УДК 681.3.06

Нгуен Ба Нгок, А.Ф. Тузовский

Обзор подходов семантического поиска

Рассмотрены подходы к поиску информации на основе семантических технологий. Кратко описаны функциональные особенности таких семантических поисковых систем, как Ask Jeeves, TrueKnowledge, Hakia. Поясняются варианты применения семантических технологий для решения задачи информационного поиска и делается сравнение подходов семантического поиска с традиционными подходами поиска по ключевым словам.

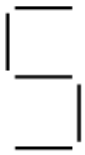
Ключевые слова: семантический поиск, семантическая технология, ключевые слова, модель, обзор, система информационного поиска.

Введение

Задача информационного поиска является классической задачей информатики. Несмотря на то, что исследования в данной области активно ведутся уже достаточно давно, эффективность современных систем поиска еще далека от совершенства. В связи с этим поисковые системы непрерывно совершенствуются, а также разрабатываются новые подходы. В последние годы большое внимание стало уделяться поиску информации на основе работы с семантикой [1–4]. Методы семантического поиска также известны, как смысловой поиск или поиск по смыслу текстов [4].

Для пояснения отличия методов семантического поиска от традиционных методов поиска по ключевым словам (в дальнейшем – традиционные методы поиска), рассмотрим следующие примеры.

Анализ содержания текста.



Предположим, что данный рисунок является аналогом некоторого текстового документа. Тогда в сценарии информационного поиска традиционные подходы видят данный рисунок как пять отдельных отрезков, расположенных в определенном порядке, а человек обычно понимает данный рисунок (его смысл) как печатную букву *S* или цифру 5, в зависимости от конкретного контекста.

Анализ смысла текста. Рассмотрим другой пример. Допустим, есть некоторый текст «*В финальном матче команда А победила команды Б со счетом 3:2*». В сценарии информационного поиска, чтобы найти информацию о победителе данного соревнования с помощью традиционных поисковых систем, пользователь должен подбирать подходящие ключевые слова. В данном случае, если пользователь использует запрос «*победитель соревнования*», то система вернее всего вернет такое сообщение, как «*по вашему запросу, ничего не найдено*». Однако пользователь сможет найти нужную ему информацию косвенным путем, используя другой запрос, например «*Финальный матч*», для получения данного текста, а затем определит победителя соревнования самостоятельно.

В отличие от традиционных подходов, в семантических поисковых системах пользователь может задавать вопросы на языке, близком к естественному. В рассматриваемой выше ситуации правильный запрос будет выглядеть примерно так: «*Кто победитель соревнования?*», при этом система должна будет выполнить анализ смысла текста и смысл запроса для формирования соответствующих ответов.

В результате сказанного выше можно дать следующее определение семантического поиска: семантический поиск – это метод информационного поиска, в котором релевантность документа запросу определяется семантически, а не синтаксически [4]. В традиционных подходах поиска релевантность документов и запросов определяется синтаксически, путем вычисления встречаемости ключевых слов в документе, без учета их семантических особенностей [6]. Семантическая релевантность оценивается по близости смыслов текстов, как это делает человек, т.е. семантические поисковые машины выполняют определение и описание смысла текста. Подходы семантического поиска используют именно такие технологии понимания текстов для улучшения качества поиска [6]. Более подробно методы семантического поиска будут пояснены на примерах семантических поисковых систем, описанных ниже.

Системы семантического поиска

Система Ask Jeeves (Ask.com). Ask Jeeves является системой вопрос–ответ, т.е. пользователь задает вопрос, а система отвечает, как это делается при общении между людьми (рис. 1).

