УДК 004.738.52

В.Н. Анохин, Б.О. Марчак, Д.В. Гринев

Харьковский национальный экономический университет, Харьков

АНАЛИЗ СИНТЕЗА И СЕМАНТИКИ ЯЗЫКОВ ЗАПРОСОВ ФОРМАТА ДАННЫХ RDF ДЛЯ ОСУЩЕСТВЛЕНИЯ ЭФФЕКТИВНОГО СЕМАНТИЧЕСКОГО ПОИСКА

Семантическая паутина — группа методов и технологий, созданных для того, чтобы позволить машинам понимать значение — или «семантику» — информации о всемирной паутине. Термин был введен на собрании World Wide Web (W3C) директором Тимом Бернерсом-Ли. Согласно исходному видению, доступность машиночитаемых метаданных позволила бы автоматизированным агентам и другому программному обеспечению более разумно получать доступ к сети. Агенты были бы в состоянии автоматически выполнять задачи и определять местоположение соответствующей информации от имени пользователя. В то время как слову «семантическая паутина» формально не дают определение, однако можно сказать, что это понятие главным образом используется для того, чтобы описать модель и технологии, предложенные W3C. Эти технологии включают платформу описания ресурсов (RDF), множество форматов обмена данными (например, RDF/XML, N3, Turtle, N-Triples), и нотации, такие как схема RDF (RDFS) и веб-язык онтологии (OWL), которые предназначены для того, чтобы обеспечить формальное описание понятий, сроки и отношения в пределах данной области знаний. Многие из технологий, предложенных W3C, уже существуют и используются в различных проектах.

***Ключевые слова**: Semantic Web (семантическая паутина), всемирная сеть, глобальная сеть, связь, ресурс.*

Вступление

В настоящее время в поисковых системах используется релевантная модель оценки соответствия исследуемого документа поисковому запросу. Данная модель практически не справляется с решением задач распознавания и поиска омонимов (грамматических, и, особенно, - лексических), синонимов и многозначных слов. Это обусловлено тем, что в основу релевантной модели поиска заложен лингвистический подход и ряд оценочных синтетических критериев (таких как положение слов на странице), а перечисленные выше языковые артефакты не могут быть распознаны без понимания смысла поискового запроса. Семантические поисковые системы пытаются привнести такой смысл в результаты запросов поиска, представленные в контекстном формате. В настоящее время семантические механизмы представляются провайдерами рекламы. В рамках статьи мы предлагаем рассмотреть данное явление с точки зрения их использования в бесплатных поисковых системах. Ключевым аспектом технологии семантической паутины также является многоуровневая структура, частью которой и является технология RDF (рис. 1) [1].

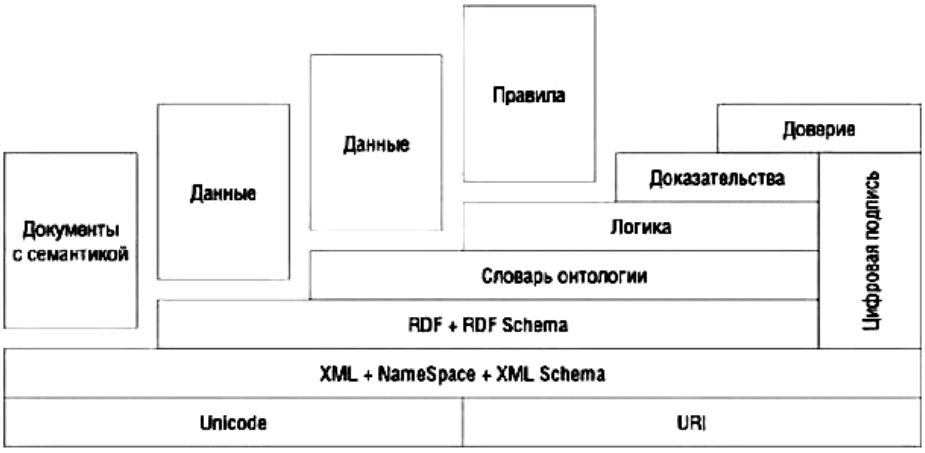

Рис. 1. Архитектура семантической сети

Crystal Semantics является разработчиком Textonomy Advance, первого в мире семантического механизма. Данный поисковый инструмент способен использовать знания человека, с которыми алгоритмы других программ не могут работать. Уникальная семантическая сеть от Crystal Semantics понимает смысл слов, выражений, а также устанавливает лингвистические связи между ними.

Textonomy, в отличие от существующих поисковых технологий, основу которых представляют статистические алгоритмы, использует лингвистические правила для определения семантической за-

висимости между словами и контекстом, в котором они встречаются. Функционирование Textonomy Advanced Engine напрямую связано с различными словарями и энциклопедиями, представленными различными источниками.

