УДК 004.89

О.Н. Половикова

Анализ способов формализаций документов для выполнения семантического поиска

O.N. Polovikova

Analysis of Methods to Formalize Documents for Semantic Search

Рассмотрены подходы к формированию представлений документов для семантических поисковых систем. Выделены особенности и достоинства использования программ-агентов Акторного Пролога, призванных извлекать знания из информационных источников глобальной сети.

Ключевые слова: семантический поиск, метаданные, метаформат, онтология, Акторный Пролог.

This article describes the approaches to the formation of semantic representations of documents for the search engines. The features and advantages of the use of software agents Actor Prolog, designed to extract knowledge from information sources of the global network are considered.

Key words: semantic search, metadata, metaformat, ontology, Actor Prolog.

В данной статье рассмотрены подходы к формированию представлений документов для семантических поисковых систем, на реализации которых основываются решения базовых задач проекта Semantic Web. Основная идея данного проекта заключается в реконструкции существующего web-пространства для последующей семантической машинной обработки хранимых документов. Опубликованные web-документы следует преобразовать, а новые создавать таким образом, чтобы в них были заложены возможности автоматизированной обработки их смыслового содержания. Такое преобразование web-пространства в пространство знаний позволит всем заинтересованным службам (запросным приложениям, агентам, модулям), работающим в сети, извлекать из хранимого множества информационных ресурсов значимую и отвечающую по смыслу предъявляемым требованиям информацию, т.е. извлекать знания [1–3].

Несмотря на то, что некоторые основополагающие задачи проекта Semantic Web уже имеют реальное практическое решение, а заявленные технологии становятся стандартами, в то же время многие исследователи отмечают, что модернизация Web-пространства еще находится в самом начале своего развития [4]. Одной из основных задач концепции Semantic Web является решение проблем связанных с индексацией и поиском информации по смысловому содержанию. Идея создания универсального средства семантического поиска информации на уровне нескольких информационных систем или для web-пространства уже не является утопической и имеет реальное практическое воплощение, в качестве примера можно выделить семантическую поисковую систему SHOE. Построенные семантические поисковые системы – скорее, демонстрационные и исследовательские разработки, а не рабочие повседневные инструменты [4, 5]. Поэтому тематика данного исследования актуальна, приведенный анализ существующих подходов к формированию представлений документов для семантических поисковых систем востребован.

Процесс извлечения знаний из совокупности ресурсов или из информационной системы может быть основан не только на поиске определенных соответствий между метаданными аннотированных документов и поискового запроса. Поиск знаний можно реализовать с привлечением логического вывода, который позволяет выстраивать (восстанавливать) цепочки триплетов (объекта, атрибута и значения) – получать новые знания. Выполнение задач преобразования web-документов и семантического поиска напрямую связаны с разработкой и использованием специализированных языков для встраивания знаний непосредственно в сам документ либо для создания отдельных от ресурса описаний-заменителей. Данные языки призваны обеспечивать реализацию всех компонентов модели поиска: способа представления информационных документов (или их заменителей), способа формирования запросов, критерия релевантности web-документов запросу. Подходы к представлению знаний для информационных ресурсов, выставляемых в глобальной сети, узкоспециализированном хранилище или распределенной информационной системе, взаимозависят от способа построения поисковых запросов и используемого критерия для определения соответствия ресурсов запросу.

В результате анализа информационных ресурсов сообщества Semantic Web были выделены следующие способы создания семантических представлений (формализации) информационных документов для их последующей обработки модулями семантической поисковой системой:

1. Аннотирование документа с использованием метаформатов (структура**,** стандарт для описания метаданных):

– метаданные встроены в сам ресурс;

– описания сохраняются и обновляются независимо от ресурсов.

2. Описание онтологии предметной области и семантическая разметка ресурса при помощи понятий семантической модели (онтологии).

3. Построение базы знаний термов, описывающих знания нескольких информационных ресурсов, с использованием программ-агентов.

На сегодняшний день разработано множество схем описания метаданных для информационного ресурса, также активно развиваются и языки для описания метаданных. Базовыми стандартами для Semantic Web в данный момент признаются стандарты Dublin Core, FOAF, SIOC и DOAP. Чтобы встроить в документ его семантическое описание с использованием выбранной схемы построения метаданных, не требуется внутреннего языка представления данных, для этих целей можно использовать язык XML, специально разработанный формат описания ресурсов RDF. Метаданные, характеризующие смысловое содержание документа (контент) и предметной области (контекст), оформляются с использованием специализированных тегов – семантических тегов. Семантические теги стандартного HTML позволяют «внести знания» прямо в страницы [1].

Метаданные, оформленные с помощью языка RDF, могут быть встроены непосредственно в ресурс (MsWord документ или HTML-страницу), а могут сохраняться и обновляться независимо от ресурсов. Многие из производителей программного обеспечения уже выпускают ряд продуктов, которые автоматически формируют некоторый небольшой блок RDF-описания внутри документа. Второй подход более универсален, так как в этом случае метаданные могут быть созданы для любого ресурса [1]. В рамках проекта Semantic Web развитие получило направление по автоматическому созданию репозитория RDF-описаний ресурсов Интернет.