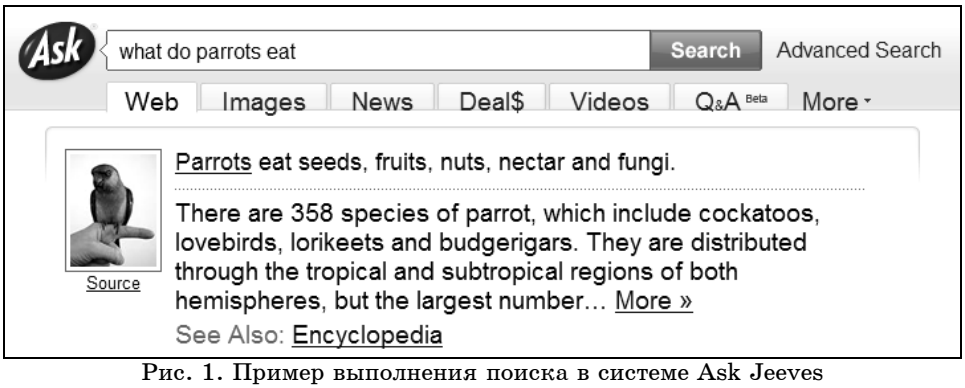


Рис. 1. Пример выполнения поиска в системе Ask Jeeves

В тексте, который система Ask Jeeves нашла на некотором сайте, выделяется та его часть, которая содержит ответ на заданный запрос, благодаря чему пользователю не надо читать всю страницу для нахождения нужной информации, как в традиционных подходах. Система Ask Jeeves функционирует на основе трех семантических технологий:

• технологии вывода прямого ответа из базы данных (direct answer from database, DADS);

• технологии вывода прямого ответа из результатов поиска (direct answer from search, DAFS);

• поисковый робот AnswerFarm, который индексирует пары вопрос–ответ (Q&A) из web-сети. Найденные в web-сети пары Q&A сохраняются в базе данных для выдачи ответа на вопросы пользователей.

Система True Knowledge (TrueKnowledge.com). True Knowledge, как и Ask Jeeves, является системой ответов на вопросы (рис. 2). В отличие от Ask Jeeves, True Knowledge использует другой подход к выдаче прямых ответов. В системе True Knowledge ответ производится на основе сохраненных фактов (триплеты в формате субъект–предикат–объект) и правила логического вывода.

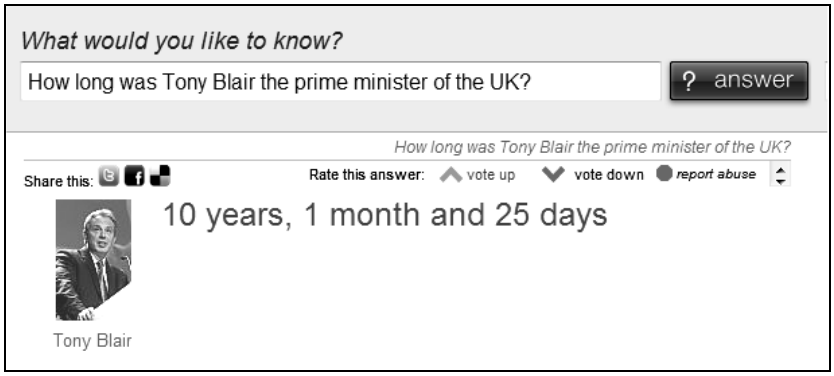


Рис. 2. Пример выполнения поиска в системе True Knowledge

Поиск в данной системе выполняется следующим образом. Вначале определяется смысл вопроса путем выделения в нем некоторых утверждений (фактов), которые содержатся в базе данных системы. После того как такие факты будут найдены, система *True Knowledge* формирует ответ на основе правила логического вывода для этих фактов. В примере показанном на рис. 2 для запроса «How long was Tony Blair the prime minister of the UK» (какое время Тони Блэр был премьер-министром Великобритании), True Knowledge использует следующие факты для вывода ответа на поставленный вопрос: «Tony Blair has been the prime minister of the UK between May 2nd 1997 and June 27th 2007» (Тони Блэр был премьер-министром Великобритании со 2 мая 1997 по 27 июня 2007); «Tony Blair has not been the prime minister of the UK until May 2nd 1997» (Тони Блэр не был премьер-министром Великобритании до 2 мая 1997); и «Tony Blair has not been the prime minister of the UK since June 27th 2007» (Тони Блэр не был премьер-министром Великобритании с 27 июня 2007).

