А.Г. Шаров

Технологии работы с текстовой и фактографической информацией

Словарь – справочник

Москва 2007

УДК 004.912 (087.2)

Технологии работы с текстовой и фактографической информацией: Словарь–справочник / А.Г.Шаров. – М.: Препринт ООО «МегаВерсия», 2007. – 40 с.

Словарь–справочник содержит более 150 терминов и понятий, а также описание методов и средств разработки, применяемых в информационном поиске.
Может использоваться в качестве справочника по терминологии, существующим прикладным системам и средствам обработки текстовой и фактографической информации.
Для специалистов, занимающихся разработкой технологических решений и программных продуктов в области информационного поиска и извлечения фактографических данных из текстовой информации, а также в области семантического Веба.

В работе использованы сведения из информационных источников сети Интернет, находящихся в свободном доступе, и не требуют специального разрешения авторов либо администрации ресурсов на их публикацию в образовательных целях.

Оглавление

1. Информация, данные, знания 7
Данные 7
Знание 7
Информация 7
Концепт 7
Метаданные 7
Понятие 7
Сведения 7
Смысл 7
Управление знаниями 7
Факт 7
2. Базы данных, базы знаний 7
Атрибут 7
База данных 7
База знаний 8
Индексирование 8
Фактографическая база данных 8
Фактографическая информационно-поисковая система 8
Фактографический информационно-поисковый язык 8
Фактографическое индексирование 8
3. Лингвистика. Термины и определения 8
Актант 8
Антонимы 8
Валентность 8
Гипероним 8
Гипоним 8
Грамматика 8
Денотат 8
Дериват 9
Десигнат 9
Дескриптор 9
Дефиниция 9
Иерархический указатель информационно-поискового тезауруса 9
Информационно-поисковый тезаурус 9
Лексема 9
Лексика 9
Мероним 9
Омоним 9
Онтология 9
Парадигма 9
Перифраз 9
Полисемия 9
Семантика 9
Семиотика 10
Синонимы 10
Синтагма 10
Синтаксис 10
Словарь предметных рубрик 10
Слово 10
Словоизменение 10
Словообразование 10

Словосочетание 10
Словоформа 10
Таксономия 10
Тезаурус 10
Терминологический словарь 10
Термины 10
Холоним 11
Элементы онтологий 11
4. Информационный поиск 11
Запрос 11
Информационный поиск 11
Модель поиска 11
Объект запроса 11
Классическая задача ИП 11
Траверс 11
Центральная задача ИП 11
4.1 Методы и алгоритмы поиска 12
Булевская модель 12
Векторная модель 12
Вероятностная модель 12
Внетекстовые критерии 12
Входные страницы 12
Дубликаты 12
Иллюзия свежести 12
Информационный анализ 12
Информационный поиск 12
Контент-анализ 12
Клоакинг 12
Обратная связь 12
Поиск информации 13
Поиск по смыслу 13
Поиск похожих документов 13
Поисковая система 13
Поисковое предписание 13
Почти-дубликаты 13
Прюнинг 13
Прямой поиск 13
Регулярное выражение 13
Спам 13
Суффиксные деревья 13
Фильтрация 13
4.2 Классификация, рубрикация 13
Аннотация 13
Иерархическая классификация 13
Информационная классификационная система 14
Каталог 14
Каталогизация 14
Классификационная таблица 14
Классификационный индекс 14
Классификация 14
Кластер 14
Кластерный анализ 14

Предметная рубрика 14
Ранжирование 14
Реферат 14
Реферативная база данных 14
Рубрикатор 14
Фасетная классификация 14
4.3 Критерии оценки поиска 15
Валидность 15
Выпадение 15
Полнота 15
Пертинентность 15
Релевантность 16
Точность 16
Шум 16
F-мера 16
4.4 Статистическая обработка текстов 16
Инвертированный файл 16
Индекс цитирования 16
Обратная встречаемость в документах 16
Частота (слова) 16
Частота термина 17
PageRank 17
TF*IDF 17
4.5 Обработка лингвистических данных 17
Автоматизированное индексирование 17
Автоматическое индексирование 17
Дизамбигуация 17
Латентно-семантическое индексирование 17
Лемматизация 17
Нисходящий разбор 17
Основа 17
Парсер 17
Различительная сила слова 17
Рекурсивный нисходящий парсер 17
Свободное индексирование 17
Стемминг 18
Стоп-слова 18
Токенизация 18
5. Интернет и всемирная паутина 18
....18
5.1.1. RDF (Resource Description Framework) – язык описания ресурсов 18
5.1.2. OWL (Ontology Web Language) – язык Веб-онтологий 19
5.1.3. SPARQL- Query Language for RDF 19
5.2. Терминология HTTP в RDF 19
5.3. Перспективы формирования Семантической Сети 20
WordNet 21
Synset 21
5.4. Семантические сети 21
5.5. Нейронные сети 22
5.6. Методы и средства обработки 22
SVD (Singular-Value Decomposition) - сингулярное разложение 22
LSA (Latent Semantic Analyses) – латентно-семантический анализ 22

SVM (Support Vector Machines) – метод опорных векторов 22
Ссылки на документацию и примеры использования SVM 23
AJAX (Asynchronous JavaScript and XML) 24
XML (eXtensible Markup Language) 24
XSL (Extensible Stylesheet Language) 25
XSLT (Extensible Stylesheet Language Transformations) 25
SOAP (Simple Object Access Protocol) 25
VML (Vector Markup Language) 25
LZW (algorithm) 26
RSS (Really Simple Syndication) 26
6. Фактографический анализ 26
Атрибут 27
Гипотеза 27
Досье 27
Объект 28
Связь 28
Тип досье 28
Факт 28
6.1. Добыча данных 28
6.2. Извлечение и структурирование фактографической информации 28
6.3. Извлечение знаний из текстовой информации 29
6.4. Обработка фактографической информации – практика 32
7. Список инструментов Semantic Web 34
7.1. Среды разработки, редакторы, системы управления контентом 34
7.2. RDF репозитории 36
7.3. API 37
7.4. Механизмы логического вывода OWL 40
7.5. Генераторы RDF 40
7.6. On-line Валидаторы 40
7.7. Серверы запросов SPARQL 40

1. Информация, данные, знания

Данные (калька от лат. data) — это представление фактов и идей в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе

Знание - Совокупность сведений в какой-нибудь области.

Информация (от лат. informatio — разъяснение, научение, сведение, изложение, осведомленность) — сведения (данные), которые воспринимаются живым существом или устройством и сообщаются (получаются, передаются, преобразуются, сжимаются, разжимаются, теряются, находятся, регистрируются) с помощью знаков символьного, иконического, жестового или звукового типа. (Википедия)

Концепт в философии и лингвистике — содержание понятия, смысловое значение имени (знака). Отличается от самого знака и от его предметного значения (денотата, объёма понятия). Отождествляется с понятием и сингификатом.

Метаданные - данные о данных: каталоги, справочники, реестры, базы метаданных, содержащие сведения о составе данных, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.
Например, значение «123456» само по себе недостаточно выразительно. А если значению «123456» сопоставлено достаточно выразительное имя «почтовый индекс» (что уже является метаданными), то в этом контексте значение «123456» более осмысленно - можно извлечь информацию о местоположении адресата, имеющего данный почтовый индекс.

Понятие - это отражение в сознании людей общих и существенных признаков явлений действительности, представлений об их свойствах. Такими признаками могут быть форма предмета, его функция, цвет, размер, сходство или различие с другим предметом и т. д.
Понятия формируются и закрепляются в нашем сознании с помощью ***слов***. Связь слов с понятием (*сигнификативный фактор*) делает слово орудием человеческого мышления. Без способности слова называть понятие не было бы и самого языка.

Сведения – Факты, данные, характеризующие кого-л., что-л.

Смысл - в повседневной речи синоним значения.

Управление знаниями (Knowledge Management) — совокупность процессов и технологий, предназначенных для выявления, создания, распространения, обработки, хранения и предоставления для использования знаний.

Факт - в информатике - это единичное значение данных, созданное или использованное бизнес-процессом.

2. Базы данных, базы знаний

Атрибут (или данное) - это некоторый показатель, который характеризует некий объект и принимает для конкретного экземпляра объекта некоторое числовое, текстовое или иное значение.

База данных (БД) — централизованное хранилище данных, обеспечивающее хранение, доступ, первичную обработку и поиск информации.

База знаний (БЗ, англ. Knowledge base, KB) — это особого рода база данных, разработанная для управления знаниями (метаданными), то есть сбором, хранением, поиском и выдачей знаний.

Индексирование - в информационном поиске - процесс описания документов и запросов в терминах информационно-поискового языка. По результатам индексирования каждому документу назначается набор ключевых слов, отражающих его смысловое содержание.

Фактографическая база данных (Source databas)e - база данных, содержащая информацию, относящуюся непосредственно к предметной области. (Глоссарий.ру)

Фактографическая информационно-поисковая система (Factographic information retrieval system) - информационно-поисковая система, обеспечивающая выдачу непосредственно фактических сведений, затребованных потребителем в информационном запросе. Поисковый массив фактографической ИПС состоит из описаний фактов, извлеченных из документов и представленных на некотором формальном языке. (Глоссарий.ру)

Фактографический информационно-поисковый язык - информационно-поисковый язык, предназначенный для индексирования описаний фактов и информационного поиска в фактографических информационных массивах.

Фактографическое индексирование - индексирование, предусматривающее отражение в поисковом образе документа конкретных сведений (фактов).

3. Лингвистика. Термины и определения

Актант - любой член предложения, обозначающий лицо или предмет, участвующий в процессе, обозначенном глаголом.

Антонимы (от греч. Anti — против, Onyma имя) — это слова, противоположные по значению. Антонимия строится на противопоставлении соотносительных понятий: друг — враг, горький — сладкий, легко — трудно и др.

Валентность (от лат. *valentia* — сила) в общем случае — способность объекта взаимодействовать с другими объектами. Точное значение термина зависит от конкретной дисциплины, в которой он используется.
Валентность предиката в математике — то же, что его **арность**.
В математике **стью** предиката, операции или функции называется количество их аргументов, или операндов. Слово образовалось из названий предикатов небольшой арности (унарный — один аргумент, бинарный — два, тернарный — три). Также для этих целей употребляется термин *валентность* (в лингвистике употребляется только он). В общем случае предикат с *n* аргументами называют *n-арным*. Также употребляются термины **местность** и **вместимость**

Гипероним (родо-видовое отношение) - это объект, являющийся *над*видом (родом) другого. *Красный* выступает как гипероним *алому*, автомобиль — гипероним для «Москвича» и т. д. Если А — гипоним для Б, то Б — гипероним для А, то есть понятия *гипоним* и *гипероним* противоположны (антонимы).

Гипоним (родо-видовое отношение) - это объект, являющийся подвидом другого. Например, понятие алый — гипоним понятия красный, автомобиль — гипоним транспортного средства. Конкретизация обозначает, что мы накладываем на объект дополнительные ограничения.

Грамматика – Основные правила, исходные положения какой-либо области знания. (Словарь Ефремовой)

Денотат в лингвистике (в логике - экстенсионал) - обозначаемый предмет, содержание знака, указывающее на его предметную соотнесенность, т. е. множество объектов действительности (вещей,

свойств, отношений, ситуаций, действий и др.), которые могут именоваться данной языковой единицей. Например, денотат имени "Утренняя звезда" – планета Венера.

Дериват м. (лат. derivatus - отведенный) - Производное от чего-л. первичного; продукт чего-л.; производное слово (в лингвистике). Нитробензол есть дериват бензола. (Толковый словарь Ушакова).

Десигнат - предметное значение, - в логике и семантике предмет, обозначаемый собственным именем некоторого языка (в формализованном языке - константой или термом), или класс предметов, обозначаемых общим (нарицательным) именем (в формализованном языке - предметной переменной).

Дескриптор - (лат. descriptor - описывающий) - лексическая единица (слово, словосочетание) информационно-поискового языка, выражающая основное смысловое содержание какого-либо текста. Используется при информационном поиске документов в информационно-поисковых системах. (Энциклопедический словарь http://slovar.plib.ru/dictionary/d5/18089.html). Ключевое слово, характеризующее блок информации. (Бизнес-словарь http://slovar.plib.ru/dictionary/d6/3498.html)

Дефиниция – Определение, истолкование понятия (словарь Ожегова). (*Определение, приведенное в данной статье, есть **дефиниция** термина* **дефиниция**).

Иерархический указатель информационно-поискового тезауруса - список дескрипторов высшего уровня иерархии, в котором для каждого из дескрипторов приводятся подчиненные нижестоящие дескрипторы, расположенные в порядке убывания общности.

Информационно-поисковый тезаурус - словарь дескрипторного информационно-поискового языка с зафиксированными в нем парадигматическими отношениями лексических единиц.

Лексема (от греч. lexis — слово, выражение, оборот речи) - в лингвистике - слово как самостоятельная единица языка, рассматриваемая во всей совокупности своих форм и значений. В одну лексему объединяются разные парадигматические формы (словоформы) одного слова (например, «словарь, словарём, словарю» и т. п.).

Лексика — раздел науки о языке, изучающий значения слов. Также под этим словом понимают совокупность слов того или иного языка, части языка или слов которые знает тот или иной человек или группа людей.

Мероним (отношение часть-целое) - это объект, являющийся частью для другого. Двигатель - это мероним для автомобиля.

Омоним – Слово, характеризующееся одинаковым написанием с другим словом, но отличающееся от него по значению (в лингвистике).

Онтология — целостная структурная спецификация некоторой предметной области, ее формализованное представление, которое включает словарь (или имена) указателей на термины предметной области и логические выражения, описывающие, как они соотносятся друг с другом.

Парадигма - Система форм изменяющегося слова.

Перифраз (перифраза) — Описательное выражение, заменяющее прямое название и содержащее в себе признаки не названного прямо предмета.

Полисемия (polysemy, homography, многозначность, омография, омонимия) - наличие нескольких значений у одного и того же слова

Семантика - раздел языкознания и логики, исследующий проблемы, связанные со смыслом, значением и интерпретацией лексических единиц.

Семиотика – теория знаковых систем, общая теория языков – естественных и искусственных.

Синонимы (от греч. Synonimos - одноименность) — слова, принадлежащие к одной и той же части речи, которые звучат и пишутся по-разному, а по смыслу тождественны или очень близки, например: Миг — момент (существительные); Бранить —ругать (глаголы); Огромный — громадный (прилагательные); Напрасно — зря (наречия); Возле — около (предлоги).

