ОБРАБОТКА ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА В МОДЕЛЯХ ПОИСКОВЫХ СИСТЕМ

В.В. Диковицкий, М.Г. Шишаев

Введение

Одной из основных функций современных информационных систем (ИС) является поиск элементов данных, удовлетворяющих некоторым признакам (информационный поиск). Специфика методических и технологических проблем, возникающих при организации такого поиска, обусловливается несколькими факторами. Прежде всего, это - характер контента, содержащегося в информационных ресурсах, входящих в систему. В современных ИС попрежнему преобладает текстовый контент, однако все большее распространение приобретают мультиформатные ресурсы, содержащие мультимедиаконтент (графика, аудио и видео информация), а также использующие для повышения эффективности функционирования различные формы структуризации контента. Результатом структуризации становится деление информации на собственно данные, мета-данные, описывающие их структуру, и даже "мета-мета-данные", определяющие различные варианты структур данных. Такие особенности контента, в явном или неявном виде, определяют подходы к организации эффективного поиска информации в рамках соответствующего набора ресурсов.

Еще одним важным обстоятельством, оказывающим существенное влияние на эффективность механизмов поиска информации, является распределенный и, как следствие, гетерогенный характер современных информационных ресурсов и систем. Ориентированные на использование в условиях однородных информационных систем и ресурсов механизмы поиска (например, на базе простых индексов) резко теряют свою эффективность в применении к распределенным гетерогенным системам, где форматы представления данных и, соответственно, метаданные отличаются от ресурса к ресурсу или от системы к системе. Это обстоятельство заставляет исследователей и разработчиков искать пути создания универсальных методов и технологий информационного поиска, адекватных требованиям современных информационных систем.

Текст является одной из основных форм обмена информацией в обществе. Текстовая информация в различных форматах составляет значительную долю информационных ресурсов информационных систем. Поэтому создание и развитие технологий обработки текста привлекали большое внимание на всех этапах развития информационных систем. Наиболее распространенными системами этой категории являются системы текстового поиска, задача которых заключается в поиске по заданной коллекции документов на естественном языке (ЕЯ) документов, удовлетворяющих информационным потребностям пользователей. В данной работе представлены основные принципы текстового поиска, методы обработки естественного языка и их использование в моделях поиска.

Методы информационного поиска и обработка текстов на естественном языке

Значительное место в технологиях текстового поиска занимает обработка ЕЯ. Под обработкой ЕЯ (Natural Language Processing, NLP) понимается решение задач, связанных с пониманием, анализом, выполнением различных операций над текстами, а так же их генерацией [6]. Примеры подобных задач: классификация, кластеризация хранимых коллекций документов, глубинный анализ текстов, перевод документов с одного языка на другой и т.д.

Все многообразие методов информационного поиска основываются на обработке и анализе текстов индексируемых документов*. Большинство ИПС являются системами с предпроцессингом - предварительной обработкой (индексированием) всех имеющихся в системе документов. Исключения составляют метапоисковые системы [9]. Перечислим основные трудности, возникающие при обработке текстов на ЕЯ:

- проблема синонимии;
- проблема омонимии;
- устойчивые сочетания слов;
- морфологические вариации.

Проблема синонимии. Одно понятие может быть выражено различными словами. В результате релевантные документы, в которых используются синонимы понятий, указанных пользователем в запросе, могут быть не обнаружены системой.

Проблема омонимии и явлений «смежных с омонимией». Грамматические омонимы - разные по значению слова, но совпадающие по написанию в отдельных грамматических формах. Это могут быть

* Под документом подразумевается некий объект, содержащий информацию в зафиксированном виде. Документы могут содержать тексты на естественном или формализованном языке, изображения, звуковую информацию и т.д.

слова одной или разных частей речи. Лексические омонимы - слова одной части речи, одинаковые по звучанию и написанию, но разные по лексическому значению.

Устойчивые сочетания слов. Словосочетания могут иметь смысл отличный от смысла, который имеют слова по отдельности.

