УДК 004.75.05

Д.В. Гринев

Харьковский национальный экономический университет, Харьков

СЕМАНТИЧЕСКИЙ ПОИСК В WEB

В статье проведен анализ таких видов семантического поиска в Web как поиск по метаданным и полнотекстовый поиск. Рассмотрена концепция развития интернет под названием Semantic Web (семантическая паутина), ее основные компоненты и стандарты описания данных. Представлены выводы о перспективности разработки методов семантического поиска без анализа метаданных и метатегов из-за низкой степени практической реализуемости проекта Semantic Web.

Ключевые слова: *семантический поиск, полнотекстовый поиск, поиск по метаданным, Semantic Web (семантическая паутина), семантическая сеть, RDF, RDF schema, OWL.*

Вступление

Постановка проблемы. Количество информации, которую создает мировое сообщество, растет с каждым годом. Открытость информационного поля теоретически обеспечивает свободный и быстрый доступ к данным. Однако у такой всеобщей доступности есть и обратная сторона – чтобы получить информацию, ее нужно сначала найти. Получение информации об интересующем объекте в подавляющем большинстве случаев сводится к использованию интернет ресурсов поисковых систем (ПС).

Поиск информации поисковым роботом представляет собой процесс выявления в индексированном множестве ПС релевантных документов, т.е. таких, которые удовлетворяют заранее определенному запросу. Основная задача состоит в том, чтобы на конкретный запрос пользователя ПС провела обработку информации с последующим ранжированием найденных web-ресурсов по их релевантности.

Пользователь не может описать системе признаки искомого объекта, поскольку принцип поиска ПС базируется на тексте и ключевых словах. Фактически пользователю сложно найти данные, о которых он еще не знает, а для их получения необходимо ввести в строку запроса информацию, содержащуюся в ответе.

Приходится различать формальную релевантность и содержательную релевантность, причем, если формальная релевантность, повторяющая в листе выдачи ПС форму запроса, но не передающая изначально заданной содержательной сути сегодня достижима, то реализация содержательного соответствия документа смыслу запроса является в большинстве случаев нерешенной.

Таким образом, основной проблемой нахождения смыслового соответствия документа пользовательскому запросу является разработка и реализация подходов, основанных на семантическом поиске.

Изложение основного материала

Семантический поиск является одним из методов информационного поиска и представляет собой процесс поиска документов по их смысловому содержанию. Основой семантического поиска служат заранее установленные отношения между символами и объектами, которые они обозначают.

Можно выделить два основных вида семантического поиска.

Полнотекстовый поиск – поиск по всему содержимому документа с использованием предварительно построенных индексов.

Поиск по метаданным – это поиск по неким атрибутам документа, которые описывают определенные объекты поддерживаемые системой. Например, автор, адрес, название организации и т. д.

Именно использование метаданных сегодня широко применяется в относительно новой концепции развития интернет под названием Semantic Web (семантическая паутина) [1]. Основной акцент концепции делается на работе с метаданными, однозначно характеризующими свойства и содержание веб-ресурсов, вместо используемого в настоящее время текстового анализа документов.

Эта концепция была принята и продвигается Консорциумом W3C [2]. Для ее внедрения предполагается создание сети документов, содержащих метаданные о web-ресурсах. Тогда как сами ресурсы предназначены для восприятия человеком, метаданные используются поисковыми роботами (агентами) для проведения однозначных логических заключений о свойствах этих ресурсов. Такой подход уже успели окрестить как Web 3.0.

Semantic Web в математической форме представляет собой разновидность графа, где роль вершин выполняют понятия базы знаний, а направленные дуги задают отношения между ними. Таким образом, строится семантическая сеть, которая отражает семантику предметной области в виде поня-

тий и отношений. Идея состоит в том, чтобы глобальной семантической сетью было подмножество систем, которые замкнуты на специфичных путях достижения достаточного удобства для агентов.

В семантической паутине предполагается повсеместное использование, во-первых, универсальных идентификаторов ресурсов (URI), а во-вторых – онтологий и языков описания метаданных.

Использование URI. Традиционная схема использования таких идентификаторов в web сводится к установке ссылок, ведущих на объект. Объектом может быть веб-страница или ее фрагмент, файл, и др., а также ресурсы, недоступные для скачивания, например, отдельные люди, города и другие географические сущности, художественные артефакты и т.д. URI должен быть уникальным и идентифицировать реально существующий объект.