Синтагма - единица речи, линейная единица, которая возникает как результат естественного членения потока речи. Синтагма состоит из слов (иногда одного слова), но в то же время она не совпадает и не равняется в синтаксическом аспекте словосочетанию.

Синтаксис - набор правил построения фраз алгоритмического языка, позволяющий определить, осмысленные предложения в этом языке.

Словарь предметных рубрик - совокупность предметных рубрик и связанного с ними ссылочно-справочного аппарата предметного каталога или указателя.

Слово является важнейшей номинативной единицей языка.

Словоизменение (inflection) – образование формы определенного грамматического значения, обычно обязательного в данном грамматическом контексте, принадлежащей к фиксированному набору форм (парадигме), характерного для слов данного типа. В отличие от словообразования никогда не приводит к смене типа и порождает предсказуемое значение. Словоизменение имен называют склонением (declension), а глаголов – спряжением (conjugation)

Словообразование (derivation) – образование слова или основы из другого слова или основы. Чаще приводит к смене типа и к образованию слов, имеющих идеосинкразическое значение

Словосочетание есть последовательность заданного числа словоформ в пределах одной синтагмы. Например, 3-мерное словосочетание "дорога в ад" в предложении "Благими намерениями вымощена дорога в ад". (http://www.rvb.ru/soft/wt/wt.htm#def_1)

Словоформа (слово) есть последовательность букв, возможно, с включением дефиса (-) или апострофа (') между знаками препинания или пробелами. Словоформа всегда начинается с буквы. Например: "state-of-art", "из-за", "О'Генри". Состав символов, разрешенных в словоформе, определяется заданной кодировкой и может быть расширен путем задания множества дополнительных (не буквенных) символов, например, цифр. (http://www.rvb.ru/soft/wt/wt.htm#def_1)

Таксономия (Taxonomy) - в лингвистике - метод исследования языка, основанный на систематизации языковых факторов путем вычисления в тексте лингвистических единиц и изучении их свойств в зависимости от их порядка и распределения.

Тезаурус (гр. Thesauros запас) - в информатике - полный систематизированный набор данных о какой-либо области знаний, позволяющий человеку или вычислительной машине в ней ориентироваться. (Современный словарь иностранных слов, С.-Петербург 1994).
Тезаурус содержит список ключевых слов, которыми может быть охарактеризовано содержание документов, с выделением слов, рекомендованных для индексирования (дескрипторов).

Терминологический словарь - словарь, содержащий термины определенной области знания и их определения (разъяснения).

Термины - слова или словосочетания, называющие специальные понятия какой-либо сферы производства, науки, искусства. В основе каждого термина обязательно лежит определение (дефиниция) обозначаемой им реалии, благодаря чему термины представляют собой точную и в то же время сжатую характеристику предмета или явления. Каждая отрасль знания оперирует своими терминами, составляющими суть терминологической системы данной науки.

Холоним (отношение часть-целое) - это объект, который включает в себя другое. Например, у дома есть крыша. *Дом* — холоним для *крыши*. Компьютер — холоним для монитора. Мероним и холоним — противоположные понятия.

Элементы онтологий - современные онтологии строятся по большей части одинаково, независимо от языка написания. Обычно они состоят из экземпляров, понятий, атрибутов и отношений.

4. Информационный поиск

Запрос — это формализованный способ выражения информационных потребностей пользователем системы. Для выражения информационной потребности используется язык поисковых запросов, синтаксис варьируется от системы к системе. Кроме специального языка запросов, современные поисковые системы позволяют вводить запрос на естественном языке.

Информационный поиск (ИП) (английский термин Information retrieval) — наука о поиске неструктурированной документальной информации. В частности это относится к поиску информации в документах, поиск самих документов, извлечению метаданных из документов, поиску текста, изображений, видео и звука в локальных реляционных базах данных, в гипертекстовых базах данных таких, как Интернет и локальные интранет-системы.

Модель поиска – это некоторое упрощение реальности, на основании которого получается формула (сама по себе никому не нужная), позволяющая программе принять решение: какой документ считать найденным и как его ранжировать. После принятия модели коэффициенты часто приобретают физический смысл и становятся понятней самому разработчику, да и подбирать их становится интересней.

Объект запроса — это информационная сущность, которая хранится в базе автоматизированной системы поиска. Несмотря на то, что наиболее распространенным объектом запроса является текстовый документ, не существует никаких принципиальных ограничений. В частности, возможен поиск изображений, музыки и другой мультимедиа информации. Процесс занесения объектов поиска в ИПС называется индексацией. Далеко не всегда ИПС хранит точную копию объекта, нередко вместо неё хранится суррогат.

Классическая задача ИП, с которой началось развитие этой области, — это поиск документов, удовлетворяющих запросу, в рамках некоторой статической коллекции документов. Но список задач ИП постоянно расширяется и теперь включает:

- Вопросы моделирования;
- Классификация документов;
- Фильтрация документов;
- Кластеризация документов;
- Проектирование архитектур поисковых систем и пользовательских интерфейсов
- Извлечение информации, в частности аннотирования и реферирования документов;
- Языки запросов и др.

Траверс – это составление запроса из терминов, лежащих между двумя заданными в сети тезауруса.

Центральная задача ИП — помочь пользователю удовлетворить его информационную потребность. Так как описать информационные потребности пользователя технически непросто, они формулируются как некоторый запрос, представляющий из себя набор ключевых слов, характеризующий то, что ищет пользователь.

4.1 Методы и алгоритмы поиска

Булевская модель (boolean, булева, булевая, двоичная) – модель поиска, опирающаяся на операции пересечения, объединения и вычитания множеств

Векторная модель – модель информационного поиска, рассматривающая документы и запросы как векторы в пространстве слов, а релевантность как расстояние между ними

Вероятностная модель – модель информационного поиска, рассматривающая релевантность как вероятность соответствия данного документа запросу на основании вероятностей соответствия слов данного документа идеальному ответу

Внетекстовые критерии (off-page, вне-страничные) – критерии ранжирования документов в поисковых системах, учитывающие факторы, не содержащиеся в тексте самого документа и не извлекаемые оттуда никаким образом

Входные страницы (doorways, hallways) – страницы, созданные для искусственного повышения ранга в поисковых системах (поискового спама). При попадании на них пользователя перенаправляют на целевую страницу

Дубликаты (duplicates) – разные документы с идентичным, с точки зрения пользователя, содержанием; приблизительные дубликаты (near duplicates, почти-дубликаты), в отличие от точных дубликатов, содержат незначительные отличия

Иллюзия свежести – эффект кажущейся свежести, достигаемый поисковыми системами в интернете за счет более регулярного обхода тех документов, которые чаще находятся пользователями

Информационный анализ - выявление в документах и фиксация в виде данных информации, относящейся к определенной предметной области.

Информационный поиск (Information Retrieval, IR) – поиск неструктурированной информации, единицей представления которой является документ произвольных форматов. Предметом поиска выступает информационная потребность пользователя, неформально выраженная в поисковом запросе. И критерий поиска, и его результаты недетерминированы. Этими признаками информационный поиск отличается от «поиска данных», который оперирует набором формально заданных предикатов, имеет дело со структурированной информацией и чей результат всегда детерминирован. Теория информационного поиска изучает все составляющие процесса поиска, а именно, предварительную обработку текста (индексирование), обработку и исполнение запроса, ранжирование, пользовательский интерфейс и обратную связь.

Контент-анализ - количественный анализ книг, эссе, интервью, дискуссий, газетных статей, исторических документов и других текстов и текстовых массивов с целью последующей содержательной интерпретации выявленных числовых закономерностей.

Клоакинг (cloaking) – техника поискового спама, состоящая в распознании авторами документов робота (индексирующего агента) поисковой системы и генерации для него специального содержания, принципиально отличающегося от содержания, выдаваемого пользователю

Обратная связь – отклик пользователей на результат поиска, их суждения о релевантности найденных документов, зафиксированные поисковой системой и использующиеся, например, для итеративной модификации запроса. Следует отличать от псевдо-обратной связи – техники модификации запроса, в которой несколько первых найденных документов автоматически считаются релевантными

Поиск информации (часто то же, что и информационный поиск) — процесс выявления в некотором множестве документов (текстов) всех таких, которые посвящены указанной теме (предмету), удовлетворяют заранее определенному условию поиска (запросу) или содержат необходимые (соответствующие информационной потребности) факты, сведения, данные.
Процесс поиска включает последовательность операций, направленных на сбор, обработку и предоставление необходимой информации заинтересованным лицам.

Поиск по смыслу – алгоритм информационного поиска, способный находить документы, не содержащие слов запроса

Поиск похожих документов (similar document search) – задача информационного поиска, в которой в качестве запроса выступает сам документ и необходимо найти документы, максимально напоминающие данный/

Поисковая система (search engine, SE, информационно-поисковая система, ИПС, поисковая машина, машина поиска, «поисковик», «искалка») – программа, предназначенная для поиска информации, обычно текстовых документов

Поисковое предписание (query, запрос) – обычно строчка текста

Почти-дубликаты (near-duplicates, приблизительные дубликаты) – см. дубликаты

Прюнинг (pruning) – отсечение заведомо нерелевантных документов при поиске с целью ускорения выполнения запроса

Прямой поиск – поиск непосредственно по тексту документов, без предварительной обработки (без индексирования)

Регулярное выражение (regualr expression, pattern, «шаблон», реже «трафарет», «маска») – способ записи поискового предписания, позволяющий определять пожелания к искомому слову, его возможные написания, ошибки и т.д. В широком смысле – язык, позволяющий задавать запросы неограниченной сложности

Спам поисковых систем (spam, спамдексинг, накрутка поисковых систем) – попытка воздействовать на результат информационного поиска со стороны авторов документов

Суффиксные деревья, суффиксные массивы (suffix trees, suffix arrays, PAT-arrays) – индекс, основанный на представлении всех значимых суффиксов текста в структуре данных, известной как бор (trie). Суффиксом в этом индексе называю любую «подстроку», начинающуюся с некоторой позиции текста (текст рассматривается как одна непрерывная строка) и продолжающуюся до его конца. В реальных приложениях длина суффиксов ограничена, а индексируются только значимые позиции – например, начала слов. Этот индекс позволяет выполнять более сложные запросы, чем индекс, построенный на инвертированных файлах

Фильтрация - анализ близости - присвоение элементу растра нового значения как некоторой функции значений окрестных элементов.

4.2 Классификация, рубрикация

Аннотация - краткая характеристика документа, его части или совокупности документов с точки зрения содержания, назначения, формы и других особенностей. Аннотация носит пояснительный или рекомендательный характер.

Иерархическая классификация - классификационная система, в которой отношения классов образуют иерархическую классификационную структуру.

Информационная классификационная система - средство формализованного представления содержания документов, данных и информационных запросов посредством кодов или описаний классов логически упорядоченного множества понятий.
Информационные классификационные системы являются одним из типов информационно-поисковых языков.

Каталог - в широком смысле - список элементов данных, файлов, серверов, принтеров, магнитных накопителей и других объектов, составленный в порядке, облегчающем их нахождение. Каталоги упорядочиваются по алфавиту, датам, размеру содержащихся в них объектов и другим признакам.

Каталогизация - совокупность процессов, обеспечивающих создание и функционирование библиотечных каталогов. В состав каталогизации входят:
- библиографическая обработка;
- ввод данных или тиражирование каталожных карточек;
- работа с каталогами: организация, ведение и редактирование каталогов.

Классификационная таблица - материальное представление классификационной системы.

Классификационный индекс - поисковый образ, построенный средствами классификационного информационно-поискового языка.

Классификация - в информационном поиске - процесс распределения документов по категориям.

Кластер - класс родственных элементов статистической совокупности.

Кластерный анализ - метод группировки экспериментальных данных в классы. Наблюдения, попавшие в один класс, в некотором смысле ближе друг к другу, чем к наблюдениям из других классов.

Предметная рубрика - элемент информационно-поискового языка, представляющий собой краткую формулировку темы на естественном языке.

Ранжирование - процесс, при котором поисковая система:
- принимает запрос пользователя;
- находит все подходящие веб-страницы; и
- выстраивает их в определенном порядке по принципу наибольшего соответствия конкретному запросу.
Выведение рейтинга зависит от алгоритма ранжирования, которым пользуется поисковая машина.

Реферат - краткое изложение содержания отдельного документа, его части или совокупности документов, включающее основные сведения и выводы, а также количественные и качественные данные об объектах описания.

Реферативная база данных - библиографическая база данных, содержащая библиографические записи, включающие аннотацию, реферат или иные указания о содержании документа.

Рубрикатор - классификационная таблица иерархической классификации, содержащая полный перечень включенных в систему классов и предназначенная для систематизации информационных фондов, массивов и изданий, а также для поиска в них.

Фасетная классификация - классификационная система, в которой:
- понятия представлены в виде ***фасетной структуры***; а
- классификационные индексы синтезируются посредством комбинирования фасетных признаков в соответствии с фасетной формулой.
Теория построения разработана индийским учёным и библиотековедом Ш. Р. Ранганатаном («Классификация двоеточием», 1933). Также известна как: классификация двоеточием, классификация

Ранганатана. ***Основой классификации*** является привычное человеку отнесение объекта к разным категориям. Примером может являться классификация фильмов:

- ***тип***: документальный, игровой, анимация
- ***жанр***: боевик, комедия, романтика, фантастика
- ***продолжительность***
- ***год***
- ***страна***
- ***автор***
- ***другие параметры***: немой/звуковой, цветной/чёрно-белый и т. п.

Таким образом, каждый фильм обладает множеством признаков. При поиске нужного фильма используется пересечение требуемых атрибутов.

4.3 Критерии оценки поиска

Валидность, Обоснованность (От лат.Validus – сильный) - степень соответствия показателя тому понятию, которое он призван отражать.

Выпадение характеризует вероятность нахождения нерелевантного ресурса и определяется, как отношение числа найденных нерелевантных документов к общему числу нерелевантных документов в базе:

$$\text{Fall-out} = \frac{|D_{nrel} \cap D_{retr}|}{|D_{nrel}|},$$

где D_{nrel} — это множество **не**релевантных документов в базе, а D_{retr} — множество документов, найденных системой.