Морфологические вариации. Во многих естественных языках слова имеют несколько морфологических форм, различающихся по написанию.

Существующие ПС используют различные методы обработки текстов ЕЯ. В современных технологиях текстового поиска используется не только аппарат лингвистики для анализа текстов, но и статистические методы, математическая логика и теория вероятностей, кластерный анализ, методы искусственного интеллекта, а так же технологии управления данными. Рассмотрим два основных подхода к обработке и анализу текстов ЕЯ – *статистический и лингвистический* (рис.1).

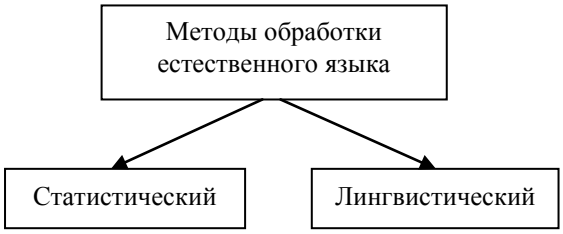

Рис.1. Методы обработки естественного языка

В основе *статистического подхода* лежит предположение, что содержание текста отражается наиболее часто встречающимися словами. Суть статистического анализа заключается в подсчете количества вхождений слов в документ. Распространенным является сопоставление каждому терму *t* в документе некоторого неотрицательного веса. Веса термов вычисляются множеством различных способов. Самый простой из них – положить «вес» равный количеству появлений терма *t* в документе *d*, обозначается *tft,d* (term frequency)[1]. Этот метод взвешивания не учитывает дискриминационную силу терма. Поэтому в случае, когда доступна статистика использования термов по коллекции, лучше работает схема *tf-idf* вычисления весов, определяемая следующим образом:

$$tf - idf_{i,d} = tf_{i,d} \times idf_i ,$$

где $idf_i = \log \frac{N}{df_i}$ - обратная документальная частота (inverse document frequency) терма *t* , *dft* - документальная частота (document frequency), определяемая как количество документов в коллекции, содержащих терм *t, N* - общее количество документов в коллекции. Схема *tf-idf* и ее модификации широко используются на практике.

Эффективным подходом, основанным на статистическом анализе, является латентно-семантическое индексирование. Латентно-семантический анализ – это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных [2]. 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.

Главный недостаток статистических методов состоит в невозможности учета связности текста, а представление текста как простого множества слов недостаточно для отражения его содержания. Текст представляет набор слов, выстроенных в определенной заданной последовательности. Преодолеть этот недостаток позволяет использование лингвистических методов анализа текста.

Существуют следующие уровни *лингвистического анализа*: графематический, морфологический, синтаксический, семантический. Результаты работы каждого уровня используются следующим уровнем анализа в качестве входных данных (рис. 2).

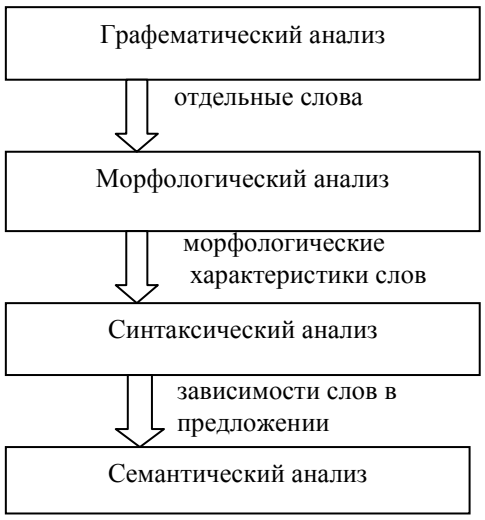

Рис.2. Уровни лингвистического анализа

Целью графематического анализа является выделения элементов структуры текста: параграфов, абзацев, предложений, отдельных слов и т. д.

Целью морфологического анализа является определение морфологических характеристик слова и его основной словоформы. Особенности анализа сильно зависят от выбранного естественного языка.

