М.А. Шумская (Минск, МГЛУ)

# ОСОБЕННОСТИ ФУНКЦИОНИРОВАНИЯ ИНТЕЛЛЕКТУАЛЬНОЙ ПОИСКОВОЙ СИСТЕМЫ

Современные информационно-поисковые системы, работающие с запросами пользователя в виде ключевых слов, характеризуются относительно неплохим уровнем полноты и точности результатов поиска. До организации процесса поиска лингвистический процессор должен «перевести» запрос пользователя с естественного языка на формализованный информационно-поисковый язык. В связи с этим ученые пытаются создать поисковые машины, которые были бы наделены изрядной долей интеллекта, что позволило бы и человеку, и компьютеру работать на естественном языке, а также решило бы ряд проблем, и значительно повысило производительность и качество работы поисковой системы. В настоящее время интеллектуальные поисковые системы могут решать следующие достаточно сложные задачи [1].

1. Автоматически определять язык запроса. Эта функция позволяет ограничить языковой сегмент Сети, в котором будет производиться поиск запрашиваемой информации, что с учетом количества информации в виртуальной среде, положительно сказывается на процессе поиска.

2. Исключать неинформативные слова (стоп-слова), т.е.. слова служебных частей речи, не несущих никакой смысловой нагрузки, либо некоторые наиболее общеупотребительные слова. Их удаление значительно сокращает объем индекса и увеличивает релевантность результатов поиска. Универсального перечня неинформативных слов не существует, так как он постоянно обновляется за счет добавления новых и исключения старых слов. Если пользователь вводит запрос, состоящий только из стоп-слов, о релевантности результатов поиска не может быть и речи. Поэтому при построении запроса пользователю нужно по возможности исключать смысловые операторы (*and, or, not*) или использовать вспомогательные средства поиска (например, символы ***!***, **–**, **~**, **:**, **|** ) и др.

3. Проводить лингвистический анализ исходного запроса пользователя, а также найденных текстовых документов. Лингвистический анализ включает следующие процедуры: 1) лексический (графематический) анализ; 2) морфологический анализ; 3) синтаксический анализ; 4) семантический анализ (см. рис.1) [2].

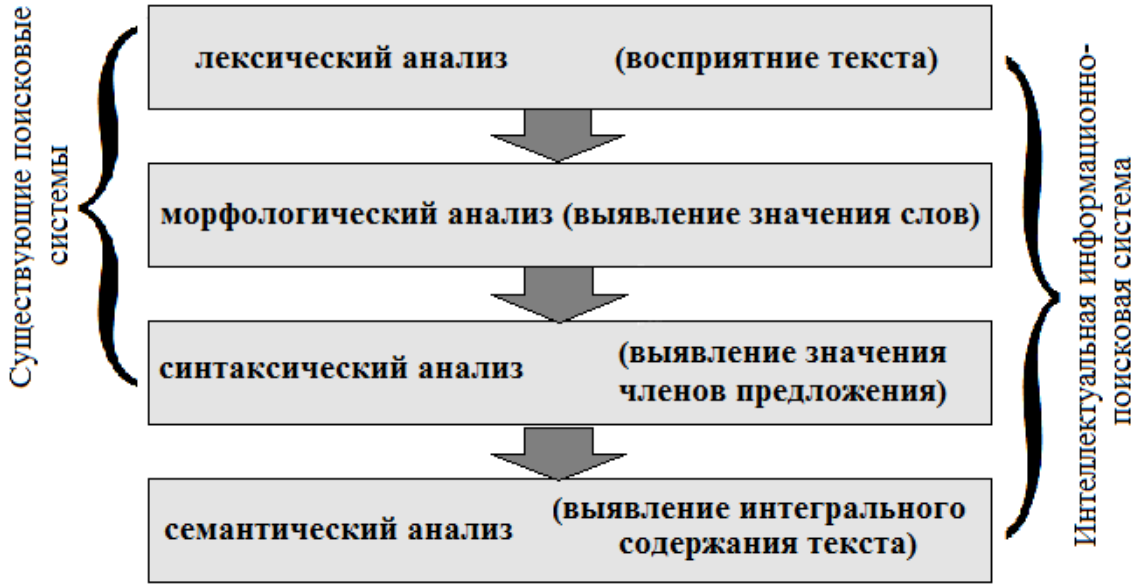


Рис. 1. Компоненты лингвистического анализа текстовой информации

Лексический анализ заключается в выявлении элементов структуры текста: параграфов, абзацев, предложений, отдельных слов и т.д. При этом проводится определение языка текста, типов содержащихся в нем предложений, и наличия лексических выражений (жаргонных слов и т.п.). Реализация такого типа анализа не вызывает затруднений. Морфологический анализ заключается в автоматическом распознавании морфологических характеристик каждого слова. При проведении данного типа анализа многое зависит от естественного языка, на котором написан текст документа. Например, информационно-поисковая система очень хорошо распознает русский язык благодаря его развитой морфологии. В случае с английским языком могут возникать трудности из-за его аналитической природы. Целью синтаксического анализа является нахождение синтаксической зависимости слов в предложении и построение его функционального дерева. Автоматически выделяются смысловые элементы фразы: логический субъект, логический предикат, объекты (прямые и косвенные дополнения), атрибуты (определения) и различные виды обстоятельств. Синтаксический анализ – достаточно трудная задача, сложность которой напрямую зависит от количества слов в предложении и использованных в нем правил. Существует несколько методов проведения такого анализа, которые воплощены в программных продуктах типа *Ergo Linguistic Technologies Parser, Link Parser, Functional Dependency Grammar* и др. Семантический анализ проводится с целью распознавания смысла текста. Решение этой задачи относится к трудно реализуемым процессам из-за необходимости существования безупречного механизма экспертной оценки качества данных. Поэтому на сегодняшний день системы, способные провести семантический анализ в полном объеме, отсутствуют [3; 4].

