам лингвистического и статистического анализа, в результате чего определяются его ключевые темы и строятся тематические резюме, а также общее резюме - реферат.

Knowledge Server (Autonomy)

http://www.autonomy.com/) Архитектура IDOL (Intelligent Data Operating Layer) сервера компании Autonomy, известной своими разработками в области статистического контент-анализа, объединяет интеллектуальный парсинг по шаблонам со сложными методами контекстного анализа и извлечения смысла для решения задач автоматической классификацию и организации перекрестных ссылок.

Knowledge Server (Autonomy)

Основное преимущество системы Autonomy - интеллектуальные алгоритмы, основанные на статистической обработке. Эти алгоритмы базируются на информационной теории Шеннона, Байесовых вероятностях и нейронных сетях.

Autonomy включает такие основные возможности:

- автоматическая классификация;
- кластеризация;
- автореферирование;
- автоматическое проставление гиперссылок;
- автоматическое создание профилей (информационных портретов);
- генерация таксонометрических деревьев;
- создание и манипулирование метаданными;
- интеллектуальная обработка XML-данных;
- персонализация;
- поиск.

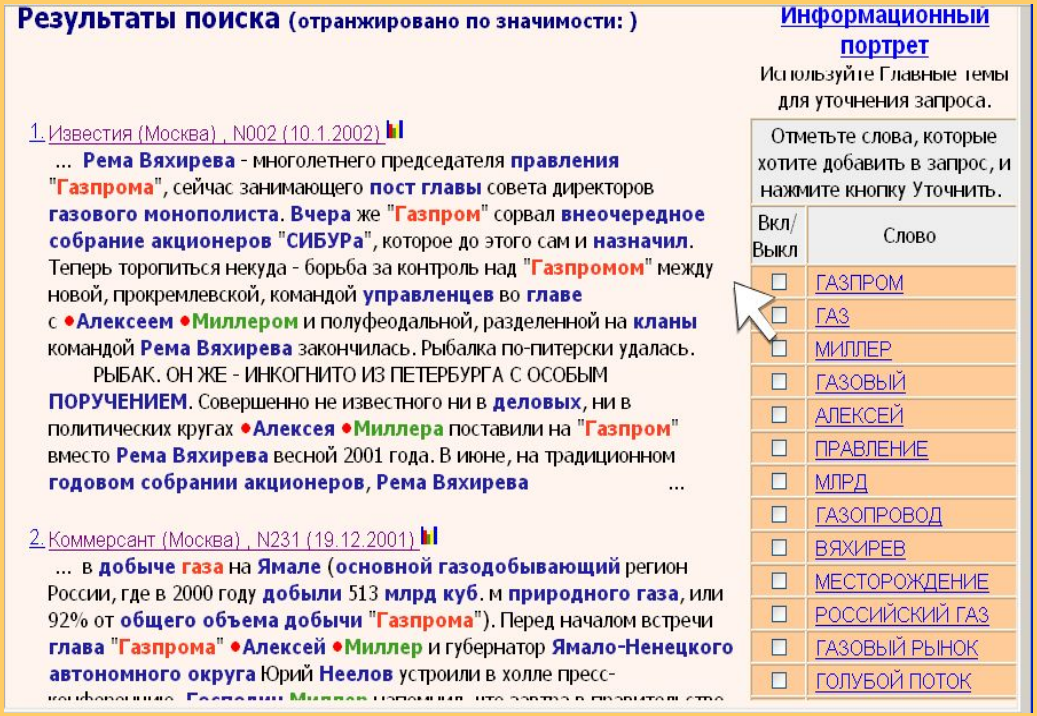
RetrievalWare (Convera)

(www.convera.com)

RetrievaWare - средство полнотекстового и атрибутивного поиска. К документам, с которыми способна работать система RetrievalWare, относятся тексты в различных форматах и кодировках в 200 форматах. Позиционируется как система добычи знаний (Knowledge Mining).

Galaktika-ZOOM ("Галактика")

(http://zoom.galaktika.ru/)
Основное назначение Galaktika-ZOOM - интеллектуальный поиск по ключевым словам с учетом морфологии, а также и формирование информационных портретов по конкретным аспектам. Ориентация на большие информационные объекты. Система содержит инструментарий для анализа смысловых связей и формирования "образа" проблемы - многомерной модели в форме списка значимых словосочетаний. Система содержит инструментарий для выявления тенденций и динамики развития проблем.