Целью синтаксического анализа является определение синтаксической зависимости слов в предложении. В связи с присутствием в русском языке большого количества синтаксически омонимичных конструкций, наличием тесной связи между семантикой и синтаксисом, процедура автоматизированного синтаксического анализа текста является трудоемкой. Сложность алгоритма увеличивается экспоненциально при увеличении количества слов в предложении и числа используемых правил.

Разработки в области семантического анализа текста связаны с областью искусственного интеллекта, делающей акцент на смысловом понимании текста. В настоящее время успехи в этом направлении достаточно ограничены. Разработанные семантические анализаторы обладают высокой вычислительной сложностью и неоднозначностью выдаваемых результатов [10].

Модели информационного поиска

В ходе развития информационно-поисковых систем было предложено множество моделей информационного поиска, далее рассмотрим основные.

Модель поиска – это сочетание следующих составляющих [6]:

1. Формат представления документов.
2. Формат представления запросов. Запрос – формализованный способ выражения информационных потребностей пользователя ИПС. Для этого используется язык поисковых запросов, синтаксис которых варьируется от системы к системе.
3. Функция соответствия документа запросу. Степень соответствия запроса и найденного документа (релевантность) – субъективное понятие, поскольку результаты поиска, уместные для одного пользователя, могут быть неуместными для другого.

В различных моделях ИПС вид критерия релевантности документов зависит от вида модели информационного поиска, например в моделях семантического поиска, точное вхождение слов запроса в документ не является основополагающим критерием, как, например, в теоретико-множественных моделях.

Вариации этих составляющих определяют множество реализаций систем поиска. Рассмотрим наиболее распространенные модели поиска.

Модели традиционного информационного поиска принято делить на три вида: теоретико-множественные (булевская, нечетких множеств, расширенная булевская), алгебраические (векторная, обобщенная векторная, латентно-семантическая, нейросетевая), вероятностные (рис.3).

Булевская модель – модель поиска, опирающаяся на операции пересечения, объединения и вычитания множеств. Запросы представляются в виде булевских выражений из слов и логических операторов. Релевантными считаются документы, которые удовлетворяют булевскому выражению в запросе. Основной недостаток булевской модели заключается в непригодности для ранжирования результатов поиска.

Векторная модель – представление коллекции документов векторами из одного общего для всей коллекции векторного пространства. Документы и запросы представляются в виде векторов в N-мерном евклидовом пространстве. Вес термина в документе можно определить различными способами. Например, можно подсчитать количество употреблений терма в документе, так называемую частоту терма, — чем чаще слово встречается в документе, тем больший у него будет вес. Если терм не встречается в документе, то его вес в этом документе равен нулю.

Все термы, которые встречаются в документах обрабатываемой коллекции, можно упорядочить. Если теперь для некоторого документа выписать по по , включая те, которых нет в этом документе, получится вектор, который и будет представлением данного документа в векторном пространстве.