Основная часть

Семантический механизм стал результатом долгих научно-исследовательских работ (в течение 8 лет) в области поисковой лингвистики и $8-миллионных вложений в этот процесс. Все началось с разработки классификационной системы, предназначенной для данных, собранных для первого издания Кембриджской энциклопедии, позднее система получила широкое распространение среди многих других энциклопедий разных издательств, например, Cambridge University Press, Penguin Book.

В то же время вся база данных принадлежала издательству Cambridge University Press, но в 1997 году она была продана голландскому электронному издательству AND, которое начало ее разработку для интерактивного использования. В течение последующих 4 лет классификационная система была преобразована в «глобальную модель данных», предполагающую несколько приложений к классификации документов и поисковым технологиям. Когда компания AND в 2001 вышла из бизнеса, вся база данных была приобретена компанией Crystal Reference Systems, созданной с целью развития глобальной модели данных и ее основного понятия «семантического механизма». Это одна из крупнейших семантических систем, постоянно развивающихся под наблюдением профессора Кристалла и его высококвалифицированной редакторской команды.

Булевый поиск представляет собой одну из самых простых поисковых программ сравнения. Ярким примером булевого поиска служит использование любой крупной поисковой системы (Google, Yahoo) со множеством слов. Это предполагает использование оператора And для поиска всех элементов. Например, введем запрос «покупка плазменного телевизора онлайн», из этого будет следовать, что поиску подлежат все слова, соответствующие запросу. Все страницы, где есть слова купить, плазменный, телевизор и онлайн будут представлены в результатах поиска [2].

Другой пример. Если пользователь хочет исключить из поиска один из элементов, например, «купить плазменный телевизор онлайн – Sony», поисковый алгоритм воспримет это следующим образом: все релевантные результаты, имеющие слова купить, плазменный, телевизор и онлайн, будут включены в результаты поиска, а вот страницы, на которых есть слово Sony, будут исключены [5].

Некоторые поисковые системы поддерживают поиск слов, которые находятся на определенном удалении от элементов запроса. Поиск с расстоянием - поиск, при котором пользователь указывает, на каком расстоянии между собой должны располагаться ключевые слова в документе. Для осуществления данного вида поиска необходимо в конце фразы использовать тильду (~). Например, чтобы задать поиск слов теплица и углерод на расстоянии 10 слов друг от друга, в строку запроса нужно ввести следующее: теплица углерод ~10 [6].

Возможно, не все знакомы с понятием «неточный поиск». В процессе неточного поиска определяются страницы, которые могут быть релевантными аргументу поиску, даже если аргумент неточно соответствует желаемой информации. Неточный поиск осуществляется посредством «Программы неточного сравнения», которая демонстрирует список результатов, составленный на основе некоторого сходства слова-аргумента с написанным вариантом. Наиболее точные и релевантные совпадения можно будет найти в начале всего списка результатов поиска. Иногда присутствует оценка относительной релевантности (в процентах) результатов поиска [7].

Программа неточного сравнения может выполнять функции корректора правописания. Например, пользователь ввел слово Misissippi неверно в Yahoo! или Google (обе системы используют данную программу), список найденных совпадений будет сопровождаться вопросом «Вы имели в виду Mississippi?». В программе представлены слова с альтернативным написанием и слова, имеющие одинаковое звучание, но разное написание. Программа неточного сравнения корректирует общие опечатки, а также ошибки, допущенные в процессе оптического распознавания знаков (OCR) печатных документов [3].

Обычно программа неточного сравнения представляет помимо релевантных совпадений и нерелевантные. Как правило, это происходит, если слово имеет много значений, одно из которых может оказаться релевантным запросу. Если у пользователя только смутное представление о том, что ему нужно найти, то следует ориентироваться по оценке относительной релевантности (у нерелевантных совпадений оценка будет низкой).

С точки зрения научных исследований, неточный поиск представляет больше возможностей, чем его точный аналог. Неточный поиск широко применяется при исследовании малоизвестных, специфических работ и работ на иностранном языке, правильное написание названий которых не известно. Данный вид поиска также используется для определения местоположения объекта, информация о котором не точна или ее не достаточно.

В оффлайн-разговоре собеседники без труда понимают друг друга, легко определяя нужное значение многозначного слова по контексту. Поиск по контексту представляет собой онлайн-попытку определять нужное значение слова в зависимости от окружающих его слов (контекста). Именно поиск по контексту лежит в основе системы Crystal Semantics' Textonomy. Данный вид поиска имеет частичное сходство

с неточным поиском, а отличие в том, что поиск по контексту предусматривает оценку содержания всей страницы в целом, а не отдельного слова [4].

Однако релевантные результаты поиска продолжают оставаться актуальной проблемой для поисковых систем. Именно по причине некорректной информации многие предпочитают не использовать Интернет с этой целью. Даже поддерживая такие программы, как Булевый поиск, Неточный поиск и даже поиск с использованием Wildcard-символов, поисковые системы не могут достичь совершенства в поиске. Многие продолжают считать, что поиск шагнет вперед только с развитием Поиска по контексту.