Для создания онтологий предметной области используют специально созданные для этих целей языки: Схема RDF (RDF Schema), OWL. Язык RDF Schema позволяет описывать структуру RDF-хранилища в терминах типов (классы, свойства) и отношений между ними, применяется для создания простых онтологий данных. Чтобы описать более сложные виды отношений, следует использовать расширенный вариант RDF Schema – язык OWL, который позволяет описывать не только классы и свойства, но также задавать ограничения на их использование.

Модели предметной области, построенные на языке OWL, могут быть опубликованы в Web и одновременно использоваться модулями различных информационных систем, для того чтобы строить актуальные знания из ресурсов специфицированных данными онтологиями. Формирование новой онтологии предметной области может базироваться на имеющихся в сети онтологиях. Для концептуализации содержания конечного ресурса можно применять существующие в сети онтологии, которые по необходимости следует настроить под его специфику. Семантическая разметка документов на основе выбранной (или построенной) онтологии характеризует не только содержание этих ресурсов, а также семантику различных сервисов, предоставляющих эти документы конечным пользователям.

Следует выделить два перспективных направления в развитии проекта Semantic Web, связанных с использованием онтологий: создание визуальных сред для работы с онтологиями (построение, модификация, сопоставление и т.д.), разработка агентов автоматического построения семантических карт. Визуальные среды позволяют специалистам непосредственно «рисовать» онтологии, что помогает наглядно сформулировать и объяснить природу и структуру явлений. Семантические карты описывают концептуализацию содержания ресурса в виде OWL онтологии.

Как уже было отмечено, все составляющие модели поиска должны быть взаимозависимы. Поэтому при поступлении в систему пользовательского запроса для него также строится соответствующее представление, а метод его построения аналогичен методу построения представлений документов. Разметка документов с помощью метаформатов или онтологических терминов позволит производить автоматическую обработку их семантического содержания. Среди специальных запросов, которые умеют работать с семантическим содержанием, следует выделить SPARQL и RDF Query, которые базируются на обработке направленных графов (RDF-графов).

Построение базы знаний термов, описывающих знания нескольких информационных ресурсов, может быть реализовано логическими языками программирования, например Акторным Прологом. Программы-агенты объектно-ориентированного Акторного Пролога позволяют извлекать данные из документов, опубликованных в сети Интернет, посредством предопределенного класса Receptor, преобразовывать их в различного рода термы (множества, списки, миры, структуры и т.д.), а затем использовать встроенный в язык механизм логического вывода для поиска информации по смысловому содержанию. Специальная стратегия логического вывода данного языка позволяет формировать новые знания для обрабатываемых документов.

Пролог-система обладает универсальным языком запросов, который полностью согласуется с БЗ термов. Таким образом, агентами Акторного Пролога обеспечивается реализация всех компонентов модели поиска.

Основные преимущества использования данного подхода для формализации документов основываются на комбинировании возможностей логического вывода и объектно-ориентированного подхода для описания взаимодействий между объектами. Применение параллельных процессов позволяет организовать независимую обработку информации из различных источников, при этом количество используемых ресурсов может быть заранее неизвестно. Если произойдет изменение ресурсов, которые были отобраны в результате работы поискового агента, это вызовет автоматическое обновление результатов поиска. При этом не нужно заново доказывать целевое утверждение, требуется повторно согласовать лишь некоторые подцели.

Необходимо заметить, что Акторный Пролог [6] разработан непосредственно для создания агентов, призванных выполнять поиск и распознавание информации в глобальной сети. Данный диалект Пролога содержит необходимый набор средства для программирования агентов, которые способны в непрерывном режиме отслеживать все изменения информационных документов в сети.

Проведенный анализ способов формализаций документов для выполнения семантического поиска позволяет сделать вывод о многонаправленности теоретических исследований и практических разработок в этой области. Выбор конкретного подхода к созданию представлений для публикуемых ресурсов зависит от различных факторов: специфики контента и контекста содержания информационного источника, характеристик информационной системы (в которой хранятся ресурсы), а также свойств технического и программного оснащения узлов сети (на которых размещена информационная система).

Библиографический список

1. Андон Ф.И., Гришанова И.Ю., Резниченко В.А. Semantic Web как новая модель информационного пространства Интернет [Электронный ресурс]. – URL: http://shcherbak.net/semantic-web-kak-novaya-model-informacionnogo-prostranstva-internet/

2. Басипов А.А., Демич О.В. Семантический поиск: проблемы и технологии. Вестник АлтГТУ. – 2012. – №1.

3. Рабчевский Е.А. Булатова Г.И. Автоматическое построение онтологий для тематических поисковых систем [Электронный ресурс]. – URL: http://shcherbak.net/avtomati-cheskoe-postroenie-ontologij-dlya-tematicheskix-poiskovyx-sistem/

4. Введение в Semantic Web // Компьютерный еженедельник UPGRATE [Электронный ресурс]. – URL: http://www.upweek.ru/vvedenie-v-semantic-web.html].

5. Зубинский А. Semantic Web // Компьютерное обозрение [Электронный ресурс]. – URL: http://ko.com.ua/semantic_web_13971].

6. Морозов А.А. Об одном подходе к логическому программированию интеллектуальных агентов для поиска и распознавания информации в Интернет [Электронный ресурс]. – URL: http://lvk.cs.msu.su/~bruzz/articles/web_retrieval/Morozov.pdf.