Полнота (recall, охват) – доля релевантного материала, заключенного в ответе поисковой системы, по отношению ко всему релевантному материалу в коллекции

Отношение числа **найденных** релевантных документов, к общему числу релевантных документов в базе:

$$\text{Recall} = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|},$$

где D_{rel} — это множество релевантных документов в базе, а D_{retr} — множество документов, найденных системой.

Пертинентность

Пертинентный документ - относящийся к делу, подходящий по сути.
Оценить пертинентность документа можно в сравнении с другими документами.
пертинетность, соотношение объема полезной для него информации к общему объему полученной информации, имеет решающее значение. При этом следует учитывать, что формальный запрос к системе является предметом творческого осмысления информационной потребности и не всегда точно отражает последнюю. Неумение большинством пользователей правильно формулировать запросы и получать приемлемые объемы отклика породило в конце 20 века мнение об Интернет, как об огромной информационной свалке. Достижение высокой пертинентности - основное поле конкурентной борьбы современных поисковых систем. Именно для максимального удовлетворения информационных потребностей пользователей информационно-поисковые системы сегодня максимально интеллектуализируются - получили широкое практическое применение теории и методы семантических сетей, контент-анализа и глубинного анализа текстов (Text Mining).

Релевантность (relevance, relevancy) – соответствие документа запросу. Под релевантностью понимается формальное соответствие информации, выдаваемой системой, запросу. Если по запросу пользователя получено N документов, представляющих собой объединение двух множества документов: соответствующих запросу (пусть их количество - N1), и не соответствующих (их количество - N2), т.е. N = N1+N2. Тогда релевантность, как степень соответствия, определяется по формуле: P = (N1/N)x100%, а **шум** - по формуле: S = (N2/N)x100% = 100% - P. Это определение характерно для формальной релевантности, однако, на практике используется другое, неформальное понятие - **пертинентность**.

Точность (precision) - доля релевантного материала в ответе поисковой системы

Отношение числа релевантных документов, найденных ИПС, к общему числу документов найденных ИПС:

$$\mathrm{Precision} = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|},$$

где D_{rel} — это множество релевантных документов в базе, а D_{retr} — множество документов, найденных системой.

Шум (процент мусора) – показатель, характеризующий количество ненужной пользователю информации, полученной по запросу.

F-мера (F-measure) - Традиционно определяется, как гармоническое среднее точности и полноты:

$$F = 2 \times \mathrm{Precision} \times \mathrm{Recall} / (\mathrm{Precision} + \mathrm{Recall}).$$

Часто ее также называют F_1 мерой, потому что точность и полнота присутствуют в этой формуле с одинаковым весом.

Более общая формула для положительного вещественного α имеет вид:

$$F_\alpha = (1 + \alpha) \times \mathrm{Precision} \times \mathrm{Recall} / (\alpha \times \mathrm{Precision} + \mathrm{Recall}).$$

4.4 Статистическая обработка текстов

Инвертированный файл (inverted file, инверсный файл, инвертированный индекс, инвертированный список) – индекс поисковой системы, в котором перечислены слова коллекции документов, а для каждого слова перечислены все места, в которых оно встретилось

Индекс цитирования (citation index) – число упоминаний (цитирований) научной статьи, в традиционной библиографической науке рассчитывается за промежуток времени, например, за год

Обратная встречаемость в документах (inverted document frequency, IDF, обратная частота в документах, обратная документная частота) – показатель поисковой ценности слова (его различительной силы); обратная говорят, потому что при вычислении этого показателя в знаменателе дроби обычно стоит число документов, содержащих данное слово

Частота (слова) в документах (document frequency, встречаемость в документах, документная частота) – число документов в коллекции, содержащих данное слово

Частота термина (term frequency, TF) – частота употреблений слова в документе

PageRank (статистическая популярность) – алгоритм расчета статической (глобальной) популярности страницы в интернете, назван в честь одного из авторов - Лоуренса Пейджа. Соответствует вероятности попадания пользователя на страницу в модели случайного блуждания

TF*IDF – численная мера соответствия слова и документа в векторной модели; тем больше, чем относительно чаще слово встретилось в документе и относительно реже в коллекции

4.5 Обработка лингвистических данных

Автоматизированное индексирование - индексирование, технология которого предусматривает использование формальных процедур, осуществляемых с помощью вычислительной техники, и включает применение интеллектуальных процедур при принятии основных решений о составе поискового образа.

Автоматическое индексирование - индексирование, технология которого предусматривает использование только формальных процедур обработки текста, осуществляемых с помощью вычислительной техники.

Дизамбигуация (tagging, part of speech disambiguation, **таггинг**) – выбор одного из нескольких омонимов с помощью контекста; в английском языке часто сводится к автоматическому назначению грамматической категории «часть речи»

Латентно-семантическое индексирование – запатентованный алгоритм поиска по смыслу, идентичный факторному анализу. Основан на сингулярном разложении матрицы связи слов с документами

Лемматизация (lemmatization, нормализация) – приведение формы слова к словарному виду, то есть лемме

Нисходящий разбор - класс алгоритмов грамматического анализа, где правила формальной грамматики раскрываются, начиная со стартового символа, до получения требуемой последовательности токенов.

Основа – часть слова, общая для набора его словообразовательных и словоизменительных (чаще) форм

Парсер – 1. Синтаксический анализатор
2. Конкретно Грамматический анализатор в составе синтаксического анализатора
3. Язык программирования разработанный студией Артемия Лебедева (Википедия)

Различительная сила слова (term specificity, term discriminating power, контрастность, различительная сила) – степень ширины или узости слова. Слишком широкие термины в поиске приносят слишком много информации, при это существенная часть ее бесполезна. Слишком узкие термины помогают найти слишком мало документов, хотя и более точных.

Рекурсивный нисходящий парсер (en:Recursive descent parser) - алгоритм грамматического разбора, реализуемый путем взаимного вызова парсящих процедур, соответствующих правилам контекстно-свободной грамматики или БНФ. Применения правил последовательно, слева-направо поглощают токены, полученные от лексического анализатора. Это один из самых простых алгоритмов парсинга, подходящий для полностью ручной реализации.

Свободное индексирование - индексирование, технология которого не предусматривает замену ключевых слов текста в соответствии с рекомендациями специального словаря.

При свободном индексировании из текста извлекаются ключевые слова без учета всех видоизменений их форм и отношений между ними.

Стемминг – процесс выделения основы слова

Стоп-слова (stop-words) – те союзы, предлоги и другие частотные слова, которые данная поисковая система исключила из процесса индексирования и поиска для повышения своей производительности и/или точности поиска

Токенизация (tokenization, lexical analysis, графематический анализ, лексический анализ) – выделение в тексте слов, чисел, и иных токенов, в том числе, например, нахождение границ предложений

5. Интернет и всемирная паутина

Интернет (от англ. Internet) — всемирная система объединённых компьютерных сетей, построенная на использовании протокола IP и маршрутизации пакетов данных. Интернет образует глобальное информационное пространство, служит физической основой для Всемирной паутины и множества систем (протоколов) передачи данных. Часто
. (Википедия)
Когда сейчас слово Интернет употребляется в обиходе, то чаще всего имеется в виду Всемирная паутина и доступная через неё информация, а не сама физическая сеть.
В английском языке, когда слово internet написано со строчной буквы, оно означает просто объединение сетей (англ. interconnected networks) посредством маршрутизации пакетов данных. В этом случае не имеется в виду глобальное информационное пространство. В русском языке такое разделение понятий иногда встречается в технической литературе.

(англ. World Wide Web) — глобальное информационное пространство, основанное на физической инфраструктуре Интернета и протоколе передачи данных HTTP. Всемирная паутина вызвала настоящую революцию в информационных технологиях и бум в развитии Интернета. Часто, говоря об Интернете, имеют в виду именно Всемирную паутину. Для обозначения Всемирной паутины также используют слово веб (англ. web) и аббревиатуру «WWW».

http://ru.wikipedia.org/wiki/Семантическая_паутина

5.1. (англ. *Semantic web*) — новая концепция развития Всемирной паутины и сети Интернет, принятая и продвигаемая Консорциумом Всемирной паутины. Иногда также упоминается как **семантический веб**. В основу Семантической Сети положены две ключевые технологии: RDF и OWL.
Семантическая паутина — это надстройка над существующей Всемирной паутиной, которая призвана сделать размещённую в сети информацию более понятной для компьютеров. Известно, что почти вся информация в Интернете находится в текстовой форме. Не секрет также, что прогресс в области обработки человеческих языков (англ. *Natural Language Processing*) идёт очень медленно. Компьютеры не могут воспринять и осмыслить словесную информацию, размещённую в Интернете, и в ближайшее время, видимо, не смогут. Тогда встаёт вопрос — как же заставить компьютеры понимать смысл размещённой в сети информации и научить компьютеры пользоваться ею? На этот вопрос и призвана ответить концепция семантической паутины. Слово «семантическая» в данном случае означает «осмысленная», «понятная».

5.1.1. RDF (Resource Description Framework) – язык описания ресурсов

http://ru.wikipedia.org/wiki/RDF
Resource Description Framework — это разработанная консорциумом W3C модель для описания ресурсов, в особенности — метаданных о ресурсах. В основе этой модели лежит идея об использовании специального вида утверждений, высказываемых о ресурсе. Каждое утверждение имеет вид «субъект — предикат — объект» и в терминологии RDF называется триплетом. Например, утверждение «Небо голубого цвета» в RDF-терминологии можно представить следующим образом: субъект — «небо», предикат — «имеет цвет», объект — «голубой». Для идентификации субъектов, предикатов и объектов в RDF используются **URI** (англ. ***Uniform Resource Identifier***).

Одной из главных целей RDF является предоставление утверждений одинаково в машинно- и человеко-распознаваемом виде.
Существует несколько синтаксисов для представления RDF-информации, самые распространённые из которых: RDF/XML, триплеты (Нотация 3) и графовая модель.
Для обработки RDF имеется несколько языков запросов: например, RQL, RDQL, SPARQL[1]
Описание языка: http://www.w3.org/RDF/

http://ru.wikipedia.org/wiki/Семантическая_паутина
RDF — это система описания сетевых ресурсов, понятная компьютеру. Формат RDF предназначен для хранения метаданных *(метаданные — это данные о данных)*. В соответствии с концепцией семантической паутины, описания в формате RDF должны прикрепляться к каждому сетевому ресурсу. Документы RDF должны обрабатываться компьютером автоматически, RDF не предназначен для прочтения и использования человеком. К настоящему времени формат RDF уже устоялся и получил широкое распространение, он служит каркасом для создания семантической паутины.

RDFS (англ. *RDF Schema*) — это важная надстройка над RDF, позволяющая создавать классы и свойства (как в объектно-ориентированном программировании в рамках конкретного приложения).

5.1.2. OWL (Ontology Web Language) – язык Веб-онтологий

http://sherdim.rsu.ru/pts/semantic_web/REC-owl-guide-20040210_ru.html
Следующим важным направлением концепции семантической паутины является язык OWL (англ. *Web Ontology Language*, произносится [ ]), который стал Рекомендацией W3C в феврале 2004 года. Этот язык построен на форматах RDF и RDFS, он предназначен для обработки информации в сети. Язык OWL имеет 3 степени детализации, что является новым словом в компьютерных технологиях. Он также легко масштабируется и совместим с самыми передовыми сетевыми стандартами.

http://ru.wikipedia.org/wiki/OWL
OWL — (англ. *Web Ontology Language*) — язык онтологии для интернета на основе XML/Web стандарта. Язык веб-онтологий OWL призван обеспечить язык, который может быть использован для описания классов и отношений между ними, которые присущи для веб-документов и приложений.
В основе языка — представление действительности в модели данных объект — свойство. По признанию создателей, owl пригоден не только для описания web страниц, но и любых объектов действительности. Каждому элементу описания в этом языке ставится в соответствие www адрес, но во всем остальном — это чудесное описание механизма reference системы на базе модели объект — свойство.
http://sherdim.rsu.ru/pts/semantic_web/REC-owl-guide-20040210_ru.html

5.1.3. SPARQL- Query Language for RDF

http://www.w3.org/TR/2007/WD-rdf-sparql-query-20070326/
SPARQL (рекурсивный акроним, SPARQL Protocol and RDF Query Language) — разрабатывающийся стандарт семантической паутины, сейчас (2006) проходящий стандартизацию RDF Data Access Working Group (DAWG) консорциума World Wide Web (W3C). Несколько реализаций для различных языков программирования уже существуют.

http://ru.wikipedia.org/wiki/Семантическая_паутина
SPARQL (англ. *Protocol And RDF Query Language*, произносится [ ]) — новый язык запросов для быстрого доступа к данным RDF. Используя обычный протокол и язык SPARQL, приложения могут анализировать RDF-описания ресурсов и получать из сети нужную информацию.

5.2. Терминология HTTP в RDF

http://www.semantictools.ru/posting/post-6.shtml
Рабочая Группа Инструментов Оценки и Восстановления консорциума W3C опубликовала обновленный черновой вариант (Working Draft) "Терминологии HTTP в RDF". Используя эти термины, заголовки HTTP, которыми обмениваются клиенты и серверы, могут быть записаны в формате RDF. Терминология включает термины для схем HTTPS, также как и термины для других расширений базовой спецификации.
Идентификация ресурсов в Сети по URI может оказаться недостаточной, для однозначного определения документа, так как другие факторы, такие как содержание HTTP сообщений начинают

играть роль. Эта ситуация оказывается особенно важной для тестирования, для требований о соответствии, и для языков отчетов, таких как Язык Оценки и Восстановления (EARL). "Терминология HTTP в RDF" описывает представление терминологии HTTP в RDF. Этот документ определяет коллекцию RDF классов и свойств для представления терминов HTTP, как они определены в спецификации HTTP. Эти RDF термины могут быть использованы для записи запросов и ответов HTTP в формате RDF.
Существует несколько возможных применений для этой технологии:

Сообщение о результатах тестов

Когда проводится тестирование веб-ресурса, может быть очень полезно сохранять точные копии заголовков сообщений, которыми обмениваются клиент и сервер. Языки отчетов RDF (такие как EARL) могут использовать эту терминологию, чтобы сохранять заголовки такого рода. Например: параметры POST, которые были посланы через форму и использованы сервером для генерации ответа.

Уточнение требований о соответствии

Требования о соответствии, например о соответствии WCAG иногда применимы только к определенной версии веб-ресурса расположенного по указанному URI. Например: документ представленный в нескольких языках. Заголовки HTTP в RDF позволяют точно указать необходимую версию веб-ресурса.