В настоящее время в базе данных системы True Knowledge содержится около 300 млн. фактов о более чем 8 млн объектах. Данная база дискретных фактов заполняется двумя способами: путем импорта из внешних баз данных и путем ручного занесения данных пользователями системы.

Система Hakia (Hakia.com). Hakia является ярким примером использования семантического подхода для поиска документов в web-сети, которые семантически релевантны поисковому запросу. Как и в системах Ask Jeeves и True Knowledge, запрос пользователя для Hakia может быть представлен на естественном языке (пример работы Hakia показан на рис. 3).

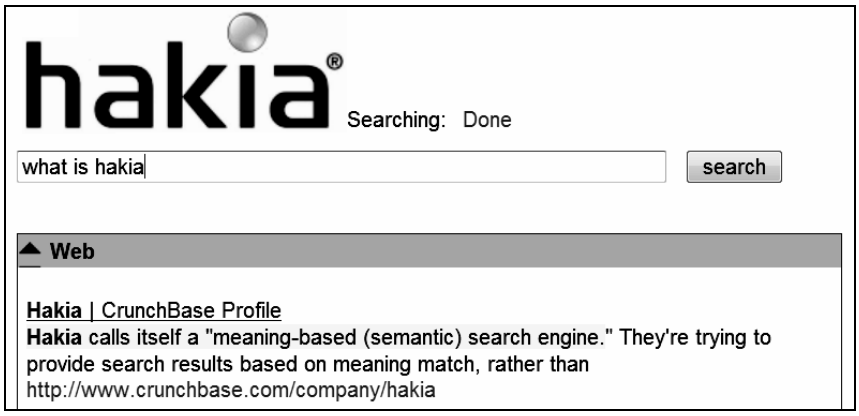


Рис. 3. Пример выполнения поиска в системе Hakia

Момент, многие специалисты считают, что именно подход семантического поиска, разработанный в системе Hakia, является новым этапом развития поисковых систем [7]. Однако, пока трудно однозначно сказать, лучше или хуже система Hakia, в сравнении с поисковой системой Google. Несмотря на это, данная система действительно является отлично разработанным, инновационным подходом к поиску информации.

Система Hakia базируется на трех технологиях: OntoSem – хранилище семантической информации, QDEX – технология индексации документов и SemanticRank – компонент ранжирования текстов по смыслу. С использованием этих технологий система Hakia достаточно успешно справляется с задачей анализа смысла текстов.

OntoSem является хранилищем отношений между концептами семантической модели, или разными терминами (словами), т.е. это лингвистическая база данных, где слова распределяются по категориям в зависимости от своих значений. QDEX является аналогом обратного индексирования в традиционных подходах информационного поиска. Для каждого документа QDEX определяет список возможных вопросов к нему и использует эти вопросы в качестве индекса при поиске. Компонент SemanticRank реализует специальный алгоритм, который используется для ранжирования результатов поиска по степени семантической близости. Для вычисления степени релевантности используется интеллектуальный алгоритм анализа выражения естественного языка и не применяются оценки соответствия по ключевому слову или по булевой логике.

Сравнение систем семантического поиска

На основе рассмотрения приведенных примеров видно, что ключевым отличием подходов семантического поиска от традиционных подходов является способность понимания смысла текста. Система True Knowledge понимает структурированные данные (триплеты [5]), а системы Ask Jeeves и Hakia понимают текст на естественных языках.

При сравнении этих двух подходов к кодированию и обработке смысла текста между собой по точности лучшим является первый подход, так как используемые в нем структурированные данные могут обрабатываться программно и имеются эффективные алгоритмы для такой обработки. Однако в масштабах такой сети, как Web, где неструктурированные тексты составляют большую часть информации, второй подход является более интересным, так как он может быть применен для любых текстов.

