

10

Поиск для Web 3.0

Технологии «семантического веба» — главного претендента на роль следующего поколения развития Сети, которое уже окрестили Web 3.0, неторопливо, но верно обосновываются на все большем количестве интернет-сервисов. Появление новых проектов связано с актуальной и серьезной проблемой извлечения полезной информации из огромного объема представленных в современной Сети данных. Сбор, обработка и сведение воедино данных из множества источников остается весьма трудоемким занятием, которое было бы весьма заманчиво переложить на плечи компьютеров. Однако интернет в его современном состоянии не самым лучшим образом приспособлен для решения этой задачи. Причины этого довольно глубоки, можно сказать, концептуальны. При всем своем многообразии подавляющее большинство современных веб-проектов объединяет то, что они ориентированы исключительно на человеческое восприятие. Под это понятное и естественное требование изначально подгонялись форма и технологии создания веб-ресурсов. В конце концов, их же должен просматривать Homo Sapiens, который способен самостоятельно выделить смысл содержимого страницы.

Проанализировать смысл такой веб-страницы в автоматическом режиме — сложная задача. Что скрывается за текстом веб-страницы — ценные сведения и выводы, сделанные знатоками темы, просто бесполезная в данный момент информация или откровенная ерунда, компьютеру не ведомо. Попытка преодолеть данный семантический барьер «в лоб», на базе технологий современной сети, требует приложений, приближенных по уровню к искусственному интеллекту, что отодвигает их перспективы в неясное будущее, по крайней мере, для обычных пользователей Сети.

На этом этапе и возникает идея «семантического веба», позволяющая решить названные проблемы и во многом изменить само представление о способах работы в интернете. Она была сформулирована известнейшим

специалистом Тимом Бернесом-Ли почти два десятка лет тому назад. Кратко ее можно изложить так: «Если создание программ, способных автоматически обрабатывать сведения из интернета, является слишком сложной задачей, то имеет смысл создать надстройку к Сети в виде технологий, приводящих сетевую информацию в дружественный для программ-агентов вид». На основе соответствующей разметки сетевых документов, которая будет отражать смысл опубликованных данных, такие программы смогут самостоятельно собирать разрозненную информацию, сопоставлять данные, принимать решение об их достоверности и даже делать некоторые логические выводы, предоставляя на выходе пользователю готовый информационный продукт.

Сильной стороной семантического веба должен стать учет не только прямого содержания опубликованных в Сети документов, но и контекста использования информации. Различные фрагменты веб-страниц могут быть полезны при решении различных задач, надо только скомбинировать имеющиеся данные нужным образом.

Комплекс технологий, необходимых для реализации семантического веба, часто представляют в виде схемы, напоминающей слоеный пирог. Ключевые элементы данного пирога: универсальные идентификаторы ресурсов (Universal Resource Identifier, URI), основанные на XML языки семантической разметки сетевых документов и язык онтологий, позволяющий устанавливать связи и отношения между объектами. Важную роль в семантическом вебе играют технологии проверки достоверности найденных программами-агентами данных. Результатом развития таких технологий должна стать возможность простой автоматической проверки степени достоверности отдельных положений. Тим Бернес-Ли в шутку назвал такой идеальный инструмент кнопкой «Да ну?», после щелчка на которой должен появиться либо перечень заслуживающих доверия положений, либо сообщение об ошибке (куда же без него?), говорящее о неправильной записи или о том, что системе не удалось найти в документе ничего, заслуживающего доверия.

Развитием семантических технологий в настоящее время достаточно активно занимается организация World Wide Web Consortium. Среди успехов отметим стандартизацию двух ключевых технологий: формата семантических данных (RDF) и языка веб-онтологий OWL.

Семантические технологии изящны в теории, однако воспринимать их как действительно перспективное и серьезное направление развития Сети ранее мешало два обстоятельства. Первое — сдержанное поначалу восприятие новых технологий действительно крупными универсальными интернет-поисковиками, что сказывалось на популярности

семантический разметки контента разработчиками веб-ресурсов, поскольку внедрение семантической разметки не несло каких-либо немедленных дивидендов. Второе — недостаток простых в использовании и полезных в повседневной работе приложений, способных стать популярными не только среди фанатов, но и в широких кругах пользователей. Сейчас можно наблюдать серьезные подвижки в обоих направлениях.

Отрадно, что такая тенденция не прошла мимо Рунета. Поисковая система Яндекс не так давно научилась работать с данными в семантических форматах. Правда, их поддержка пока ограничивается только поиском в блогах. Для индексации профилей и выявления социальных связей используется формат FOAF. Более того, Яндекс разработал собственное расширение FOAF, позволяющее точнее описывать персональные данные пользователей своих социальных сервисов. Добавлять FOAF-разметку к веб-страницам предлагают самим пользователям блогов. Конкретные рекомендации можно найти в справке по прикладному программному интерфейсу (API) поиска в блогах Яндекса.