Разработка Веб-приложений

Инструментальные средства веб, например: инструменты для разработки сложных приложений с динамическим содержанием, могут использовать эту технологию, чтобы помочь разработчикам отлаживать скрипты и приложения. Например, при разработке AJAX приложений, когда веб-ресурс представляет собой совокупность небольших порций контента.

5.3. Перспективы формирования Семантической Сети

http://www.semantictools.ru/conception/two_approach.shtml

Существует два возможных способа формирования Семантической Сети: снизу вверх и сверху вниз.
При первом способе мы начинаем с самого низа, то есть мы добавляем семантическую разметку в документы опубликованные в Сети. Таким образом, пользовательские агенты получают доступ к метаданным. Этот процесс понемногу начинает набирать темп. Все чаще и чаще можно встретить данные в формате RDF, встроенные в те или иные странички. Каковы перспективы этого подхода?

- Во-первых, нужно отметить, что существует огромная разница в психологии людей, занимающихся созданием контента. Большинство людей крайне скептически воспринимают перспективу не просто излагать свои мысли виде обычного текста, но еще и предпринимать особые шаги для того, чтобы объяснить свои идеи бездушному (или безмозглому) компьютеру. Тем не менее, многие склонны видеть эту ситуацию в ином свете. Они готовы часами приводить в порядок свои данные, расставлять метки и писать комментарии, составлять каталоги и рейтинги. Все ради того, чтобы обеспечить удобный, хорошо структурированный доступ к информации.
- Во-вторых, все больше и больше в Интернете публикуется автоматически генерируемой информации. Всевозможные базы данных, отчеты, прогнозы погоды, списки и т.д. и т.п. Конечно, добавление семантической информации в автоматически генерируемые документы требует значительно меньших усилий.
- В-третьих, сейчас активно развиваются инструменты для семантической разметки документов. Нужно понимать, что семантическая информация, которую вы добавляете в свой документ, способна немедленно оказать вам помощь. Возьмем простой случай, вы пишете письмо другу Пете с предложением пойти выпить пива. Если вы даете понять своему компьютеру о чем идет речь, то вы можете тут же получить контекстную информацию необходимую вам. Например: когда у Петра сегодня заканчивается работа, список пивных баров наиболее подходящих для вас (куда удобно добраться и вам и ему, с учетом пробок на улицах города), какое пиво предпочитает ваш друг, какие важные события в его жизни должны произойти в ближайшее будущее, и так далее. Причем, у компьютера появляются уникальные возможности для того, чтобы подстраиваться именно под ваши интересы, предпочтения и стиль работы, а возможность кооперации с другими компьютерами в Сети позволит ему выполнять эту работу весьма качественно. Таким образом, пользователь будет стимулироваться к тому, чтобы наполнять семантической информацией все что он делает. Более того, вполне можно представить себе ситуацию, когда пользователь предпочтет указывать информацию только в виде понятном машине, предоставляя компьютеру всю остальную работу связанную с формулированием и оформлением данных для потенциального читателя. Сколько разного рода формальных бумаг

нам приходится создавать: справки, счета, отчеты, заявления. Значительную часть этой рутины компьютер может взять на себя.

- В-четвертых, большое количество метаданных создается неявным образом. Сервисы социальных закладок, такие как del.icio.us, с одной стороны стали весьма популярны в современной Сети, а с другой стороны они активно собирают метаданные в виде тегов, описаний и оценок сайтов. Если определить семантические отношения между отдельными тегами, создав, тем самым, некую онтологию, то мы получим огромное количество вполне релевантных семантических данных. В том же направлении могут двигаться и системы коллаборативной фильтрации, такие как, MovieLens и Last.fm. Они уже давно зарекомендовали себя как весьма и весьма эффективные инструменты. Еще дальше пошли авторы проекта DBin Они предлагают новую парадигму работы в Сети: "Сообщества Семантического Веба". Используя систему DBin пользователи могут обмениваться релевантной информацией так же, как пользователи пиринговых сетей обмениваются файлами.

Другой подход предполагает использование средств анализа текстов на естественных языках (Natural Language Processing - NLP). Такие инструменты должны прочитать и обработать существующие в Сети документы, чтобы извлечь из них семантические данные. К сожалению средства NLP еще далеки от совершенства. Сегодня, они не способны, в автоматическом режиме, семантически размечать документы. Однако, не надо недооценивать возможностей таких инструментов. Например: современные системы извлечения фактов позволяют найти в тексте (для английского языка) до 96% именованных объектов, то есть имен людей, названий компаний, адресов, телефонов, названий технологий, брендов и т.д. Программы синтаксического разбора русского языка позволяют правильно определить подлежащее и сказуемое примерно в 60% предложений. Уже этого достаточно, для того, чтобы извлечь из текста огромное количество семантически значимой информации. При этом, следует отметить, что технологии Семантического Веба начинают, в свою очередь, оказывать влияние на развитие инструментов NLP. Например: уже сейчас создана и внедряется в Совете Федерации Федерального Собрания Российской Федерации первая очередь информационной системы "Семантический контроль текстов редактируемых документов", которая активно использует технологии Семантического Веба. С идеей использования информации об окружающем мире (что, по сути, предполагает применение хранилищ знаний аналогичных RDF репозиториям) связаны огромные надежды в области систем распознавания текстов, синтеза и распознавания речи. Очевидно, что прогресс в сфере Семантического Веба будет способствовать и развитию систем распознавания образов.
Сегодня можно утверждать, что оба подхода к созданию среды, наполненной семантической информацией, будут развиваться параллельно, дополняя друг друга.

WordNet – лексическая база данных, содержащая:

- основную лексику языка (существительные, глаголы, прилагательные и наречия), организованную в виде синсетов.
- таксономию отношений между синсетами (например, гипонимия, меронимия) и между лексемами (например, антонимия).
- определение семантических классов – *TopOntology*

Synset (синсет) – основная структура, представляющая словарную статью в ***WordNet***. Синсет представляет множество лексем с одинаковым значением.
synset - A synonym set; a set of words that are interchangeable in some context.

5.4. Семантические сети

http://ru.wikipedia.org/wiki/

— один из способов представления знаний. В названии соединены термины из двух наук: семантика в языкознании изучает смысл предложений, а сеть в математике представляет собой разновидность графа — набора вершин, соединённых дугами. В семантической сети роль вершин выполняют понятия базы знаний, а дуги (причем направленные) задают отношения между ними. Таким образом, семантическая сеть отражает семантику предметной области в виде понятий и отношений.
Ни в коем случае нельзя смешивать понятия «Семантическая сеть» (англ. *Semantic Network*) и «Семантическая паутина» (англ. *Semantic Web*). Это несоответствие возникает как раз из-за неточного перевода.
Семантическая сеть есть набор элементов, представляющих понятия предметной области (слова и словосочетания), которые связаны между собой ассоциативными связями, и может быть описана матрицей весов связей:
$W = [w_{ij}]$,
где w_{ij} может интерпретироваться как степень ассоциированности и отражать вероятность появления понятия j в смысловой связи с понятием i в рамках предметной области, описываемой сетью.

5.5. Нейронные сети

http://ru.wikipedia.org/wiki/

— это математическая модель, а также устройства параллельных вычислений, представляющие собой систему соединённых и взаимодействующих между собой простых процессоров (искусственных нейронов).

Нейроподобная структура – вычислительная структура, построенная на базе нейронной сети (либо аппарата нейронных сетей), работающая над входной информацией, имеющей упорядоченный характер (базы данных, базы знаний). В «чистом» виде нейронной сети на данном этапе развития техники и технологии быть не может. Осуществление моделирования НС возможно лишь на математическом аппарате и в вычислительных средах, основанных на конечных автоматах и работающих по алгоритмическому принципу. Поэтому имеет смысл говорить о ***нейроподобных структурах*** как результате практической реализации нейронных сетей.

5.6. Методы и средства обработки

SVD (Singular-Value Decomposition) - сингулярное разложение

http://www.dialog-21.ru/trends/?id=15539&f=1

Самая популярная модель, работающая по смыслу. В теории информационного поиска данную модель принято называть *латентно-семантическим индексированием* (иными словами, выявлением скрытых смыслов). Эта алгебраическая модель основана на *сингулярном разложении* прямоугольной матрицы, ассоциирующей слова с документами. Элементом матрицы является частотная характеристика, отражающая степень связи слова и документа, например, TF*IDF. Вместо исходной миллионно-размерной матрицы авторы метода [furnas], [deerwester] предложили использовать 50-150 «скрытых смыслов»[3], соответствующих первым главным компонентам ее сингулярного разложения.

Сингулярным разложением действительной матрицы A размеров m*n называется всякое ее разложение вида A = USV, где U - ортогональная матрица размеров m*m, V - ортогональная матрица размеров n*n, S - диагональная матрица размеров m*n, элементы которой s_{ij}= 0, если *i* не равно *j*, и s_{ii} = s_i >= 0. Величины s_i называются **сингулярными числами** матрицы и равны арифметическим значениям квадратных корней из соответствующих собственных значений матрицы AA^T. В англоязычной литературе сингулярное разложение принято называть **SVD-разложением**.

LSA (Latent Semantic Analyses) – латентно-семантический анализ

http://meta.math.spbu.ru/~igor/thesis/node5.html#SECTION00232000000000000000

Латентно-семантический анализ (LSA) -- это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных. В течение нескольких последних лет этот метод не раз использовался как в области поиска информации, так и в задачах фильтрации и классификации.

Латентно-семантический анализ основывается на идее, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые в значительной степени позволяют определить похожесть смысловых значений слов и множеств слов между собой.

В качестве исходной информации LSA использует матрицу термы-на-документы, описывающую используемый для обучения системы набор данных. Элементы этой матрицы содержат частоты использования каждого терма в каждом документе.

Наиболее распространенный вариант LSA основан на использовании разложения матрицы по **сингулярным значениям** (SVD). Используя SVD, огромная исходная матрица разлагается во множество из, обычно от 70 до 200, ортогональных матриц, линейная комбинация которых является неплохим приближением исходной матрицы.

SVM (Support Vector Machines) – метод опорных векторов

http://en.wikipedia.org/wiki/Support_vector_machine

SVM (Support Vector Machines) – метод опорных векторов – метод построения классификатора, минимизирующего верхнюю оценку ожидаемой ошибки классификации (в том числе и для неизвестных объектов, не входивших в тренировочный набор).

В исходном виде SVM представляет собой алгоритм, обучающийся ровно одной задаче: различению объектов двух классов. И этому SVM обучается очень быстро по сравнению, например, с нейронными сетями, поскольку вместо сложной нелинейной минимизации SVM решает задачу минимизации всего лишь квадратичного функционала, правда при огромном количестве линейных ограничений. И классифицирует обученная SVM быстро. Но при попытке применить SVM к многоклассовой задаче качество и скорость работы резко падают.

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. They can also be considered a special case of Tikhonov regularization. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as **maximum margin classifiers**.
http://www.dtreg.com/svm.htm

Ссылки на документацию и примеры использования SVM

General

- www.pascal-network.org *(EU Funded Network on Pattern Analysis, Statistical Modelling and Computational Learning)*
- www.kernel-machines.org *(general information and collection of research papers)*
- www.kernel-methods.net *(News, Links, Code related to Kernel methods - Academic Site)*
- www.support-vector.net *(News, Links, Code related to Support Vector Machines - Academic Site)*
- www.support-vector-machines.org *(Literature, Review, Software, Links related to Support Vector Machines - Academic Site)*
- www.support-vector.ws *(Free educational MATLAB based software for SVMs, NN and FL , Links, Publications downloads, Semisupervised learning software SemiL, Links)*

Software

- Java applet with an SVM implementation
- The Kernel-Machine Library (GNU) C++ template library for Support Vector Machines
- Lush -- a Lisp-like interpreted/compiled language with C/C++/Fortran interfaces that has packages to interface to a number of different SVM implementations. Interfaces to LASVM, LIBSVM, mySVM, SVQP, SVQP2 (SVQP3 in future) are available. Leverage these against Lush's other interfaces to machine learning, hidden markov models, numerical libraries (LAPACK, BLAS, GSL), and builtin vector/matrix/tensor engine.
- SVMlight -- a popular implementation of the SVM algorithm by Thorsten Joachims; it can be used to solve classification, regression and ranking problems.
- SVMProt -- Protein Functional Family Prediction.
- LIBSVM -- A Library for Support Vector Machines, Chih-Chung Chang and Chih-Jen Lin
- YALE -- a powerful machine learning toolbox containing wrappers for SVMLight, LibSVM, and MySVM in addition to many evaluation and preprocessing methods.
- LS-SVMLab - Matlab/C SVM toolbox - well-documented, many features
- Gist -- implementation of the SVM algorithm with feature selection.
- Weka -- a machine learning toolkit that includes an implementation of an SVM classifier; Weka can be used both interactively though a graphical interface or as a software library. (The SVM implementation is called "SMO". It can be found in the Weka Explorer GUI, under the "functions" category.)
- OSU SVM - Matlab implementation based on LIBSVM
- Torch - C++ machine learning library with SVM
- Shogun - Large Scale Machine Learning Toolbox that provides several SVM implementations (like libSVM, SVMlight) under a common framework and interfaces to Octave, Matlab, Python, R
- Spider - Machine learning library for Matlab
- e1071 - Machine learning library for R
- SimpleSVM - SimpleSVM toolbox for Matlab
- SVM and Kernel Methods Matlab Toolbox
- PCP -- C program for supervised pattern classification. Includes LIBSVM wrapper.
- TinySVM -- a small SVM implementation, written in C++

- pcSVM is an object oriented SVM framework written in C++ and provides wrapping to Python classes. The site provides a stand alone demo tool for experimenting with SVMs.
- PyML -- a Python machine learning package. Includes: SVM, nearest neighbor classifiers, ridge regression, Multi-class methods (one-against-one and one-against-rest), Feature selection (filter methods, RFE, multiplicative update, Model selection, Classifier testing (cross-validation, error rates, ROC curves, statistical test for comparing classifiers).