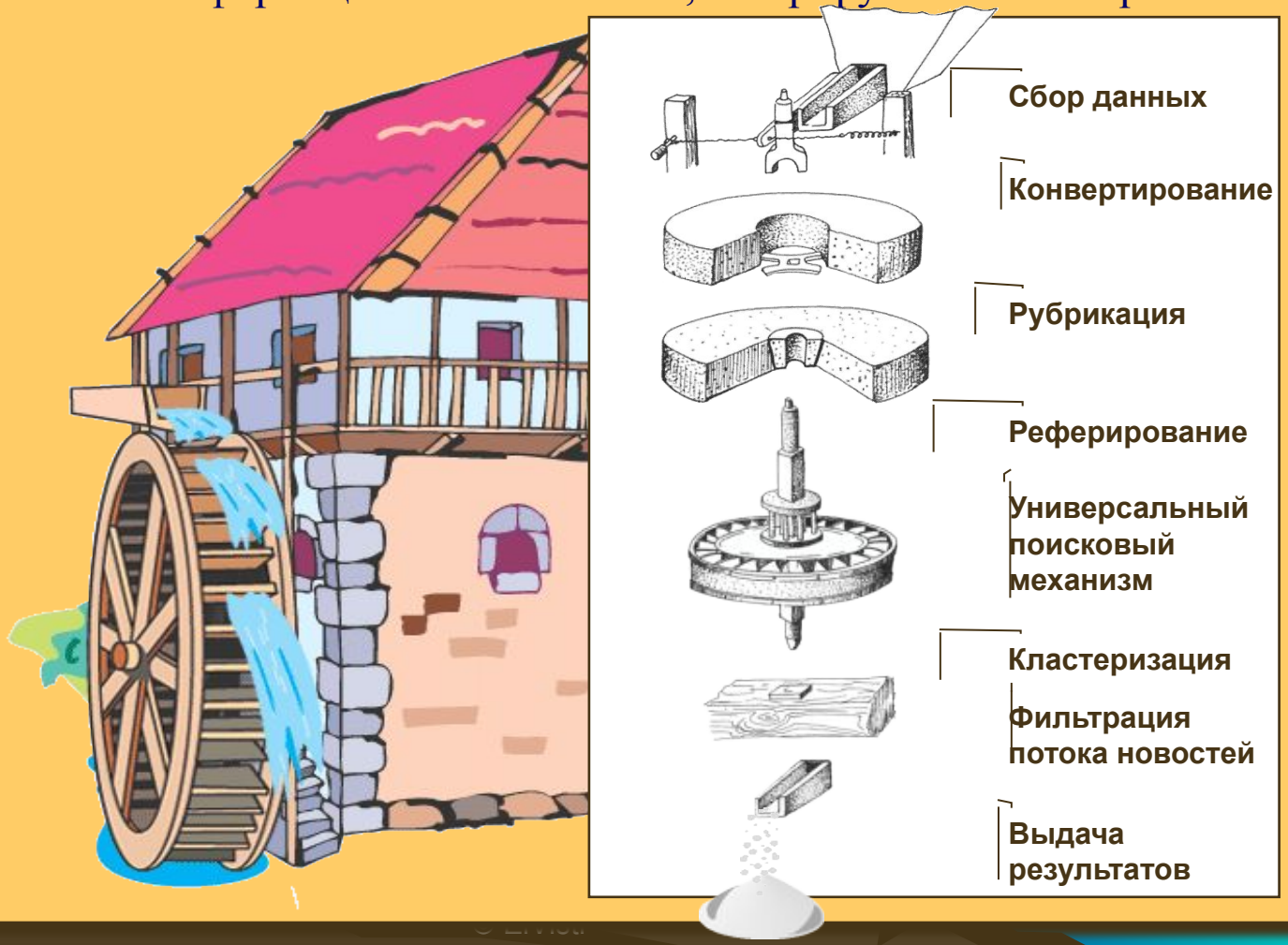


InfoStream (ИЦ "ЭЛВИСТИ")

ЗАВЖДИ У КУРСІ ВСІХ ПОДІЙ

(http://infostream.ua)

Система InfoStream создана для охвата и обобщения динамических новостных информационных массивов, генерируемых в Интернет.



InfoStream (ИЦ "ЭЛВИСТИ")

ЗАВЖДИ У КУРСІ ВСІХ ПОДІЙ

(http://infostream.ua)

Система *InfoStream* обеспечивает:

- ✔ Доступ к оперативной информации (более 2700 источников) с единого интерфейса в поисковом режиме с учетом возможного дублирования и семантической близости документов, языковых версий, размеров документов их цифровой насыщенности и т. д.
- ✔ Доступ к уникальному ретроспективному фонду, превышающему 30 млн. записей.
- ✔ Поддержку аналитической работы в режиме реального времени: построение сюжетных цепочек, дайджестов, диаграмм встречаемости и таблиц взаимосвязей понятий, медиа-рейтингов.

Спасибо за внимание!

Ландэ Д.В
dwl@visti.net
http://poiskbook.kiev.ua

МЕЖДУНАРОДНЫЙ СОЛОМОНОВ УНИВЕРСИТЕТ
Киев, Украина