Постепенный рост объема данных в семантических форматах выявил потребность в своеобразных «точках доступа» — сервисах, аккумулирующих сведения о свободно доступных данных в семантических форматах, к которым могли бы обращаться как приложения-агенты, так и пользователи-люди. В «обычном» вебе эту роль играют поисковые системы и каталоги. Аналогичным путем пошла эволюция и в современном семантическом вебе. Такие поисковые системы иногда даже называют «семантическими приложениями второго поколения». Этим подчеркивается значительное улучшение удобства работы с информацией в новых форматах, в том числе для широких кругов пользователей. Данная глава посвящена лучшим и наиболее показательным представителям семейства поисковиков для семантического веба.

Требования к представленным в этой главе поисковым машинам были следующие. Все предлагаемые сервисы должны быть общедоступными. Поисковики обязательно должны были уметь работать с документами в основных форматах семантического веба. Поскольку такие ресурсы претендуют на роль универсальных точек доступа к семантическим данным, поисковики должны обладать удобным пользовательским интерфейсом, предназначенным для пользователя-человека, а также предусматривать возможность подключения к поисковику внешних приложений — программ-агентов, ориентированных на автоматические сбор и систематизацию данных.

Семантическая разметка веб-документов предлагает широкие возможности по составлению сложных запросов, во многом отличающиеся от

привычных опций «обычных» интернет-поисковиков. Эффективные поисковики должны обладать соответствующими инструментами составления сложных запросов, а также различными вспомогательными средствами последующего уточнения запросов и фильтрации результатов.

Swoogle

Ведущие роли в разработке инструментов для поиска в семантической сети играют исследовательские проекты различных научных и учебных заведений. Сейчас это настоящие локомотивы семантического веба. Выбор и возможности академических проектов пока заметно превосходят общедоступные предложения коммерческих разработчиков. Обзор поисковиков для Web 3.0 начнем рассказом о сервисе Swoogle, разработанном специалистами университета Мэриленда (США).

Название системы: Swoogle	URL: swoogle.umbc.edu
Язык интерфейса	Английский
Тип поиска	Поиск по Semantic Web
Удобство использования	Удовлетворительно
Качество поиска	Хорошо
Разработчик	Ebiquity group/UMBC

Swoogle индексирует документы в форматах N-Triples, RDF/XML и N3 (RDF). Обработываются как полностью составленные с помощью этих языков документы, так и «обычные» веб-страницы, содержащие фрагменты семантического кода. Индексная база у Swoogle собственная, построенная на информации из открытых сетевых источников, и на сегодняшний

день содержит сведения примерно о трех миллионах документов, о миллиарде отдельных высказываний-триплетов, а также десяти тысячах онтологий. Это достаточно серьезные цифры.

Воспользоваться поисковиком может любой желающий. На сервисе предусмотрена бесплатная регистрация. Без регистрации по запросу демонстрируется только первая сотня результатов. По всей видимости, это призвано несколько снизить нагрузку на серверы проекта, поскольку основное преимущество обладателей аккаунта — получение именно полных результатов поиска. Это немаловажно, поскольку Swoogle поддерживает подключение программ-агентов. Данные им выдаются в формате RDF.

Предназначенный для просмотра человеком интерфейс Swoogle весьма прост и лишен каких-либо излишеств. Доступно несколько специфических режимов поиска. Наиболее строгим является режим *Ontology*. При

поиске в этом режиме в результаты попадают только документы, содержащие определения свойств семантических объектов. Пожалуй, наиболее удобным при личном обращении пользователя к ресурсу является режим Document, в котором поиск ведется среди всех проиндексированных документов. Третий режим — Term — предлагает поиск отдельных триплетов. Он выдает максимальное количество результатов, однако удобен скорее для машинной обработки. Swoogle обладает собственным языком составления сложных запросов, который состоит из собственных операторов, а также поддерживает синтаксис запросов Apache Lucene. Перечень операторов Swoogle можно найти в приложении А. По умолчанию поисковый оператор и его значение разделяются двоеточием. Поддерживается поиск по фразе с использованием кавычек.

Страница выдачи Swoogle максимально проста (рис. 10.1). В списке предоставляются только ссылки на ресурсы и краткие комментарии. Каких-либо дополнительных инструментов просмотра или фильтров не предусмотрено. Компенсацией может служить достаточно высокая скорость работы сервиса.