Использование онтологий и языков описания метаданных. Современные методы автоматической обработки данных, как правило, основаны на частотном и лексическом анализе текстового содержимого. В семантической паутине предлагается использовать форматы описания, доступные для машинной обработки, например, семейство форматов, часто упоминаемое в литературе как "Semantic Web family", в свою очередь, использующие URI для адресации описываемых и описывающих объектов, а также онтологии и дескриптивные логики в качестве базовых математических формализмов.

Когда агенты смогут понимать смысл той информации, с которой работает пользователь, ПС смогут предоставлять более релевантные списки ссылок на документы. Для достижения этого необходимо чтобы определение типов данных и связей между объектами проводилось самими авторами веб-страниц. Такую концепцию еще в 2001 году предложил не кто иной, как Тим Бернерс-Ли – создатель Всемирной паутины [1].

В 2001 году была сформулирована информационная коммуникационная модель Semantic Web, аналогичная семиуровневой модели OSI и ориентированная на обмен в первую очередь информацией, а не данными [3]. В 2005 году появилась новая редакция этой модели. Стек стандартов Semantic Web описывает интерфейсы между уровнями. Но кроме стандартов нужны еще и средства для реализации семантической паутины, поэтому кроме самого стека должны активно развиваться сервисы, обеспечивающие работу Semantic Web. С практической точки зрения наибольший интерес представляет процесс сближения идей семантической паутины и веб-сервисов. Развитие в этом направлении может привести к созданию нового поколения сервисов, которые пока условно называют "интеллектуальными веб-сервисами".

Основные компоненты Semantic Web рекомендуемые W3C представлены на рис. 1 [4].

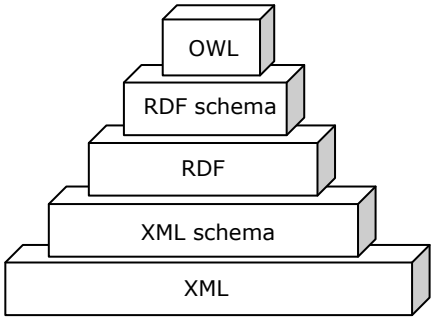

Рис. 1. Рекомендации W3C

XML предоставляет синтаксис для определения структуры документа, подлежащего машинной обработке. Синтаксис XML не несет семантической нагрузки, т.е. он дает возможность пользователям снабжать свои документы произвольной структурой, однако данный язык ничего не говорит о том, что означает эта структура.

XML Schema определяет ограничения на структуру XML-документа, для того, чтобы обеспечить предсказуемость обработки. Стандартный синтаксический анализатор языка XML в состоянии проверить произвольный XML-документ на соответствие его структуры, так называемой схеме документа, описанной в XML Schema.

RDF (Resource Description Framework) является универсальным языком представления знаний в Web и включает в себя принципы описания ресурсов. В то время как XML схемы просто описывают структуру документа, RDF имеет дело со знаниями как таковыми. Это позволяет значительно расширить область применения данных, представленных в таком формате.

RDF формирует базовый слой для создания семантической сети за счет определения управляемых графов связей, представленных триплетами объект-атрибут-значение. Триплеты задаются с помощью URI с применением тэгов языка XML. URI-идентификаторы гарантируют то, что каждое понятие, используемое в документе – это не просто слово, а нечто, привязанное к единому определению. Таким образом, в языке RDF документ состоит из утверждений о том, что нечто имеет определенное отношение с некоторым определенным значением.

Для сериализации данных, представленных в RDF, разработан и рекомендован W3C стандартный формат обмена данными – RDF XML.

RDF Schema – это семантическое расширение RDF, которое обеспечивает механизмы описания связанных ресурсов и их связей. Система классов и свойств RDF Schema похожа на систему типов языков объектно-ориентированного программирования (ООП) с некоторыми отличиями. Так, описательный язык словаря RDF определяет свойства в терминах того класса ресурсов, к которому эти свойства отно-

сятся, в отличие от языков ООП, описывающих класс в терминах свойств своих элементов.

Существование стандартов для описания данных (RDF) и их атрибутов (схема RDF) позволяет создавать инструменты обработки информации из многочисленных источников. То, насколько глубоко различные приложения могут обмениваться данными и использовать их, принято называть синтаксическим взаимодействием сетей. Чем более стандартизированными и распространенными являются эти инструменты работы с данными, тем выше степень синтаксического взаимодействия сетей.