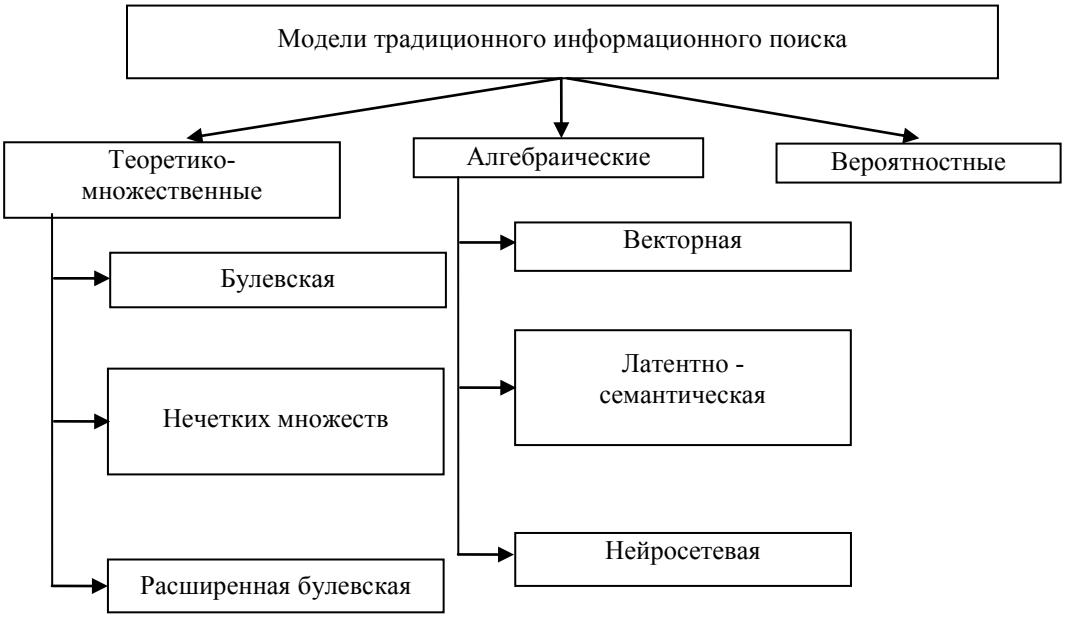

Рис.3. Модели традиционного информационного поиска

Размерность этого вектора, как и размерность пространства, равна количеству различных термов во всей коллекции, и является одинаковой для всех документов. Релевантность в данной модели выражается через подобие векторов. Для вычисления подобия векторов используется косинусная метрика. Учитывать частотные характеристики слов предложили в 1957 году Joyce и Needham, и в 1968 году векторная модель была реализована Джерардом Солтоном* в поисковой системе SMART (Salton's Magical Automatic Retriever of Text) [11]. Векторно-пространственная модель связана с расчетом массивов высокой размерности и малопригодна для обработки больших массивов данных.

В 1977 году Robertson и Sparck-Jones реализовали вероятностную модель [12]. Релевантность в этой модели рассматривается как вероятность того, что данный документ может оказаться интересным пользователю. При этом подразумевается наличие уже существующего первоначального набора релевантных документов, выбранных пользователем или полученных автоматически при каком-нибудь упрощенном предположении. Вероятность оказаться релевантным для каждого следующего документа рассчитывается на основании соотношения встречаемости терминов в релевантном наборе и в остальной, «нерелевантной» части коллекции. Вероятностная модель характеризуется низкой вычислительной масштабируемостью, необходимостью постоянного обучения системы.

Семантический поиск

Одно из перспективных направлений развития информационно-поисковых систем – построение моделей «семантического» поиска. Семантический поиск — вид автоматизированного полнотекстового информационного поиска с учетом смыслового содержания слов и словосочетаний запроса пользователя и предложений текстов проиндексированных информационных ресурсов. Семантический поиск, например, позволяет найти документы, вовсе не содержащие слов из поискового запроса, но имеющие к ней отношение. Попытки реализации семантического поиска начались в конце 20 века. В 2000 г. P. Vakkari [15] предложил способ поиска схожих по семантике документов на основе сопоставления их лексических векторов.

Существующие системы семантического поиска

В трудах Гавриловой Т.А., Хорошевского В.Ф. [17, 18] исследуется вопрос о применении онтологического подхода для информационного поиска. Онтологии являются методами представления и обработки знаний и запросов, и предназначены для описания семантики данных для некоторой предметной области и решения проблемы несовместимости и противоречивости понятий.

Онтологии обладают собственными средствами обработки (логического вывода), соответствующими задачам семантической обработки информации. Поэтому онтологии получили широкое распространение в решении проблем представления знаний и инженерии знаний, семантической интеграции информационных ресурсов, информационного поиска и т.д.

Определение онтологии дано в работе Gruber T.R «A Translation Approach to Portable Ontology Specifications»[13]: *эксплицитная, т.е. явная спецификация концептуализации, где в качестве концептуализации выступает описание множества объектов и связей между ними.*

В работе Wielinga B., Schreiber A.T., Jansweijer [14], сделана попытка дать математические определения понятий "модель концептуализации предметной области", "база знаний предметной области" и "модель онтологии предметной области".