Interactive SVM applications

- ECLAT classification of Expressed Sequence Tag (EST) from mixed EST pools using codon usage
- EST3 classification of Expressed Sequence Tag (EST) from mixed EST pools using nucleotide triples

AJAX (Asynchronous JavaScript and XML) — это новая технология, применяемая в самых современных проектах, таких как gmail.com. Однако среди популярных поисковых систем AJAX еще не применялся. Nigma.ru, например, решила использовать эту технологию для отображения результатов кластеризации в виде "облака" тэгов. Тэги-"облачка" разлетаются по синему "небу"; на каждый тэг можно кликнуть и получить подтэги, связанные с этим тэгом, а под облаком тэгов отображаются соответствующие тэгу результаты поиска.

http://ru.wikipedia.org/wiki/Ajax

AJAX (от англ. *Asynchronous JavaScript and XML* — «асинхронный JavaScript и XML») — это подход к построению интерактивных пользовательских интерфейсов веб-приложений. При использовании AJAX веб-страница не перезагружается полностью в ответ на каждое действие пользователя. Вместо этого с веб-сервера догружаются только нужные пользователю данные. AJAX — один из компонентов концепции DHTML.

AJAX базируется на двух основных принципах:

- использование DHTML для динамического изменения содержания страницы;
- использование технологии динамического обращения к серверу «на лету», без перезагрузки всей страницы полностью, например:
  - с использованием XMLHttpRequest;
  - через динамическое создание дочерних фреймов;
  - через динамическое создание тега <script>.

http://ru.wikipedia.org/wiki/Xml

XML (eXtensible Markup Language) — расширяемый язык разметки.

Рекомендован Консорциумом Всемирной паутины язык разметки, фактически представляет собой свод общих синтаксических правил. XML предназначен для хранения структурированных данных (взамен существующих файлов баз данных), для обмена информацией между программами, а также для создания на его основе более специализированных языков разметки (например, XHTML), иногда называемых *словарями*. XML является упрощённым подмножеством языка SGML.

Целью создания XML было обеспечение совместимости при передаче структурированных данных между разными системами обработки информации, особенно при передаче таких данных через Интернет. Словари, основанные на XML (например, RDF, RSS, MathML, XHTML, SVG), сами по себе формально описаны, что позволяет программно изменять и проверять документы на основе этих словарей, не зная их семантики, то есть не зная смыслового значения элементов. Важной особенностью XML также является применение так называемых пространств имён (англ. *namespace*).

http://ru.wikipedia.org/wiki/Xml

XSL является технологией, описывающей как форматировать или трансформировать данные XML документа. Документ трансформируется в формат, подходящий для отображения в браузере. Процесс аналогичен применению CSS к HTML документу для отображения. Браузер это наиболее частое использование XSL, но не стоит забывать, что с помощью XSL можно трансформировать XML в любой формат, например VRML, PDF, текст.

Без использования CSS или XSL, XML-документ отображается как простой текст в большинстве web-браузеров. Некоторые браузеры, такие как Internet Explorer, Mozilla и Mozilla Firefox отображают структуру документа в виде дерева, позволяя сворачивать и разворачивать узлы с помощью нажатий клавиши мыши.

http://ru.wikipedia.org/wiki/Xsl

XSL (Extensible Stylesheet Language) — расширяемый язык таблиц стилей.

- Стилевые таблицы XSL позволяют определять оформление элемента в зависимости от его месторасположения внутри документа, то есть к двум элементам с одинаковым названием могут применяться различные правила форматирования.
- Языком, лежащем в основе XSL, является XML, а это означает, что XSL более гибок, универсален и у разработчиков появляется возможность использования средства для контроля за корректностью составления таких стилевых списков(используя DTD или схемы данных)
- Таблицы XSL не являются каскадными, подобно CSS, так как чрезвычайно сложно обеспечить «каскадируемость» стилевых описаний, или, другими словами, возможность объединения отдельных элементов форматирования путём вложенных описаний стиля, в ситуации, когда структура выходного документа заранее неизвестна и он создаётся в процессе самого разбора. Однако в XSL существует возможность задавать правила для стилей, при помощи которых можно изменять свойства стилевого оформления, что позволяет использовать довольно сложные приёмы форматирования.

Разработчики обычно не задумываются о разнице в использовании терминов XSL и XSLT. На самом деле, спецификация XSL состоит из двух довольно-таки независимых частей:

- XSL-T (XSL Transformations), язык для преобразования XML и
- XSL-FO (XSL Formatting Objects), язык для вёрстки XML.

http://ru.wikipedia.org/wiki/Xslt

XSLT (Extensible Stylesheet Language Transformations) — часть спецификации XSL, задающая язык преобразований XML-документов. Спецификация XSLT является рекомендацией W3C.
При применении *таблицы стилей* XSLT, состоящей из набора *шаблонов*, к XML-документу (*исходное дерево*) образуется *конечное дерево*, которое может быть как XML-структурой, так и обычным текстом. Запросы выбора данных из исходного дерева пишутся на языке запросов XPath.
XSLT находит множество различных применений, в основном в области web-программирования.
Консорциум W3 определяет три составные части языка XSL (от англ. eXtensible Stylesheet Language — Расширяемый Язык Стилей): XSLT, XPath (язык путей и выражений, используемый в XSLT для доступа к отдельным частям XML-документа) и XSL Formatting Objects — словарь, определяющий семантику форматирования документов.

http://ru.wikipedia.org/wiki/SOAP

SOAP (Simple Object Access Protocol) — протокол обмена структурированными сообщениями в распределённой вычислительной среде. Первоначально SOAP предназначался, в основном, для реализации удалённого вызова процедур (RPC), а название было аббревиатурой: *Simple Object Access Protocol* — простой протокол доступа к объектам. Сейчас протокол используется для обмена произвольными сообщениями в формате XML, а не только для вызова процедур. Официальная спецификация последней версии 1.2 протокола никак не расшифровывает название SOAP. SOAP является расширением языка XML-RPC.
SOAP может использоваться с любым протоколом прикладного уровня: SMTP, FTP, HTTP и др. Однако его взаимодействие с каждым из этих протоколов имеет свои особенности, которые должны быть определены отдельно. Чаще всего SOAP используется поверх HTTP.
SOAP является одним из стандартов, на которых базируется технология веб-сервисов.

VML (Vector Markup Language) – язык для рисования геометрических фигур, использующий векторную графику
http://en.wikipedia.org/wiki/Vector_Markup_Language#See_also
Vector Markup Language (**VML**) is an XML language used to produce vector graphics. VML was submitted as a proposed standard to the W3C in 1998 by Microsoft, Macromedia, and others. VML was rejected as a web standard because Adobe, Sun, and others submitted a competing proposal known as PGML [1]. The two standards were joined and improved upon to create SVG.
Even though rejected as a standard by the W3C, and largely ignored by developers, Microsoft still implemented VML into Internet Explorer 5.0 and higher and in Microsoft Office 2000 and higher.
Google Maps currently uses VML for rendering vectors when running on Internet Explorer 5.5+[2].

Другие средства для графики:
http://en.wikipedia.org/wiki/Vector_graphics_markup_language

LZW (algorithm)

тм мпеля — ива — лча (**Lempel-Ziv-Welch**, **LZW**) — это универсальный алгоритм сжатия данных без потерь, созданный Абрахамом Лемпелем (*Abraham Lempel*), Якобом Зивом (*Jacob Ziv*) и Терри Велчем (*Terry Welch*). Он был опубликован Велчем в 1984 году, в качестве улучшенной реализации алгоритма LZ78, опубликованного Лемпелем и Зивом в 1978 году. Алгоритм разработан так, чтобы его можно было быстро реализовать, но он не обязательно оптимален, поскольку он не проводит никакого анализа входных данных.
http://ru.wikipedia.org/wiki/LZW

RSS (Really Simple Syndication) (в RSS 2.x дословно — очень простое приобретение информации) — это формат обмена информацией на базе языка разметки XML, который разрабатывался для публикации новостей на новостных и подобных сайтах, хотя данная технология позволяет публиковать любой материал, который можно разбить на отдельные части. Одна из главных проблем сегодняшнего Интернета — недостаточная стандартизация, вследствие чего новостные ленты на разных сайтах настолько отличаются друг от друга по своему оформлению и расположению, что обрабатывать их можно только вручную. Для решения автоматизации этой задачи и был предложен формат RSS.
RSS — семейство XML-форматов, предназначенных для описания лент новостей, анонсов статей, изменений в блогах и т. п. Информация из различных источников, представленная в формате RSS, может быть собрана, обработана и представлена пользователю в удобном для него виде специальными программами-агрегаторами.
RSS-агрегатор — программа для автоматического сбора сообщений из источников, экспортирующих в форматы RSS или Atom.
Агрегаторы бывают двух типов. Web-агрегаторы и программные агрегаторы. Задачи их одинаковы — работа с RSS и получение обновлений.
Программный агрегатор - Это компьютерные программы, которые устанавливаются на компьютер для работы с RSS. Они могут быть встроены в браузеры или почтовые программы, или даже в операционную систему, но могут быть и отдельными программами. Браузеры Opera и FireFox, Internet Explorer (с версии 7.0) поддерживают агрегацию. iTunes — агрегатор для подкастов.
Веб-агрегатор - Агрегатор находящийся в Интернете, таким образом к нему можно получать доступ с любого компьютера, где есть интернет. Примеры таких агрегаторов: Яндекс.Лента, Google Reader

6. Фактографический анализ

Автоматизированные информационные системы (АИС) создавались как **фактографические системы** с представлением информации пользователям в виде регламентированных форм, в которых фактографическая информация была сгруппирована в соответствии с решаемыми на ее основе прикладными задачами.
В большинстве случаев и ввод информации в целях удобства сбора данных осуществлялся с помощью предварительно заполняемых форм. И теоретически АИС можно считать документально-фактографическими ИПС. Однако, как правило, эта терминология в практике разработки АИС не использовалась. (http://www.interface.ru/home.asp?artId=1636)
В последнее время появился широкий спектр специализированных ИС - экономические информационные системы (ЭИС), бухгалтерские информационные системы (БУИС), банковские информационные системы (БИС), информационные системы рынка ценных бумаг, маркетинговые ИС (МИС) и т.п.
До 85% новых знаний аналитики до сих пор получают, изучая тексты. В ближайшем будущем наиболее востребованными станут системы с максимально автоматизированными ETL-процессами структурирования контента (extract, transfer, load — «извлечение, преобразование, загрузка»). Важной чертой таких систем будет функция **оперативного анализа** информации, полученной по запросу для выбора дальнейшего направления исследования документов (автопилотирование направления исследования), выполняемой с помощью методов интеллектуального анализа текста.
К наиболее актуальным средствам интеллектуального анализа текстов относятся технологии выделения **фактографической информации** об объектах с учетом анафорических ссылок на них (ссылочные местоимения на объект, поименованный в тексте); нечеткий поиск; тематическое и тональное (точное и полное) рубрицирование; кластерный анализ хранилищ и подборок документов; выделение ключевых тем; построение аннотаций; построение многомерных частотных распределений документов и их

исследование с помощью OLAP-технологий; использование методов интеллектуального анализа текста для определения направления исследования больших подборок документов и извлечения новых знаний.
В современных системах используется двухфазная технология аналитической обработки. В первой фазе (ETL) производится автоматизированный анализ отдельных документов, структуризация их контента и формирование хранилищ исходной и аналитической информации. Во второй фазе (OLAP, Text Mining, Data Mining) — извлечение в оперативном режиме знаний из хранилища или из полученной по запросу подборки документов. На наш взгляд, к наиболее интересным системам аналитической обработки относятся ClearForest, Convera RetrievalWare, Hummingbird KM, IBM Text Miner, инструменты компании IQMen, Inxight Smart Discovery Extraction Server, Ontos Miner, Oracle Text, ODB-Text, TextAnalyst, инструменты компании Smartware, XANALYS Link Explorer, X-Files, инструменты компании «Гарант-Парк-Интернет» и «Медиалогия». Попробуем проанализировать современное состояние дел в области аналитических технологий на примерах конкретных систем.
В фактографических ИС регистрируются **факты** - конкретные значения данных атрибутов об объектах реального мира. Основная идея таких систем заключается в том, что все сведения об объектах (фамилии людей и названия предметов, числа, даты) сообщаются компьютеру в каком-то заранее обусловленном формате (например, дата - в виде комбинации ДД.ММ.ГГ).
Информация, с которой работает фактографическая ИС, имеет четкую структуру, позволяющую машине отличать одно данное от другого, - например, фамилию от должности человека, дату рождения от роста и т.п. Поэтому фактографическая система способна давать однозначные ответы на поставленные вопросы, например: “Сколько мониторов марки HITACHI продал магазин “Компьютеры” в июне 2002 г.?”, “Кто из работников фирмы с датой рождения не ранее 1 января 1980 г. имеет высшее образование?”, “Какие культурно-исторические памятники г.Тулы посетили летом 2002 г. Иностранные туристы?” и т.д.

Атрибут — структурный элемент типа досье, предназначенный для накопления фактов одного типа (биографические данные, сведения о поездках и др.). Атрибут имеет семантический фильтр для выделения «своих» фактов из потока документов. Один атрибут может входить в досье нескольких типов.
Атрибут – это комплексная характеристика группы объектов, которая описывает определенную сторону их деятельности. Атрибут позволяет сгруппировать в один класс совокупность различных фактов, имеющих одинаковый смысл для пользователя.
Каждый атрибут имеет свое название и описание атрибута, в соответствии с которым производится отнесение найденных фактов к атрибуту - классификация.
После нахождения в тексте упоминания факта, относящегося к атрибуту, программой формируется значение атрибута, которое представляет собой имя одного из фигурантов факта. Имена других найденных участников могут формировать комментарий к атрибуту. Например, значениями атрибута с названием “Договора” могут быть слова “Александр Иванов”, “коммерческий банк”, “ряд организаций”, а комментариями к этому атрибуту – “сотрудничество”, “постройка корабля”.
Описание атрибута представляет собой список фактов. Например, атрибут “владение предприятиями” может включать в себя факты “купля-продажа акций”, “купля-продажа предприятий”, “владение акциями”, “владение предприятиями”. При включении факта в атрибут указывается роль целевого участника, которая позволяет отнести факт к атрибуту, связанному с объектом, если объект фигурирует в этой роли. Например, в атрибут “владение предприятиями” может быть дважды включен факт “купля-продажа акций” с указанием роли “продавца” и роли “покупателя”, так как оба фигуранта на какой-то момент времени являются владельцами предприятия. Аналогично указываются роли других участников, которые формируют значение атрибута. Прочие найденные участники выносятся в комментарий к атрибуту.