Синтаксическое взаимодействие сетей требует определенного преобразования между терминами, для чего, в свою очередь, необходим контентный анализ. Два поисковых агента могут использовать различные идентификаторы для обозначения одного и того же понятия и им необходимо объяснить, что два конкретных термина используются ими для обозначения одного и того же.

Такой контентный анализ требует формальных и подробных спецификаций моделей доменов, которые определяют используемые термины и их связи. Подобные формальные модели доменов принято называть онтологиями. Они определяют модели данных в терминах классов, подклассов и свойств. Проще говоря, онтология – это документ, формально задающий отношения между терминами.

Наиболее типичными видами онтологий в Web являются таксономия и набор правил вывода.

Таксономия определяет классы объектов и отношения между ними. Например, понятие адрес может быть определено как разновидность понятия местонахождение, а код города можно задавать применительно лишь к местонахождениям и так далее. Большое количество отношений между индивидами можно задать путем приписывания классам определенных свойств и позволяя подклассам наследовать эти свойства.

Правила вывода, задаваемые в онтологиях, дают еще больше возможностей. В рамках онтологии можно записать такое правило: «Если объект А соответствует некоторому объекту В, а в объекте С фигурирует объект А, то этому объекту С тоже соответствует объект В». Поисковый робот не «понимает» в полном смысле этого слова ничего из всей этой информации, но теперь он уже может манипулировать терминами гораздо более эффективно с тем, чтобы стать полезным и осмысленным для пользователя.

Онтологический язык Web (Web Ontology Language), рекомендуемый W3C, помогает в выражении онтологий. Рабочий онтологический язык Ontology Working Language (OWL) добавляет больше словарных возможностей для описания свойств и классов, чем RDF или схема RDF. В частности, он позволяет описывать связи между классами, мощность множества, равенство, более богатую типологию свойств и их характеристики.

Рабочая Группа W3C по доступу к данным разработала язык запросов SPARQL [5], имеющий SQL-подобный синтаксис, определяет запросы в терминах шаблонов графа, которые сравниваются с направленным графом, представляющим данные RDF. SPARQL предоставляет возможности, для запроса необходимых и необязательных шаблонов, а также для их объединения и разделения. Результат сравнения также может быть использован для конструирования нового графа RDF с использованием отдельного шаблона. Используя такие точки доступа SPARQL, агенты могут запрашивать удаленные RDF данные и, даже, формировать новые RDF графы, без какой-либо локальной обработки.

Одним из первых серьезных и популярных проектов, основанным на принципах семантической паутины, стал проект "Дублинское ядро", реализуемый инициативной организацией Dublin Core Metadata Initiative (DCMI) [6]. Это открытый проект, цель которого разработать стандарты метаданных в формате RDF, независящие от платформ и подходящие для широкого спектра задач.

В то время как совокупность ресурсов и их метаданных можно считать статической частью семантической паутины, ее динамическую часть представляют т. н. семантические веб-сервисы – законченные элементы программной логики с однозначно описанной семантикой, доступные через интернет и пригодные для поиска, композиции и выполнения.

Консорциум W3C предполагает использование для описания веб-сервисов тех же языков разметки, что и для статической части семантической паутины, а также онтологии OWL-S, описывающей базовую терминологию предметной области. Онтология OWL-S состоит из четырех онтологий – онтологии сервиса, онтологии модели сервиса, онтологии процесса и онтологии базы.

Потенциальная выгода от использования семантических веб-сервисов заключается в возможности автоматического поиска программными агентами подходящих сервисов для решения поставленных задач. Тем не менее, сложность этой задачи в ее общей формулировке пока позволяет добиваться некоторых положительных результатов только в узкоспециализированных отраслях, явным образом выигрывающих от внедрения сервисо-ориентированной архитектуры (SOA), например, в интеграции корпоративных приложений.

Несмотря на очевидную актуальность Semantic Web существуют сложности ее практической реализуемости.

Во-первых, необходимость описания метаданных приводит к дублированию информации. Правда, этот недостаток семантической паутины был главным толчком к созданию микроформатов, с помощью которых можно семантически размечать сведения о разнообразных сущностях непосредственно в коде HTML или XHTML.