Онтология определяет общий словарь для ученых, которым нужно совместно использовать информацию в предметной области. Она включает машинно-интерпретируемые формулировки основных понятий предметной области и отношения между ними.

В России информационно-поисковая система с использованием онтологии была впервые реализована авторами Добров Б.В., Лукашевич Н.В., Сыромятников С.В., Загоруйко Н.Г. в информаци-онно-поисковой системе УИС «РОССИЯ» (Университетская информационная система). Поступающие на вход информационной системы потоки документов подвергаются автоматической лингвистической обработке, включающей в себя следующие этапы: морфологический анализ, терминологический анализ, рубрицирование, аннотирование [4]. Терминологический анализ реализован на основе Тезауруса по общественно-политической тематике. На базе Тезауруса осуществляется автоматическое концептуальное индексирование входящего потока текстов и производится процедура разрешения многозначных терминов.

Основная проблема при реализации применении онтологического подхода - отсутствие достаточно больших и качественных онтологий предметных областей, особенно на русском языке.

Осипов Г.С. и соавторы предложили собственную модель семантического поиска, реализовав ее в информационно-поисковой системе «Exactus», в которой объединены статистические и лингвистические методы поиска. Из статистических характеристик текста Exactus учитывает TF*IDF веса термов и значимость фрагментов текстов (на основе HTML-разметки документов). Лингвистическая составляющая – значения синтаксем (минимальных семантико-синтаксических единиц текста) и их сочетаемость в конкретном предложении [5].

В теории коммуникативной грамматики [8] русского языка опровергается традиционное противопоставление синтаксиса семантике, которое

* Gerard Salton (Sahlman) 1927-1995 гг.

предполагает разделение знаний о законах формирования связной речи на два уровня: знания о форме (синтаксис) и знания о значении (семантика).

Основополагающая идея коммуникативной грамматики заключается в том, что синтаксис должен изучать именно осмысленную речь, а синтаксические правила должны учитывать категориальные значения слов, чтобы иметь возможность определять обобщенные значения любой синтаксической конструкции – от слова до словосочетания и простого предложения. Очевидно, что одних морфологических характеристик недостаточно, чтобы слово стало конструктивной единицей синтаксиса. Слово-лексема еще не является синтаксической единицей, слово – единица лексики, а в разных его формах могут реализоваться или актуализироваться разные стороны его общего значения. Таким образом, решающую роль здесь играет обобщенное значение, то есть категориально-семантический класс слова. Обобщенное значение определяет синтаксические возможности слова и способы его функционирования. Формируя и изучая связную речь, синтаксис имеет дело с осмысленными единицами, несущими свой не индивидуально-лексический, а обобщенный, категориальный смысл в конструкциях разной степени сложности. Эти единицы характеризуются всегда взаимодействием морфологических, семантических и функциональных признаков. Эти единицы получили название *синтаксем*. Важно подчеркнуть, что семантическое значение складывается в результате соединения категориального значения и морфологической формы, реализуется в определенной синтаксической позиции. Рассмотрение слова изолированно, в отрыве от текста, не позволяет установить синтаксическое значение, а следовательно – осуществлять семантический поиск.[8]

Методы семантического поиска в информационно-поисковой системе «Exactus» применяются к обработке текстов запросов пользователей и возвращаемых документов. Семантическая обработка включает в себя построение семантического поискового образа запроса, построение семантического образа документов и сравнение получившихся образов. В результате вычисляются дополнительные виды релевантности, позволяющие фильтровать документы, не соответствующие поисковому запросу в указанном понимании, т.е. отбирать только те тексты, в которых семантическое значение синтаксемы совпадает с ее семантическим значением в запросе (что невозможно в обычных статистических методах).