Гипотеза — аналитическое высказывание (полученное в результате аналитической обработки данных, например, прогнозирования) относительно состояния атрибута досье, которое сопровождается аргументирующей информацией, ссылками на источники и др. Гипотеза может порождаться разными подсистемами извлечения знаний или экспертами.

Досье — реализация типа досье для конкретного объекта.
Вся найденная информация, связанная с объектами, автоматически классифицируется и собирается в структуру, называемую досье.
Досье представляет собой таблицу со столбцами: название объекта, название атрибута, значение атрибута, число документов, частота встречаемости, комментарий и подкрепления в тексте. Каждая запись-строка досье содержит информацию, относящуюся к одному из атрибутов одного из объектов - множество однотипных фактов, имеющих общего участника-фигуранта.
Структура досье позволяет приписать объектам желаемые атрибуты и определяет, какие строки могут присутствовать в досье. Структура досье представляет собой дерево, в котором объекты и атрибуты могут быть сгруппированы в логически связанные узлы для удобства работы. Это позволяет просматривать досье по частям, относящимся к выбранным объектам и их совокупностям.

Объект — сущность, информация о которой накапливается в системе. Объект имеет семантический фильтр для самоидентификации в тексте.
Объект - целевая персона или организация, по которой производится поиск фактов в тексте.

Связь — направленное или ассоциативное отношение определенного типа между объектами системы. Связь представляется специальным типом атрибута в каждом досье связываемых объектов.

Тип досье — описание проблемной области, представленное в виде иерархии атрибутов. Для каждого объекта должен быть определен хотя бы один тип досье.

Факт — событие (как правило, зафиксированное и произошедшее), сопровождаемое временной и географической метками, аргументирующей информацией, ссылками на источники и др. Факт может быть извлечен из текста документов либо определен экспертом. Он может определять как свойства объекта, так и его связь с другими объектами.
Фактом называется некоторая элементарная ситуация, представляющая интерес для пользователя, описание которой может быть найдено в тексте. Элементарные ситуации представляют тот базис, из элементов которого складывается экспертное суждение о более сложных понятиях и явлениях предметной области. Например, понятие “политический профиль” не выражается в тексте напрямую, однако может формироваться из множества элементарных фактов вида “кого поддержал на выборах?”, “кого и за что критикует?” и т.п.
Каждый факт имеет свое название (например, “покупка акций”) и список ролей возможных участников-фигурантов (“покупатель”, “продавец”, “эмитент акций”). Выделение факта и его участников-фигурантов происходит в соответствии с лингвистическими описаниями, которые описывают возможные способы выражения факта в тексте и роли участников.

6.1. Добыча данных

Широкое применение методов искусственного интеллекта позволяет порождать гипотезы — предложения по дальнейшему исследованию. Типичная технология анализа взаимосвязей проблем содержит следующие фазы:
- получение подборки документов по запросу;
- получение ее семантической карты;
- просмотр документов о связи выделенной пары тем;
- кластерный анализ этих документов;
- анализ документов нужных кластеров;
- резюме о структуре связи тем.

Типичная технология анализа динамики развития проблемы в регионе (стране) включает следующие фазы:
- получение подборки документов по запросу;
- получение двумерного частотного распределения рубрик-проблем по регионам;
- выделение значимой проблемы в исследуемом регионе;
- получение частотного распределения рубрики-проблемы в регионе по времени;
- анализ документов в пиковые периоды времени;
- кластерный анализ этих документов;
- предложения по нормализации проблемы.

К примеру, многие ежедневно ездят на работу по Москве, но эти факты еще не свидетельствуют о наличии связи между ними, однако если два дипломата работали в одно время в небольшой стране, то с большой вероятностью следует, что они могли быть знакомы. Система должна уметь предлагать аналитику такого типа гипотезы.

6.2. Извлечение и структурирование фактографической информации

Одна из главных целей систем извлечения фактографической информации из делового текста состоит в нахождении, интерпретации и стандартизации числовых и, более широко, количественных данных.

К таким данным могут относиться следующие типы:
- дата и время (включая интервалы, возраст, период и т.п.)
- физический размер (габариты: длина, высота, ширина; объем)
- координаты (географические, относительные, местоположение, направление, и.т.п.)
- физические и химические характеристики (плотность, мощность, вес и т.п.)
- поименованные идентифицирующие характеристики (ФИО, наименование, состояние и т.п.)
- количественные идентификаторы (номер, артикул, и.т.п.)

В общем случае набор таких данных будет зависеть от конкретной предметной области.

Для выделения объектов и их свойств (адреса, поездки, встречи, бизнес и т. п.) используются компоненты управления фактографической информацией и ведения досье. Например, в терминах системы Xfiles факт об объекте является структурированным представлением фрагментов текста документа в виде значения факта: его суть, время и место совершения, его участники. Факты выделяются из предложений, содержащих упоминания объектов или ссылки на них. Технология выделения фактов основана на использовании специальных семантико-лингвистических методов, которые дают возможность получить точность и полноту фактов, сравнимую с экспертными.
Зачастую факты содержат информацию о взаимосвязях объектов и классифицируются как прямые (имеется факт о связи двух объектов); нечеткие (нет фактов); общего места и времени (для пары различных фактов различных объектов); косвенные, или транзитивные (через общий третий объект-связь у пары фактов различных объектов); рефлективные (между парой атрибутов досье, связанных семантически). Если в одном из них появляется факт с определенным объектом-связью, то в симметричном атрибуте для объекта-связи также появляется этот факт. Скажем, атрибут «продажа акций» имеет симметричный атрибут «покупка акций». Симметричные атрибуты «срабатывают» по прямым связям. Свойство симметричности задается при создании атрибутов независимо от того, в какие досье они входят. При включении атрибута в другое досье свойство симметричности сохраняется.
Все эти свойства необходимы в системах аналитической разведки, немыслимых без следующих сервисов: автоматическое выявление прямых и косвенных (т. е. через третье лицо) связей объекта; автоматическое выявление связей объектов по месту и времени (когда события произошли с разными объектами в одном месте или в близкое время); типизация связей, представленных различной лексикой; формирование групп объектов, связанных между собой общностью фактов (например, место, время, содержание факта); построение карты связей объектов для различных типов связей, визуализация и фильтрация связей; поиск оптимальных (обычно кратчайших) связей между заданными объектами; построение многомерных частотных распределений фактов. Сегодня системы извлечения фактов являются наиболее эффективным инструментом выделения нужной для принятия решений информации, заменяя ее поиск.

6.3. Извлечение знаний из текстовой информации

К наиболее актуальным методам сегодня можно отнести: семантические сети тем и объектов текста документов, выделение фактографической информации с учетом анафорических ссылок, возможность параллельной обработки распределенных архивов документов, различные стратегии нечеткого поиска, тематическое и тональное рубрицирование, кластеризация документов, аннотирование, анализ многомерных частотных распределений документов.
Технология выделения фактов из текста основана на использовании специальных семантико-лингвистических методов, которые дают возможность получить точность и полноту фактов, сравнимую с экспертными. Вкратце суть метода обработки каждого документа заключается в следующем.
Сначала строится дайджест объекта, который содержит все предложения документа, содержащие ссылки на объект. Дайджест учитывает анафорические ссылки между предложениями (кореферентные связи). Затем строится информационный портрет документа на основе смысла элементов текста, извлекаемых средствами синтаксического анализа и синтеза. Далее он преобразуется в семантическую сеть, обеспечивающую инвариантность представления смыслов относительно ряда особенностей поверхностно-синтаксической организации текста. Например, семантическая сеть позволяет абстрагироваться от малоинформативных элементов формально-синтаксической структуры текста (порядка слов, залога и т.п.) и представляет его пропозициональную структуру в терминах описываемых ситуаций (предикатов) и их участников (аргументов) в определенных семантических ролях. Для решения задачи выделения фактов полное представление смысла текста в форме семантической сети является избыточным и непродуктивным, оно имеет большой объем (превышающий объем документа), а его утилизация требует высокопроизводительного оборудования и развитых нетривиальных средств для поиска и сравнения структур на графах.
Будучи дополнен правилами для генерации канонической формы синтагм, синтаксический анализ-синтез позволяет описать каждый смысловой атрибут текста в виде строки, инвариантной к его

грамматическому выражению в различных фразах. Например, фразам "Транспорт был арендован террористом у автобазы", "Террорист арендует у автобазы транспорт" и "Аренда транспорта террористом у автобазы" будут соответствовать одинаковые элементы смысла: "террорист арендует", "аренда транспорта", "аренда у автобазы".
Выделяемые связи между элементами смысла можно разделить на следующие основные классы:
- связи между ситуациями и их участниками - предикатно-аргументные связи, например: (сделать, покупка), (продажа, акции).
- связи внутри именных групп (генитивные цепочки), обычно называющих участников ситуации, - атрибутивные связи, например, акт (террористический, боевиков), предприятие (прибыльное, город).
- связи между ситуациями - предикатно-предикатные, например, покупать (учиться), бороться (искореняя).
- связи ситуаций с обстоятельствами или дополнительными атрибутами.
В последней технологической фазе извлечения а-фактов движок фактографических правил на основе семантической сети дайджеста производит поиск шаблонов фактов и сохраняет структурированное описание выделенных фактов в базе данных системы. Хотелось бы ещё раз отметить, что выделенный факт - это не контекст, а выделенные из него свойства.
Алгоритм выделения фактов из текстов наиболее глубоко проработан для русского языка, для большинства других языков могут использоваться источники документов (например, Hummingbird SS 2004), поддерживающие многоязычный поиск.
Следовательно, процесс извлечения знаний является достаточно трудоемким. Формализуя представленную систему, имеем следующее. S1 = Лингвистический эшелон = {сбор ретроспекти-вных данных по ряду характеристик, унификация форм представления, формирование исходного массива}. S2 = Анализ фактографической информации на основе многомерных методов статистики = {факторный анализ, кластерный анализ, построение когнитивной модели}. S3 = Гносеологический эшелон = {Определение методов исследования, разработка методики прогнозирования, исследование на модели}. Понимание системы возрастает при последовательном переходе от одной страты к другой: чем ниже мы опускаемся по иерархии, тем более детальным становится раскрытие системы, чем выше мы поднимаемся, тем яснее становится смысл и значение всей системы

1. Количественная информация и количественные группы
Рассмотрим стандартный пример контекста, в котором, несомненно, имеется количественная информация:
Жесткие диски емкостью до 100 ГБ.
Здесь можно выделить следующий набор значимых для результата анализа элементов [1]:
1) имя объекта: *жесткие диски;*
2) наименование признака: *емкость;*
3) количественное (в данном примере числовое) значение: *100;*
4) единица измерения: *ГБ*
5) модификатор значения: *до.*
Некоторые из элементов могут отсутствовать: *Жесткие диски до 100 ГБ.*
Собственно **количественной группой** мы будем называть фрагмент, содержащий элементы 2 – 5 (не включая сюда имя характеризуемого объекта.). Пару, составленную из имени (классификационного описания) объекта и количественной группы, естественно назвать **количественным фактом.**
Разные варианты количественных групп могут быть охвачены следующей простой классификацией, выстроенной по различению типа используемой оценки упомянутого в группе признака (для каждого типа указана возможная логическая интерпретация):

А. Числовые группы
1) Точечные:
мощностью 100 вт - МОЩНОСТЬ_вт (x, v) & v = 100
2) Типа "пятно":
*мощностью **около** 100 вт - МОЩНОСТЬ_вт (x, v) & v ≈ 100*
3) Интервальные:
- зона, ограниченная снизу: *мощностью свыше 100 вт*;
- зона, ограниченная сверху: *мощностью до 100 вт*;
- собственно диапазон
мощностью от 100 до 1000 вт - МОЩНОСТЬ_вт (x, v) & v >= 100 & v <= 1000
4) Парциальные: *В Латвии сейчас до трети населения неграждане*.
5) Представляющие числовую оценку динамики изменения:
- «на сколько» - абсолютная оценка:
мощность увеличена на 100 вт - МОЩНОСТЬ_вт (x, v) & Увеличение_на (v, 100)
- «на сколько_%» - относительная оценка:
мощность упала на 20 % - МОЩНОСТЬ_вт (x, v) & Уменьшение_на_% (v, 20)

- «во сколько»:
мощность выросла в 1,5 раза - МОЩНОСТЬ_вт (x, v) & Увеличение_в (v, 1,5)

Б. Нечисловые группы.
1) Нормативно-оценочные:
большой мощности - МОЩНОСТЬ_вт (x, v) & Большая_величина (v)
2) Представляющие вербальную оценку динамику изменения:
мощность (резко) растет - МОЩНОСТЬ_вт (x, v) & Увеличение (v)

Приведенное выше и есть перечень тех основных различений, которые должен уметь проводить алгоритм анализа, имея дело с количественными группами.

2. Основные задачи, решаемые анализатором
1) Преобразование вербальных и вербально-цифровых значений в числовой формат (с предварительным восстановлением сокращенных обозначений элементов числа):
тысяча сто □ 1100; 10 млн. - 10 000 000 и т. д.
2) Интерпретация именованного числа как значения признака; пересчет значения к стандартной единице измерения:
10 квт - 10 000 вт (мощность)
3) Разграничение ***величины, количества и парциальной оценки***:
*20 кг **vs** 20 человек **vs** треть населения*

В первом случае должен быть построен признак *ВЕС_г (x, v) & v = 100,*
во втором – выражение *ЧЕЛОВЕК (x) & ЧИСЛЕННОСТЬ_СОВОКУПНОСТИ (x, v) & v=100,*
в третьем - выражение *НАСЕЛЕНИЕ (x) & ДОЛЯ_СОВОКУПНОСТИ (x, v) & v = 0,33*
4) Присваивание признаку значения; уточнение наименования признака:
толщиной 100 мкм (первоначально восстановленный по единице измерения обобщенный признак *линейный размер* уточняется как *толщина*).
5) Устранение смысловой избыточности. Учет лексически опосредованных связей между элементами количественных групп:
Финансовая помощь ЕС Литве в 2004-2006 гг. планируется на уровне 2,5 млрд. евро (Здесь число является оценкой признака *финансовая помощь,* но не признака *уровень*)
6) Прикрепление количественной группы к имени объекта: (например, связь между ИГ *жесткие диски* и количественной группой *емкостью до 100 ГБ*, интерпретируемая на уровне ЯПЗ как конъюнктивная связь.)