Выводы

Идея семантической паутины, которая в состоянии описать и связать значение с данными, имеет больше чем простой код разметки XHTML. Это основано на предположении, что для того, чтобы можно было обеспечить машины возможностью точно интерпретировать контент, нужны не простые упорядоченные отношения, включающие в себя буквы и слова, а необходима базовая инфраструктура. Иначе большая часть благосклонной функциональности была бы доступна в Web 2.0 (и более ранних версиях), и было бы возможно получить семантически способную сеть с незначительными, инкрементными дополнениями.

Дополнения к инфраструктуре, чтобы поддерживать семантическую функциональность, включают в себя скрытые динамические сетевые модели, которые при определенных условиях могут изучить значение, основанное на данных порядка в отношениях ‘изучения’ процесса с порядком (своего рода элементарная рабочая грамматика).

Список литературы

1. Н.В. Рябова, С.С. Щербак Развитие технологий Semantic Web: обработка RDF–графов на основе XSLT [Электронный ресурс]. – Режим доступа: http://shcherbak.net/rdf_xslt_tech/

2. Стин Декер, Сергей Мельник, Франк ван Хермелен, Semantic Web: роли XML и RDF [Электронный ресурс]. – Режим доступа: http://www.publish.ru/os/1072570/text/180411/_p3.html

3. Даниил Кальченко Семантическая сеть [Электронный ресурс]. – Режим доступа: http://www.compress.ru/article.aspx?id=12195&iid=468

4. Дмитрий Ландэ Семантический веб: от идеи – к технологии [Электронный ресурс]. – Режим доступа: http://poiskbook.kiev.ua/sw.html

5. Dan Brickley, R.V. Guha RDF Vocabulary Description Language 1.0: RDF Schema [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/2002/WD–rdf–schema–20021112

6. Graham Klyne, Jeremy Carroll Resource Description Framework (RDF): Concepts and Abstract Data Mode [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/2002/WD–rdf–concepts–20020829

7. Diane Hillmann Using Dublin Core [Электронный ресурс]. – Режим доступа: http://dublincore.org/documents/usageguide

Поступила в редколлегию 4.04.2011

Рецензент: канд. экон. наук, проф. И.А. Золотарева, Харьковский национальный экономический университет, Харьков.

АНАЛІЗ СИНТЕЗУ І СЕМАНТИКИ МОВ ЗАПИТІВ ФОРМАТУ ДАНИХ RDF ДЛЯ ЗДІЙСНЕННЯ ЕФЕКТИВНОГО СЕМАНТИЧНОГО ПОШУКУ

В.М. Анохін, Б.О. Марчак, Д.В. Гріньов

Семантична павутина — група методів і технологій, створених для того, щоб дозволити машинам розуміти значення, — або «семантикові» — інформації про усесвітню павутину. Згідно початковому баченню, доступність машиночитаемых метаданих дозволила б автоматизованим агентам і іншому програмному забезпеченню розумніше діставати доступ до мережі. Агенти були б в змозі автоматично виконувати завдання і визначати місцеположення відповідної інформації від імені користувача. Тоді як слову «семантична павутина» формально не дають визначення, проте можна сказати, що це поняття головним чином використовується для того, щоб описати модель і технології, запропоновані W3C. Ці технології включають платформу опису ресурсів (RDF), безліч форматів обміну даними (наприклад, RDF/XML, N3, Turtle, N-Triples), і нотації, такі як схема RDF (RDFS) і веб-сервер-мова онтології (OWL), які призначені для того, щоб забезпечити формальний опис понять, терміни і відношення в межах даної галузі знань. Багато хто з технологій, запропонованих W3C, вже існує і використовується в різних проектах.

Ключові слова: *Semantic Web (семантична павутина), всесвітня мережа, глобальна мережа, зв'язок, ресурс.*

ANALYSIS OF SYNTHESIS AND SEMANTICS OF LANGUAGES OF QUERIES OF RDF DATA LAYOUT FOR REALIZATION OF EFFECTIVE SEMANTIC SEARCH

V.N. Anohin, B.O. Marchak, D.V. Grinev

A semantic spider web is a group of methods and technologies, created in an order to allow machines to understand a value — or to «семантику» — information about a world spider web. In obedience to initial vision, availability of metadatas would allow the automated agents and other software it is more reasonable to get access to the network. Agents would be able automatically to execute tasks and determine the location of the proper information on behalf of user. While to the word «semantic spider web» legalistically does not give determination, however possible it is to say that this concept is mainly utillized in an order to describe a model and technologies, offered W3C. These technologies include the platform of description of resources (RDF), great number of formats of exchange by information (for example, RDF/XML, N3, Turtle, N-Triples), and notations, such as a chart of RDF (RDFS) and web-language of ontology (OWL), which are intended in an order to provide the formal specification of concepts, terms and relation within the limits of this area of knowledges. Many of technologies, offered W3C, already exist and utillized in different projects.

Keywords: *Semantic Web (Semantic Web), a global network, global network, communications and resource.*