Заключение

Приведенные традиционные модели поисковых систем изначально предполагали рассмотрение документов как множества отдельных слов, не зависящих друг от друга. Вероятностная модель характеризуется низкой вычислительной масштабируемостью, необходимостью постоянного обучения системы. Наиболее распространенными являются алгебраические теоретико-множественные модели, т.к. их практическая эффективность обычно выше. Следует отметить, что предлагаемые в последнее время новые реализации проектов информационного поиска зачастую являются гибридными моделями и обладают свойствами моделей разных классов. Одно из перспективных направлений развития информационно-поисковых систем – построение моделей семантического поиска, основная задача которых заключается в анализе текста, т.е. извлечение смысла из текста и отображение его в формальную модель, которая позволяет находить смысловую близость двух текстов. Стоит признать, что потенциал у таких систем действительно большой, однако в настоящее время реализованы далеко не все возможные семантические технологии. По сути, сейчас они только помогают выделить ключевые слова из фраз, построенных на естественном языке и подобрать дополнительные словоформы для составления корректного поискового запроса. Данное направление методов поиска требует развития.

Литература

1. Brin, S. The Anatomy of a Large-Scale Hypertextual Web Search Engine / Sergey Brin, Lawrence Page// – Режим доступа: http://infolab.stanford.edu/pub/papers/google.pdf
2. Некрестьянов, И.С. Латентно-семантический анализ: Введение в латентно-семантический анализ. - Режим доступа: http://meta.math.spbu.ru/~igor/papers/lsa-prg/node2.html
3. Studer, R. Knowledge Engineering: Principles and Methods/ Studer R., Benjamins V.R., Fensel D. // In Data & Knowledge Engineering, 25, 1998. – P. 161 – 197.
4. Журавлев, С.В. УИС «РОССИЯ».Автоматическое тематическое индексирование полнотекстовых документов / С.В. Журавлев, Б.В. Добров //Материалы научно-практической конф. «Проблемы обработки больших массивов неструктурированных текстовых документов», 2001.
5. Осипов, Г.С. Семантический поиск в сети интернет средствами поисковой машины Exactus /Г.С. Осипов, И.А. Тихомиров, И.В. Смирнов. – Режим доступа: http://www.raai.org/cai-08/files/cai-08_exhibition_31.doc
6. Когаловский, М.Р. Перспективные технологии информационных систем / М.Р. Когаловский. –М.: Компания АйТи, 2003. – 288 с.
7. Baeza-Yates R. Modern Information Retrieval / R. Baeza-Yates , B. Ribeiro-Neto // ACM Press Series/Addison Wesley, New York, 1999. – 513 p.
8. Золотова, Г.А. Коммуникативная грамматика русского языка / Г.А. Золотова, Н. К. Онипенко, М.Ю. Сидорова //Институт русского языка РАН им. В.В. Виноградова. - М., 2004. – 544 с.

9. Тихонов, В. Архитектура метапоисковых систем – Режим доступа: http://www.cmsmagazine.ru/library/items/internet_info/metasearch/
10. Калиниченко, А.В. Сущность проблемы анализа текста в полнотекстовых поисковых системах. Подходы и пути решения. – Режим доступа: http://www.jurnal.org/articles/2010/inf12.html
11. Солтон, Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979.
12. Лифшиц, Ю. Модели информационного поиска. – Режим доступа: http://yury.name/internet/03ianote.pdf
13. Gruber, T.R. A Translation Approach to Portable Ontology Specifications. – Режим доступа: http://tomgruber.org/writing/ontolingua-kaj-1993.pdf.
14. Wielinga, B. Framework and Formalism for Expressing Ontologies / B. Wielinga etc.// ESPRIT Project 8145 KACTUS, Free University of Amsterdam Deliverable, DO1b.1, 1994.
15. Vakkary, P. eCognition and changes of search terms and tactics during task performance // RIAO'2000.
16. Гаврилова, Т.А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем / Т.А. Гаврилова // Новости искусственного интеллекта, 2003. – №2. – С. 24-30.
17. Гаврилова, Т.А. Использование онтологии в системах управления знаниями. – Режим доступа: http://big.spb.ru/publications/bigspb/kni/use_ontology_m_suz.shtml
18. Гаврилова, Т.А. Базы знаний интеллектуальных систем /Т.А. Гаврилова, В.Ф. Хорошевский. -СПб.: Изд-во «Питер», 2001. - 382 с.