В некоторых ситуациях это может представлять нетривиальную проблему, что можно проиллюстрировать сопоставлением следующих двух фраз:
*(а) Возьмите деревянный брусок с отверстием **диаметром 30 мм**.*
*(б) Возьмите деревянный брусок с отверстием **весом 300 г**.*
Разрешение такого рода коллизий требует не только определенной различительной силы алгоритмов, но, прежде всего, специальной поддержки со стороны концептуального словаря.

3. Алгоритмический аспект
В разработанной нами системе анализа все названные алгоритмические задачи (и не только эти) решаются внутри единого переборного механизма, реализующего все процедуры семантической интерпретации. На вход интерпретатора подается синтаксически размеченный текст. Наличие разметки наиболее критично для тех фрагментов входного текста, которые, как в случае количественных групп, должны получить точную интерпретацию. Однозначность разметки не предполагается. Разрешение синтаксической и оставленной синтаксисом лексической омонимии производится единообразно путем перебора и оценки результатов интерпретации всех имеющихся вариантов. Выбор процедуры интерпретации управляется семантическими характеристиками синтаксического хозяина и синтаксического слуги.

4. Словарная поддержка процедур анализа
Функциональность и структуру концептуального словаря (онтологии), востребованную процедурами анализа можно кратко охарактеризовать следующим перечнем: связи *признак – единицы измерения*; *стандартные – нестандартные единицы измерения* (с указанием алгоритма пересчета); *признак – релевантный класс объектов*; наличие функциональных термов, характеризующих всю номенклатуру возможных значений (при этом нужно позаботиться о достаточно полном отражение лексических способов их выражения в толковом словаре системы). Так, для одного только "функционального" смысла *(очень) большая величина* нужно иметь весьма обширный список текстовых эквивалентов: *большой, весомый, высокий, крупный, немалый, астрономический, безграничный, бесчисленный,*

*большущий, великий, гигантский, гипертрофированный, головокружительный, грандиозный, громадный, здоровенный,…*и т.д., и т.п.

5. Вопросы реализации
В фактографической системе логическое представление, как правило, упрощается до сетевой нотации. (Семантическую сеть можно интерпретировать как логический язык с существенными ограничениями допустимых правил построения высказываний.)
В сетевой нотации различаются объектные и предикатные узлы; они соединяются ролевыми дугами, либо дугами, представляющими типовые отношения предметной области (*часть-целое, локализация, предмет-функция* и т. п.). Каждый узел характеризуется множеством терминов, разделяемых на два подкласса - термины-свойства и термины, представляемые парой < признак, значение >.
Количественные группы обычно характеризуют именно объектные узлы (хотя возможны и признаки, характеризующие предикатные узлы – процессы (*скорость, длительность* и т.п.)
В используемой нами нотации рабочим языком представления является табличный язык стандартной РСУБД. Семантическая сеть представляется тремя основными таблицами. Все термы, опознанные в тексте и релевантные целевой системе знаний, отображаются в ***таблице термов***; отнесение терма к тому или иному узлу маркируется в поле *Номер узла*, используемом как эквивалент референциальных индексов в логической записи.
Вторая таблица хранит ***значения признаков***. В ней представлены все значения (в частности, количественные).
Понятно, что помимо собственно значения, таблица должна, во-первых, специфицировать и его тип – в номенклатуре и различиях, приведенных в разделе 1. И, во-вторых, связывать это значение с определенным наименованием признака, записанном в *таблице термов*. Еще одна таблица должна представлять (именованные) связи между узлами сети.
Словарь и собственно анализатор ориентированы на анализ как количественной, так и чисто вербальной информации и реализованы средствами *CBuilder 6.0.* Текущее состояние – тестирование разработчиком на текстах свободного стиля при одновременном пополнении словарей. Результативность анализа, разумеется, полностью зависит от наличия в словарях соответствующей лексики и от точного соответствия ее описания принятым в модели анализа спецификациям. При этих условиях анализ гарантированно успешен, так что многократное тестирование однотипных конфигураций, в сущности, лишено смысла. Более осмысленный подход состоит в поиске в потоке текстов лексико-грамматических конфигураций не охваченных алгоритмами анализа.

(Труды международной конференции «Диалог 2006»
РАСПОЗНАВАНИЕ КОЛИЧЕСТВЕННОЙ ИНФОРМАЦИИ В ЕЯ- ТЕКСТАХ
В. Ш. Рубашкин (vrub@mail.nw.ru), Б. Ю. Чуприн (boris@vr4591.spb.edu) СПбГУ, С.-Петербург)

6.4. Обработка фактографической информации – практика

Особенности аналитической обработки
Первичная аналитическая обработка в фазе ETL требует значительных вычислительных ресурсов. Опыт эксплуатации систем с объемом фондов 5-10 млн. единиц хранения показывает, что если объем входных документов и время построения индекса принять за единицу, а запросы дополнительной памяти на диске и времени, требуемые на каждом из следующих видов обработки, как dV и dT соответственно, то получается следующая картина:

- выполнение индексирования: dV = 0,3-2, dT = 1;
- построение семантической сети: dV = 0,2-0,4, dT = 2-3;
- построение рубрик: dV = 0,001, dT = 0,1;
- создание аннотации и ключевых тем: dV = 0,1, dT = 1-2;
- терминологические векторы документов: dV = 0,1, dT = 0,02;
- хранилище аналитических данных: dV = 0,3, dT = 0,5;
- база данных фактографической информации, объединенной в досье: dV = 0,3, dT = 3.

Объем вторичных данных может быть в 3-4 раза больше объема документов, а время, необходимое на извлечение новых знаний, больше времени индексирования в семь–девять раз.

В ходе аналитической обработки происходит выделение текста фактографической информации об объекте, причем с учетом всех ссылок. Для этого сначала выделяются все предложения с упоминаниями об объекте (создается дайджест), в которых могут встречаться названия объекта («Иванов»), ссылки на него (анафория: «он», «который»…), а также обобщающие определения (кореференты: «воин»,

«семьянин»…). Нахождение и разрешение кореферентов и анафор дает увеличение объема дайджеста на 15-30%, а значит, и объема фактографической информации.

В начале исследования аналитики в первую очередь стремятся к полноте запроса, а не к его точности, поэтому объем релевантной подборки документов составляет сотни или тысячи единиц. Дальнейшее исследование проблемы производится уже после получения подборки документов с помощью кластерных, семантических карт или других методов. Такая технология работы аналитика сегодня типична как для работы в Internet, так и при работе со специализированными системами. Русский язык плохо поддается описанию формализмами различных уровней: морфологией, синтаксисом, семантикой. Например, для идентификации морфологических признаков лексемы на русском языке необходимо выполнить также предсинтаксический анализ предложения для снятия омонимии. В любом случае реализации этих формализмов используют нечеткую модель анализа текста.

К наиболее актуальным направлениям извлечения знаний из текста на сегодняшний день относятся:
- аналитическая обработка фактов; ведение досье;
- извлечение и структурирование фактографической информации;
- поиск информации по запросам на естественном языке с использованием тезаурусов;
- направления поиска информации, объектов в хранилище документов, в подборке документов;
- аннотирование документов, построение дайджестов по объектам;
- проведение тематического анализа документов (кластеризация и рубрицирование);
- построение и динамический анализ семантической структуры текстов;
- выделение ключевых тем и информационных объектов;
- определение общей и объектной тональности сообщений;
- исследование частотных характеристик текстов.

Технологии формирования досье
При коллективной работе зачастую несколько фактов вводятся в один атрибут одного объекта, после чего возникает необходимость в экспертной оценке достоверности введенных (возможно, противоречивых) фактов. Для этого в базе досье хранится дополнительная информация, подтверждающая факты в форме цитат из документов, а также прикрепленных к факту документов, почтовых сообщений, заключений экспертов, видеофрагментов и графических файлов. Каждый факт в системе имеет статус достоверный или недостоверный. На основе дополнительной подтверждающей информации из базы данных эксперт может принять решение об изменении статуса факта либо его удалить.
В системе должен быть реализован трекинг фактов — для любого факта пользователи имеют возможность вводить и просматривать комментарии и фрагменты контента, а также сами информационные объекты.

Технология **пакетного формирования досье** весьма актуальна в компаниях, имеющих распределенную систему офисов, каждый из которых может порождать информацию, например о действиях конкурентов в их регионе. При этом рыночная политика формируется в центральном офисе на основании в том числе досье на конкурентов. Для разметки удаленно сформированных сообщений, содержащих новые факты об объектах мониторинга, используется язык XML. Он удобен по нескольким причинам. Во-первых, состав атрибутов для каждого типа досье постоянно изменяется. Во-вторых, необходимо обеспечить возможность ввода новых типов досье. Встроенные в шаблон средства контроля над целостностью документа позволяют передавать только правильные факты. Автоматический ввод поступающих фактов производится с помощью программы-агента. Она выполняет мониторинг поступления новых сообщений, анализ корректности и структурный разбор XML-сообщения, формирование списка фактов, содержащихся в сообщении, и ввод фактов в базу данных.

7. Список инструментов Semantic Web

http://www.semantictools.ru/tools/tools_list.shtml

В этом разделе собраны ссылки на инструменты для работы с RDF и OWL. Список, конечно, не претендует на полноту, так как Semantic Web развивается очень быстро и новые средства появляются буквально каждый день.

7.1. Среды разработки, редакторы, системы управления контентом

Adobe's XMP

Adobe's XMP технология позволяющая встроить данные о файле как метаданные в сам файл.

Altova's SemanticWorks

Altova's SemanticWorks 2006 визуальный редактор RDF/OWL от создателей XMLSpy.

Amilcare

University of Sheffield's Amilcare адаптивный инструмент для извлечения информации, поддерживает метаданные Semantic Web

Arity's LexiLink

LexiLink - инструмент для создания словарей и онтологий, и управления ими в одном Веб-приложении уровня предприятия. Технология основана на RDF и OWL.

Cerebra Server

Cerebra Server - технологическая платформа используемая предприятиями для создания основанных на моделях приложений и высоко-адаптивной инфраструктуры интеграции информации.

Cypher

Cypher генерирует .rdf (RDF граф) и .serql (SeRQL запросы) представление исходного текста на естественном языке, позволяя пользователям делать запросы к базам данных на обычном языке.

DERI Ontology Management Environment (DOME)

DOME включает в себя инструменты для Редактирования и Просмотра, Управления Версиями и Развития, а также Отображения и Слияния, поставляется в виде свободно комбинируемых плагинов к Eclipse.

Graphl

Graphl - инструмент для совместного редактирования и визуализации графов RDF.

GrOWL

GrOWL - графический браузер и редактор OWL онтологий, может использоваться как автономное приложение или встраиваться в веб-браузер.

IBM's IODT

IODT, инструментарий от IBM для разработки управляемой онтологиями.

IBM's Web Ontology Manager

IBM's Web Ontology Manager - легкий, основанный на интернет-технологиях инструмент для управления онтологиями выраженными на Языке Веб Онтологий (OWL).

IBM Semantic Layered Research Platform

IBM SLRP - семейство открытых (open-source) программных компонентов Семантического Веба включающее в себя репозиторий RDF, инструмент для формирования запросов, среду для веб-приложений, библиотеки RCP, и много другое.

Intellidimension's RDF InferEd

Intellidimension's RDF InferEd - мощная среда, которая дает вам средства для навигации и редактирования RDF (Resource Description Framework) документов.

IsaViz

IsaViz - инструмент для просмотра и разработки моделей RDF представленных в виде графов.

Language & Computing's LinKFactory

Language & Computing's LinKFactory инструмент управления онтологиями, обеспечивает эффективный, ориентированный на пользователя способ создания, поддержки и развития исчерпывающих многоязычных терминологических систем и онтологий (Английский,

Испанский, Французский и др.). Инструмент спроектирован для создания, управления и поддержки больших, сложных и независимых от языка онтологий.

Metatomix M3t4.Studio Semantic Toolkit

M3t4.Studio Semantic Toolkit - свободно распространяемый набор плагинов к Eclipse для создания и управления OWL онтологиями и RDF документами.

Model Futures OWL Editor

Model Futures OWL Editor свободно распространяемый инструмент, простой в использовании и установке. Предлагает простой пользовательский интерфейс и может работать с очень большими OWL файлами. Также имеет средства для импорта XMI, Thesaurus Descriptor, и ErWin(TM), а также может экспортировать онтологии как документы MS Word(TM).

OpenLink's Data Spaces Platform

OpenLink Data Spaces (ODS) - распределенная платформа для приложений для использования средств Семантической Сети совместно с приложениями Web 2.0, такими как: Блоги, Вики, Агрегаторы RSS, Менеджеры закладок, Форумы, Галереи Фотографий, Социальные Сети, и т.д. Обеспечивает прозрачный доступ к данным приложения через встроенную поддержку SPARQL и использование онтологий таких как SIOC, FOAF, и Atom OWL. ODS - приложение OpenLink Virtuoso, и доступна и как Open Source, и в коммерческом варианте.

OWL verbalizer

OWL verbalizer - on-line инструмент, который вербализирует OWL онтологии на (ограниченном) английском языке.

pOWL

pOWL предоставляет PHP и основанное на веб-технологиях решение для редактирования и управления онтологиями.

Profium's Semantic Information Router

Profium's Semantic Information Router (SIR) - система управления контентом использующая стандартные метаданные, что улучшает повторное использование информации и позволяет пользователю обрабатывать и распространять дальше информацию собранную из нескольких источников и в разных форматах.

RDFe

RDFe - построенный на Schema редактор RDF, основан на pyrple.

Semantic Web Client

Semantic Web Client Library - представляет всю Семантическую Паутину как единый граф RDF. Библиотека позволяет приложениям делать запросы к этому глобальному графу используя SPARQL. Для поиска ответа на запрос библиотека динамически извлекает информацию из Семантической Сети разрешая HTTP URIs и следуя ссылкам rdfs:seeAlso. Библиотека написана на Java и базируется на среде Jena.

Siderean's Seamark Navigator

Siderean's Seamark Navigator - обеспечивает мощные средства для просмотра всей информации вашего предприятия вами и вашими клиентами. Страницы поиска по Интернету могут быть объединены с базами данных ваших продуктов, с серверами документов, и с другой цифровой информацией как изнутри, так и извне компании. Также предоставляется SPARQL API для получения данных непосредственно.

