

Зберігання, аналіз та захист даних в інформаційних системах

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, повторяющаяся 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 поль-

зовательскому 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 содержанию 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 веб-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 понятий 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 содержания. 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. Стекло 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 при-

вести к созданию нового поколения сервисов, которые пока условно называют "интеллектуальными веб-сервисами".

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

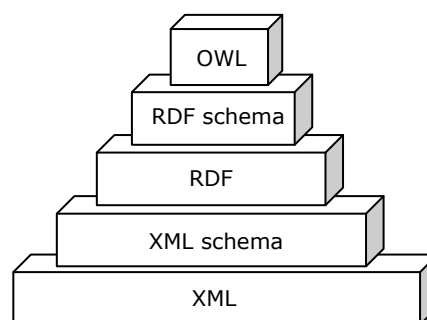


Рис. 1. Рекомендации W3C

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 язы-

ков об'єктно-орієнтованого програмування (ООП) з некоторими відмінностями. Так, описателів мови словаря RDF визначає властивості в термінах того класу ресурсів, до якого ці властивості відносяться, на відміну від мов ООП, описуваних клас в термінах властивостей своїх елементів.

Існування стандартів для описання даних (RDF) і їх атрибутів (схема RDF) дозволяє створювати інструменти обробки інформації з багаточисельних джерел. Тоді, наскільки глибоко різні застосування можуть обмінюватися даними і використовувати їх, прийнято називати синтаксичним взаємодією мереж. Чим більше стандартизованими і поширеними є ці інструменти роботи з даними, тим вище ступінь синтаксичного взаємодіювання мереж.

Синтаксичне взаємодіювання мереж вимагає визначеного перетворення між термінами, для чого, в свою чергу, потрібний змістовий аналіз. Два пошукових агенти можуть використовувати різні ідентифікатори для позначення одного і того ж поняття і їм необхідно пояснити, що два конкретних терміна використовуються ними для позначення одного і того ж.

Такий змістовий аналіз вимагає формальних і детальних специфікацій моделей доменів, які визначають використовувані терміни і їх зв'язи. Подібні формальні моделі доменів прийнято називати онтологіями. Вони визначають моделі даних в термінах класів, підкласів і властивостей. Проще кажучи, онтологія – це документ, формально задає відносини між термінами.

Найбільш типовими видами онтологій в Web є таксономія і набір правил виводу.

Таксономія визначає класи об'єктів і відносини між ними. Наприклад, поняття адрес може бути визначено як різновидність поняття місцезнаходження, а код міста можна задавати застосовно тільки до місцезнаходженням і так далі. Велика кількість відносин між індивідами можна задати шляхом приписування класам визначених властивостей і дозволяючи підкласам успадковувати ці властивості.

Правила виводу, задавані в онтологіях, дають ще більше можливостей. В межах онтології можна записати таке правило: «Якщо об'єкт А відповідає деякому об'єкту В, а в об'єкті С фігурує об'єкт А, то цьому об'єкту С теж відповідає об'єкт В». Пошуковий робот не «розуміє» в повному сенсі цього слова нічого з всієї цієї інформації, але тепер він може маніпулювати термінами значно більш ефективно з тим, щоб стати корисним і осмисленим для користувача.

Онтологічний мови Web (Web Ontology Language), рекомендуваний W3C, допомагає в вираженні онтологій. Робочий онтологічний мови Ontology Working Language (OWL) додає більше словарних можливостей для описання властивостей і

класів, ніж RDF або схема RDF. В частині, він дозволяє описувати зв'язи між класами, потужність множини, рівність, більш багату типологію властивостей і їх характеристики.

Робоча Група W3C по доступу до даних розробила мову запитів SPARQL [5], яка має SQL-подібний синтаксис, визначає запити в термінах шаблонів графа, які порівнюються з напрямленим графом, що представляє дані RDF. SPARQL надає можливості, для запиту необхідних і необов'язкових шаблонів, а також для їх об'єднання і розділення. Результат порівняння також може бути використаний для побудови нового графа RDF з використанням окремого шаблону. Використовуючи такі точки доступу SPARQL, агенти можуть запитувати віддалені RDF дані і, навіть, формувати нові RDF графи, без якої-небудь локальної обробки.

Одним з перших серйозних і популярних проєктів, заснованих на принципах семантичної мережі, став проєкт "Дублінське ядро", реалізуваний ініціативною організацією Dublin Core Metadata Initiative (DCMI) [6]. Це відкритий проєкт, метою якого розробити стандарти метаданих в форматі RDF, незалежні від платформ і підходящі для широкого спектру завдань.

В той час як сукупність ресурсів і їх метаданих можна вважати статичною частиною семантичної мережі, її динамічну частину представляють т. н. семантичні веб-сервіси – завершені елементи програмної логіки з однозначно описаною семантикою, доступні через інтернет і придатні для пошуку, композиції і виконання.

Консорціум W3C передбачає використання для описання веб-сервісів тих же мов розмітки, що і для статичної частини семантичної мережі, а також онтології OWL-S, описуваних базову термінологію предметної області. Онтологія OWL-S складається з чотирьох онтологій – онтології сервісу, онтології моделі сервісу, онтології процесу і онтології бази.

Потенціальна вигода від використання семантичних веб-сервісів полягає в можливості автоматичного пошуку програмними агентами підходящих сервісів для рішення поставлених завдань. Тим не менше, складність цієї задачі в її загальній формулюванні поки дозволяє досягати деяких позитивних результатів тільки в спеціалізованих галузях, явним чином виграваних від впровадження сервісо-орієнтованої архітектури (SOA), наприклад, в інтеграції корпоративних застосувань.

Незважаючи на очевидну актуальність Semantic Web існують складності її практичної реалізації.

В першу чергу, необхідність описання метаданих призводить до дублювання інформації. Правда, цей недолік семантичної мережі був

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://dublin-core.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.