Во-вторых, в семантической паутине для получения ответа на некоторые вопросы, совсем не обязательно переходить по ссылке на сайт. Доля поискового трафика на сайты может значительно снизиться, т.к. ПС будут сами отбирать и предоставлять нужную пользователю информацию. Соответственно отпадает необходимость посещать сайт, на котором опубликованы материалы и реклама, а значит, коммерческая выгода от привлечения пользователей на сайт уменьшается в разы. Здесь же можно сказать о том, что сохранение анонимности или же авторских прав на текстовую информацию становится весьма проблематичным.

Полнотекстовый семантический поиск является альтернативным подходом к концепции семантической паутины. Разрабатываются алгоритмы, которые самостоятельно анализируют содержание интернет и переводят его из текстового представления в объектное. Запросы на человеческом языке являются для пользователя естественными и актуальность анализа таких запросов очевидна. Однако это непростая задача, ведь нужно создать онтологию каждого процесса, а для этого необходимо знать структуру взаимоотношений объектов, правила их существования и интерпретации.

Выводы

Поисковые системы, использующие поиск по метаданным, работают с объектами, а не с фрагментами текста, а, следовательно, подобный подход позволяет осуществлять более эффективный поиск. Однако слабость такого подхода заключается в том, что сейчас практически вся информация в интернете представляет собой как раз текст, и, чтобы столь же эффективно решать задачи глобального поиска, нужно научиться из текста выделять объекты.

Проведя анализ семантической паутины, становится очевидно, что сегодня перспективно разрабатывать методы семантического поиска без анализа метаданных и метатегов из-за низкой степени практической реализуемости проекта Semantic Web. Надеяться на то, что семантический поиск должен быть основан только на поиске по метаданным в полной мере нельзя, а значит, разработка новых методов полнотекстового семантического поиска будет являться дополнительным аргументом в пользу ускорения развития различных систем искусственного интеллекта.

Список литературы

1. Berners-Lee T. The Semantic Web [Электронный ресурс] / T. Berners-Lee, J. Hendler, O. Lassila // Режим доступа к статье: http://www. scientificamerican.com/ article.cfm?id=the-semantic-web.

2. W3C Semantic Web Activity [Электронный ресурс]. – Режим доступа к статье: http://www.w3.org/2001/sw/wiki/ Main_Page.

3. Черняк Л. О стеке стандартов Semantic Web [Электронный ресурс] / Л. Черняк // Computerworld Россия. –2006. – № 12. – Режим доступа к статье: http://www.w3.org/DesignIssues/LinkedData.html.

4. Андреев А.М. Использование технологии Semantic Web в системе поиска несоответствий в текстах документов [Электронный ресурс] / А.М. Андреев, Д.В. Березкин, В.С. Рымарь, К.В. Симаков. – Режим доступа к статье: http://www.inteltec.ru/publish/articles/textan/rimar_ RCDL2006.shtml.

5. Технологии Semantic Web [Электронный ресурс]. – Режим доступа к статье: http://www.semantictools.ru/ technology.html.

6. Dublin Core Metadata Initiative (DCMI) [Электронный ресурс]. – Режим доступа к статье: http://dublincore.org/.

Поступила в редколлегию 21.09.2012

Рецензент: д-р техн. наук, проф. В.А. Краснобаев, Полтавский национальный технический университет имени Кондратюка, Полтава.

СЕМАНТИЧНИЙ ПОШУК В WEB

Д.В. Гриньов

У статті проведено аналіз таких видів семантичного пошуку в Web як пошук по метаданих і повнотекстовий пошук. Розглянуто концепцію розвитку Інтернет під назвою Semantic Web (семантична павутина), її основні компоненти і стандарти опису даних. Представлені висновки про перспективність розробки методів семантичного пошуку без аналізу метаданих і метатегів через низький ступень практичної реалізованості проекту Semantic Web.

Ключові слова: *семантичний пошук, повнотекстовий пошук, пошук по метаданим, Semantic Web (семантична павутина), семантична мережа, RDF, RDF schema, OWL.*

SEMANTIC SEARCH IS IN WEB

D.V. Grinev

In the article the analysis of such types of semantic search is conducted in Web as a search to on to the metadatas and fulltext search. Conception of development is considered the internet under the name Semantic Web (semantic spider web), its basic components and standards of definition of data. Conclusions are presented about perspective of development of methods of semantic search without the analysis of metadatas and memategs from the low degree of practical realized of project Semantic Web.

Keywords: *semantic search, fulltext search, search to on to the metadatas, Semantic Web (semantic spider web), semantic network, RDF, RDF schema, OWL.*