Software AG's Enterprise Information Integrator (EII)

EII версии 2.1 - глобально доступный продукт информационной интеграции, который использует технологии Семантического Веба. Динамически объединяя смысл и контекст бизнес данных с правилами которые им управляют, Enterprise Information Integrator обеспечивает руководителей бизнеса ресурсами для более быстрого принятия решений основанных на текущей информации. См. пресс-релиз для дополнительной информации

Stanford's Protégé

Stanford University's general Protégé 2000 - редактор онтологий, имеет архитектуру основанную на плагинах, что обеспечивает разработку целого ряда Semantic Web инструментов. Например: OWL плагин (называется Protégé-OWL) для редактирования RDF и OWL онтологий, а также SWRL правил, визуальный редактор для OWL (называется OWLViz), хранилище для Jena и Sesame, а также OWL-S плагин, который обеспечивает некоторые специфические возможности для редактирования OWL-S описаний Web-сервисов.

SWOOP

SWOOP от University of Maryland - основанный на Гипермедиа редактор OWL онтологий

Teranode's Experiment Design Automation

Teranode's Experiment Design Automation (XDA) - мощная платформа, которая позволяет ученым автоматизировать лабораторные эксперименты и управлять данными внутри и между лабораториями, для увеличения скорости и качества проектов R&D.

Thetus Publisher

Thetus обеспечивает программную инфраструктуру для моделирования и получения знаний, что позволяет организациям описывать, структурировать, искать, связывать, моделировать, разделять, и повторно использовать информацию независимо от схем и устройств.

Top Quandrant's TopBraid Composer

Top Quandrant's TopBraid Composer - полная основанная на стандартах платформа разработки, тестирования и сопровождения приложений Семантического Веба. Также реализует RDFa и GRDDL.

VisualKii

VisualKii - многоцлевая платформа визуального программирования основанная на Java. Содержит библиотеки для обработки моделей RDF, N3 и N-TRIPLES с помощью визуального определения потока данных и установки шагов обработки. Также включает поддержку запросов SPARQL.

@Semantics' Enterprise Information Integration

@Semantics Enterprise Information Integration (EII) - интегрированный подход к управлению данными. Подход EII полностью основан на открытых стандартах, и использует RDF/S для описания информации.

7.2. RDF репозитории

Aduna Metadata Server

Aduna Metadata Server - автоматически извлекает метаданные из источников информации, таких как файловые серверы, интранет или общедоступные веб-сайты. Aduna Metadata Server - мощное и хорошо масштабируемое хранилище метаданных. Metadata Server базируется на сервере Sesame.

Boca

Boca - RDF репозиторий уровня предприятия основан на Java и клиентских библиотеках, которые реализуют хранилище RDF, управление доступом, поддержку версий, репликацию и локальное сохранение данных для автономного доступа, и уведомления (события) для распределенных клиентов. Boca является частью IBM Semantic Layered Research Platform (SLRP).

D2RQ и D2R Server

D2RQ - библиотека Java, которая обеспечивает доступ к содержимому реляционных баз данных через SPARQL, Jena API, и Sesame API. D2R Server - SPARQL и RDF сервер на базе D2RQ.

Dojo Data

Dojo.data - Dojo JavaScript модуль, который включает хранилище RDF (dojo.data.RdfStore).

Franz Inc's AllegroGraph

AllegroGraph - система для загрузки, хранения и обеспечения доступа к RDF данным. Она включает SPARQL интерфейс и систему логического вывода RDFS. Она имеет Java и Prolog интерфейсы.

Intellidimension's RDF Gateway

Intellidimension's RDF Gateway - база данных RDF Троек с системой логического вывода RDFS и SPARQL интерфейсом.

Jena's Joseki

Jena's Joseki layer предлагает средства хранения RDF Троек с SPARQL интерфейсом (см. также Jena)

Kowari

Kowari Metastore - открытая (Open Source), хорошо масштабируемая, безопасная с точки зрения транзакций, специализированная база данных для хранения RDF, написанная на Java. Kowari не поддерживается с декабря 2005. См. Mulgara.

Mulgara

Mulgara Semantic Store - открытая (Open Source), хорошо масштабируемая, безопасная с точки зрения транзакций, специализированная база данных для хранения RDF, написанная на Java. Она пришла на смену Kowari.

OpenLink Virtuoso

Virtuoso SQL-ORDBMS и Web Application Server гибрид (Универсальный Сервер), обеспечивает управление данными SQL, XML, и RDF в одном многопоточном серверном приложении. Доступ к Хранилищу обеспечивается через: SPARQL, SIMILE Semantic Bank API, ODBC, JDBC, ADO.NET, XMLA, WebDAV, и Virtuoso/PL (SQL Stored Procedure Language). Продукт доступен как Open Source или как коммерческое приложение

Oracle Spatial 10g

Oracle Spatial 10g включает открытую, масштабируемую, безопасную и надежную платформу. Основана на графах, RDF тройки сохраняются, индексируются и запрашиваются аналогично другим объектно-ориентированным типам данных. База данных Oracle 10g RDF гарантирует, что разработчики приложений выиграют от масштабируемости Oracle 10g

OWLIM

OWLIM - высоко-производительный семантический репозиторий, упакованный как Storage and Inference Layer (SAIL) для базы данных RDF Sesame.

RDFStore

RDFStore - RDF репозиторий с поддержкой Perl и C API, и SPARQL доступа.

RAP's RDF сервер

RDF сервер среды PHP RAP.

SemWeb для .NET

SemWeb поддерживает хранилища в MySQL, Postgre, и Sqlite; протестирован на наборах в 10-50 миллионов троек; поддерживает SPARQL.

Sesame

Sesame - это открытая (open source) база данных RDF с поддержкой для логического вывода RDF Schema и запросов. Она предлагает большой набор инструментов для разработчиков для использования RDF и RDF Schema.

Tucana Suite

Northrop Grumman's Tucana Suite - версия Kowari Metastore промышленного уровня качества.

YARS

YARS (Yet Another RDF Store) (Еще одно хранилище RDF) - хранилище RDF на Java с поддержкой запросов RDF основанных на декларативном языке запросов, которое предлагает некий более абстрактный уровень, чем обычные API.

3Store

3Store - хранилище троек на базе MySQL. Сам сервер не предлагает пользователю непосредственного интерфейса, но к нему можно делать запросы используя несколько сервисов, а том числе column based view и непосредственный браузер RDF

7.3. API

Среды поддерживающие множество языков

Euler

Euler - механизм логического вывода. Существуют реализации на Java, C#, Python, Javascript и Prolog. Через N3 может взаимодействовать с W3C Cwm.

Redland RDF Application Framework

The Redland RDF Application Framework - набор бесплатных программных библиотек обеспечивающих поддержку RDF. Он предоставляет парсеры для RDF/XML, Turtle, N-triples, Atom, RSS; имеет SPARQL и GRDDL реализации, и имеет интерфейсы на C#, Python, Obj-C, Perl, PHP, Ruby, Java и Tcl

Java

Corese

Corese Corese расшифровывается как Conceptual Resource Search Engine. Это движок RDF основанный на Conceptual Graphs (CG). Он обеспечивает обработку RDF Schema и RDF выражений в рамках CG формализма, обеспечивает движок основанный на правилах и механизм запросов понимающий SPARQL синтаксис.

DartGrid

DartGrid - Среда для разработки приложений на Java для интеграции гетерогенных реляционных баз данных с использованием технологий Семантического Веба.

Jena

Jena Java RDF API и инструментарий - инфраструктура для конструирования приложений Семантической Сети. Обеспечивает программную среду для RDF, RDFS и OWL, SPARQL и

включает систему логического вывода. Также может быть использована как база данных RDF через подсистему Joseki. См. список обсуждений jena для дополнительной информации.

JRDF

JRDF Java RDF Binding - попытка создать стандартный набор API и базовые реализации для RDF на Java.

OWLJessKB

OWLJessKB - система логического вывода для OWL. Семантика языка реализована с использованием Jess (Java Expert System Shell). В настоящее время реализовано большинство возможностей OWL lite, плюс некоторые еще и минус некоторые.

RDFSuite

ICS-FORTH RDFSuite открытые (open source), хорошо масштабируемые инструменты для Семантического Веба. Этот набор включает Validating RDF Parser (VRP), RDF Schema Specific DataBase (RSSDB) и поддерживает RDF Query Language (RQL).

YARS

YARS (Yet Another RDF Store) - хранилище данных RDF на Java, обеспечивает запросы к RDF основанные на декларативном языке запросов, который предлагает более высокий уровень абстракции чем другие API.

См. также Kowari и Allegro системы, и Euler engine

Python

CWM

Closed World Machine (CWM) манипулятор данными, обработчик правил и система запросов в основном использующая Notation 3 текстовый синтаксис RDF. Также имеет не полную реализацию OWL Full и SPARQL.

pyrple

pyrple парсер для RDF/XML, N3, и N-Triples. Имеет хранилище в памяти с запросами на уровне API, экспериментальный маршаллинг, множество утилит, небольших и в минимальной степени зависимых друг от друга. Поддерживает тесты на изоморфность графов, и многое другое.

RDFLib

RDFLib, RDF библиотека для Python, включает SPARQL API. Библиотека также содержит хранилище для графов в памяти и на диске.

4Suite 4RDF

4Suite 4RDF открытая (open-source) платформа для обработки XML и RDF реализованная на Python с расширениями на C.

См. также Euler engine и Redland Framework

C

См. RDFStore и Redland Framework

C++

Brahms

Brahms быстрое RDF/S хранилище в оперативной памяти, способное хранить и обеспечить доступ к большим онтологиям. Реализовано как набор C++ классов.

C# и .Net

Drive

Drive - RDF парсер написанный на C# и платформе .NET

SemWeb

SemWeb - RDF библиотека на C# с поддержкой RDBMS хранилища, чтения/записи XML и N3, SPARQL, и механизма логического вывода RDFS.

См. также Euler engine и Redland Framework

Javascript

AJAX Клиент для SPARQL

AJAX Клиент для SPARQL - простой AJAX клиент, который может быть использован для запуска SELECT запросов к сервису и для последующей их интеграции с Javascript кодом на стороне клиента.

Dojo Data

Dojo.data - Dojo Javascript модуль с поддержкой хранилища RDF (dojo.data.RdfStore).

Javascript RDF/Turtle парсер

Javascript RDF/Turtle парсер, может быть использован совместно с Jibbering

Jibbering

Jibbering - простой Javascript RDF Парсер и средство формирования запросов.

RDFParser

RDFParser - парсер реализующий в полном объеме RDF/XML, может быть использован с браузерами поддерживающими DOM Level 2

SPARQL JavaScript Library

SPARQL JavaScript Library реализует SPARQL Protocol и интерпретацию возвращаемых значений как часть AJAX среды.

См. также Euler engine

Tcl/Tk

См. также Redland Framework

PHP

ARC

ARC - легкая RDF система с поддержкой SPARQL для реализации мэйнстримных веб-проектов. Написана на PHP и оптимизирована для совместно используемых веб-сред.

RAP

RAP - PHP пакет для манипулирования RDF моделями с поддержкой постоянного хранилища RDF/XML данных. Включает интегрированные парсеры для RDF/XML, n3, n-triple, TriX, GRDDL, и RSS, движок запросов SPARQL и клиентскую библиотеку SPARQL и интегрированный RDF сервер.

См. также Redland Framework

Lisp

Wilbur

Wilbur Lisp инструментарий для программирования приложений Семантического Веба. Wilbur - инструментарий от Nokia Research Center и использует RDF написанные на Common Lisp.

Obj-C

См. также Redland Framework

Prolog

dlpconvert

dlpconvert - инструмент для преобразования Horn фрагментов OWL (называемых DLP) из XML или RDF синтаксиса в Prolog (см. также Kaon2)

SWI-Prolog

SWI-Prolog - всеобъемлющая Prolog среда, которая также включает хранилище RDF Троек. Также имеется отдельная Prolog библиотека для обработки OWL.

См. также Euler engine и систему AllgroGraph.

Perl

CARA

CARA - RDF API написанный на Perl. CARA опирается на модель графов RDF и поддерживает хранилища RDF графов как в памяти, так и на диске. Парсер RDF также входит в API.

См. также RDFStore и Redland Framework

Ruby

ActiveRDF

ActiveRDF - библиотека для доступа к данным RDF из программ на Ruby. Может быть использована как уровень данных в Ruby-on-Rails. Вы можете обращаться к ресурсам RDF, классам, свойствам, и т.д. программно, без формирования запросов.

См. также Redland Framework

Haskell

Swish

Swish - среда для выполнения дедуктивных логических операций над данными в RDF. По возможностям напоминает CWM.

Weso

Weso - набор инструментов Семантического Веба разработанный как часть курса Декларативное Программирование в Университете Овиедо.

7.4. Механизмы логического вывода OWL

Bossam

Bossam - основанный на правилах механизм логического вывода OWL (бесплатный, хорошо документированный, с закрытыми исходниками).

FaCT++

FaCT++ - механизм логического вывода OWL DL реализованный на C++.

KAON2

KAON2 - инфраструктура для управления онтологиями OWL-DL, SWRL, и F-Logic. Запросы могут быть сформулированы на SPARQL.

Pellet

Pellet - механизм логического вывода OWL DL с открытыми исходниками написанный на Java. Может быть использован совместно с Jena или библиотеками OWL API. Также может быть включен в состав других приложений.

RacerPro

RacerPro - механизм логического вывода OWL и сервер для Семантической Паутины.

7.5. Генераторы RDF

Cypher

Cypher генерирует RDF и SeRQL представление предложений и фраз на естественном языке.

FOAF-o-matic

FOAF-o-matic - онлайновый генератор FOAF.

7.6. On-line Валидаторы

BBN OWL Validator

BBN OWL Validator

OWL Consistency checker

OWL Consistency checker (based on Pellet)

WonderWeb OWL-DL Validator

WonderWeb OWL-DL Validator

W3C's RDF Validator

W3C's RDF Validator

RDF/XML and N3 Validator

rdfabout.com's Validator

VIStology's ConsVISor OWL Consistency checker

ConsVISor

7.7. Серверы запросов SPARQL

SPARQLer

SPARQLer; см. описание.

SPARQLette

Демонстрационный сервис запросов SPARQL

XML Army Knife

XML Army Knife; см. описание.

OpenLink Virtuoso

Live SPARQL Query Service Endpoint ; см. http://demo.openlinksw.com/sparql_demo для деталей и примеров использования (в том числе и удаленных запросов к другим сервисам запросов SPARQL).