В некоторых интеллектуальных информационно-поисковых системах используется метод «неточного поиска», позволяющий определять веб-страницы, которые могут быть релевантными запросу на поиск даже при их неточном совпадении. Системы могут корректировать ряд опечаток в запросе пользователя на основе входящего в их состав массива омонимов и слов, имеющих альтернативное написание. Например, пользователь ввел в поисковую систему *Google* или *Yahoo!* неверно написанное слово *Onomatopia.* В результате список найденных совпадений будет сопровождаться вопросом: *Вы имели в виду Onomatopoeia?* В сравнении с точным поиском подобный подход значительно увеличивает качество поисковых результатов [5].

Необходимо также отметить метод семантико-синтаксического анализа, положенный в основу работы интеллектуальной поисковой системы *Exactus* и значительно повышающий точность ее работы (рис. 2) [6]. При этом слово рассматривается как единица «словарного состава языка в совокупности его конкретных грамматических форм и выражающих их флексий, а также возможных конкретных смысловых вариантов» (цит. по [6]). Слово-лексема не является синтаксической единицей, слово – единица лексики, а в разных его формах могут реализоваться или актуализироваться разные стороны его общего значения, разные семы, предопределяющие различия и в синтаксическом употреблении. Таким образом, синтаксис имеет дело с осмысленными единицами, несущими свой обобщенный категориальный смысл в конструкциях с разной степенью сложности. Исходя из этого, подобные единицы получили название синтаксем. В процессе поиска целью должна стать не лексема, а синтаксема, не только лексическое, но и производное от него синтаксическое значение компонента запроса. Важно подчеркнуть, что синтаксическое значение складывается в результате соединения категориального значения и морфологической формы и реализуется в определенной синтаксической позиции. Рассмотрение слова изолированно, в отрыве от текста не позволит установить его синтаксическое значение [6].


Рис. 2 Схема работы интеллектуальной системы *Exactus*

К числу интеллектуальных информационно-поисковых систем можно отнести первую версию системы, разработанной в исследовательском центре корпорации *Microsoft Research* и получившей название *Ask MSR* [7, с. 189–190]. Система способна не только производить поиск в Сети, но и извлекать из найденных веб-страниц полезную информацию, текстовые фрагменты с фактами, которые используются для ответа на вопрос пользователя. При этом ответ системы представляет собой одно слово или предложение. Процедура обработки вопроса пользователя и поиска релевантной информации осуществляется следующим образом. Система *Ask MSR* анализирует структуру вопроса, определяет основной объект поиска, преобразует вопрос в поисковой запрос, отправляет его в обычную (классическую) информационно-поисковую систему, получает результат поиска, а затем интеллектуально отфильтровывает найденные страницы и выдает требуемый ответ. Для создания собственной базы знаний система *Ask MSR* проанализировала свыше 1 млрд. веб-страниц, выбрав 2.3 млн. адресов часто задаваемых вопросов. Разработанная версия системы *Ask MSR* пока обеспечивает корректные ответы только на 40% вопросов, что, тем не менее, можно считать большим успехом.

Таким образом, основные отличия интеллектуальной поисковой системы от классической заключаются в умении системы обрабатывать запросы пользователя, сформулированные в произвольной форме на естественном языке, а также представлять результаты поиска не в виде ранжированных адресов сайтов, а в виде конкретного текстового фрагмента, содержащего нужную пользователю информацию.

ЛИТЕРАТУРА

1. Алгоритмы поисковых систем [Электронный ресурс].— Режим доступа: http://xbb.uz/SEO/Algoritmy-poiskovy-h-sistem. — Дата доступа: 12.11.2015.

2. Описание технологии лингвистического анализа [Электронный ресурс]. — Режим доступа: http://asknet.ru/Technology/techdescr.htm. — Дата доступа: 23.10.2015.

3. Диковицкий, В.В., Шишаев, М.Г. Обработка текстов естественного языка в моделях поисковых систем [Электронный ресурс]. — Режим доступа: http://cyberleninka.ru/article/n/ob-rabotka-tekstov-estestvennogo-yazyka-v-modelyah-poiskovyh-sistem. — Дата доступа: 12.11.2015.

4. Концептуальные положения создания программного обеспечения поисковой машины, реализующей морфологическую, синтаксическую и семантическую обработку текстовой информации [Электронный ресурс]. — Режим доступа: http://www.shkola-gym.ru/obls/konceptualeni e-polojeniya-sozdaniya-programmnogo-obespecheniya. — Дата доступа: 21.12.2015.

5. Семантический поиск [Электронный ресурс]. — Режим доступа: http://www.seo-news.ru/analytics/semanti-cheskiy-poisk. — Дата доступа: 21.12. 2015.

6. Осипов, Г.С. Интеллектуальная поисковая система «Exactus»: опыт участия в семинаре РОМИП [Электронный ресурс].— Режим доступа: http://romip.ru/romip-2005/addons/Exactus.pdf. — Дата доступа: 23.10.2015.

7. Ландэ, Д.В. Поиск знаний в Internet. Профессиональная работа: Пер. с англ. — М., 2005.