При сравнении подходов семантического поиска с традиционными подходами поиска по ключевым словам можно отметить, что теоретически они имеют ряд преимуществ над традиционными подходами в смысле повышения релевантности получаемых результатов. Это связано с тем, что релевантными результатами являются документы, удовлетворяющие информационные потребности пользователей и релевантность оценивается по смыслу текстов.

Главным недостатком подходов семантического поиска в сравнении с традиционными подходами поиска является тот факт, что алгоритмы обработки смысла текстов зависят от особенностей конкретного анализируемого естественного языка, т.е. требуется создание специальных алгоритмов для разных естественных языков. При этом для каждого

естественного языка должны учитываться его синтаксические и семантические особенности, отношения между словами и т.п. В связи с этим реализация подходов семантического поиска в многоязычных системах является очень сложной и трудоемкой работой.

Заключение

Трудно однозначно ответить на вопрос, как лучше формулировать поисковый запрос: с использованием набора ключевых слов или вопроса на естественном языке. Вероятнее всего, лучшим решением будет их комбинация, т.е. поисковые системы должны хорошо понимать разные формы запросов. Для этого есть смысл преобразовать вопрос пользователя на естественном языке в соответствующие ему набор ключевых слов.

В настоящее время основными используемыми системами информационного поиска являются представители традиционных подходов поиска по ключевым словам (Google, Bing, Yahoo и т.д). Это говорит о несомненной эффективности таких подходов для решения данной задачи. Однако даже крупнейшим компаниям, предоставляющим сервисы информационного поиска трудно отказаться от огромных возможностей, которые могут дать семантические поисковые системы. Примерами применения семантических технологий в глобальных масштабах можно считать проекты SearchMonkey от Yahoo, Rich Snippets от Google, или Bing Powerset. Такие системы подтверждают тот факт, что семантический метод является перспективным направлением развития поиска информации.

Литература

1. Heflin J. Searching the web with SHOE / Hendler J. // Artificial Intelligence for Web Search: American Association for Artificial Intelligence (AAAI). – Menlo Park, CA: WS-00-01, AAAI Press, 2000. – P. 35–40.

2. Stojanovic N. On analysing query ambiguity for query refinement: the librarian agent approach // Conceptual Modeling: 22nd International Conference on Conceptual Modeling. – Chicago, USA, 2003. – P. 490–505.

3. Squiggle: a semantic search engine for indexing and retrieval of multimedia content / I. Celino, E.D. Valle, D. Cerzza, A. Turati // Proceedings of SAMT. – 2006. – P. 20–34.

4. Guha R. Semantic search / R. Guha, R. McCool, E. Miller // Proceedings of the 12th international conference on World Wide Web. – N.Y. ACM Press, 2003. – P. 700–709.

5. Mihalcea R. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity / R. Mihalcea, C. Corley, C. Strapparava // American Association for Artificial Intelligence (AAAI'06). – Boston:USA, 2006.

6. Christopher D.M. Introduction to information retrieval / D.M. Christopher, R. Prabhaka, S. Hinrich. – Cambridge University press, 2008. – 504 p.

7. Mangold C. A survey and classification of semantic search approaches // Metadata, Semantics and Ontology. – 2007. – Vol. 2, № 1. – P. 23–34.

Нгуен Ба Нгок
Аспирант каф. оптимизации систем управления
Национального исследовательского Томского политехнического университета (НИТПУ)
Тел.: +7-913-868-58-99
Эл. почта: ngocrs1984@sibmail.com

Тузовский Анатолий Федорович
Д-р техн. наук, проф. каф. оптимизации систем управления НИТПУ
Тел.: +7 (3822) 42-14-85
Эл. почта: tuzovskyaf@tpu.ru

Ba Ngoc Nguyen, Tuzovsky A.F.
A review of semantic search approaches

The innovative approaches to information retrieval, semantic search approaches,are presented. The following semantic search approaches as Ask Jeeves, TrueKnowledge and Hakia are briefly discussed. Modifications of the semantic technologies application are analyzed, and comparison of the semantic search approaches with the traditional keywords search approaches is carried out.
Keywords: semantic search, semantic technology, keywords search, relevance model, review, information retrieval system.