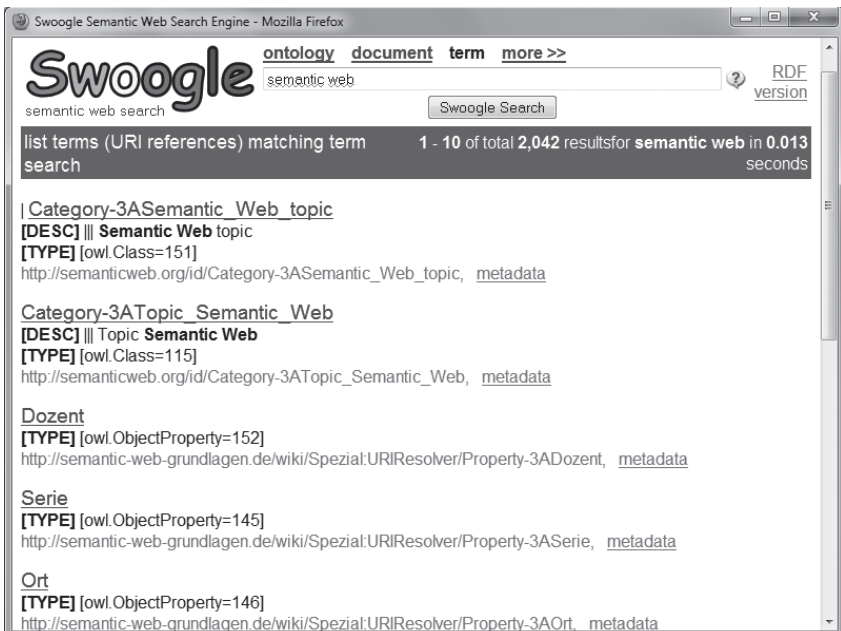


Рис. 10.1. Страница выдачи поисковой системы Swoogle

VisiNav

VisiNav — это еще один университетский проект поиска в семантическом вебе. У данного сервиса ирландские корни. Принадлежит он институту Digital Enterprise Research Institute (DERI). Размер индексной базы заявлен в объеме примерно двух миллионов документов и тринадцати с половиной миллионов высказываний-триплетов. Собирается база собственными роботами VisiNav. Главный недостаток VisiNav — редкое обновление базы.

Название системы: VisiNav	URL: visinav.deri.org
Язык интерфейса	Английский
Тип поиска	Поиск по Semantic Web
Удобство использования	Хорошо
Качество поиска	Удовлетворительно
Разработчик	DERI

Начинается поиск указанием ключевого слова, после чего система выводит начальный набор объектов, подходящих под условия запроса. Данные в VisiNav представлены в виде объектов-ссылок, которые обладают атрибутами и связями с другими объектами. Щелчками на связях или перетаскиванием объектов (такие ссылки

подчеркиваются красным) можно уточнять свой запрос. Поскольку семантическая разметка довольно подробно, одним движением мыши можно найти, например, всех партнеров или подрядчиков какой-либо фирмы. Разработчики старались сделать этот процесс наглядным, однако это обусловило и некоторые требования к браузерам. Пользователи Firefox и Opera последних версий могут не опасаться ошибок, а вот в отношении браузеров Internet Explorer и Chrome их отсутствие не гарантировано.

Поисковик предлагает несколько режимов просмотра результатов поиска. В режиме **List** найденные ссылки выводятся в виде простейшего списка. Отметим возможность экспорта результата в RSS-канал. В режиме просмотра **Detail** экран делится на две области. Справа выводятся теги и текстовые выдержки, а слева — свойства и ссылки. Заметим, что набор данных при этом на панелях одинаковый, это просто два варианта просмотра найденного. Система умеет находить тематические изображения и выводить их в виде миниатюр. Здесь также доступен экспорт данных в семантический формат RDF.

Кроме этих режимов, VisiNav предлагает еще два варианта просмотра списка выдачи, причем они у конкурентов не повторяются. Режим **Timeline** выстраивает все найденные ресурсы на интерактивной «линии времени», что позволяет отследить развитие той или иной темы. Интересно, что при

щелчках на объектах, располагающихся на линии времени, вызываются не исходные ресурсы, а окошки с конкретными текстовыми выдержками. При желании результаты поиска можно сразу же загрузить в формате iCal. В режиме Table результаты выводятся в виде таблицы. Ее колонки — теги семантической разметки, строки — найденные ресурсы. Такой режим просмотра, как и любая таблица, хорош для сравнения похожих значений у различных объектов. Результаты также можно загрузить в виде CSV-файла.

К сожалению, система VisiNav плохо работает с запросами, набранными кириллицей, хотя в результатах поиска попадают и русскоязычные ресурсы.

Watson

Проект Watson разработан в Великобритании. Заметную долю в его финансировании занимают гранты Евросоюза. Особенность данного поисковика — ориентация на достижение максимального качества поиска за счет отбора включаемых в индекс ресурсов. Особое внимание в данном проекте уделяется установлению смысловых отношений различных онтологий: поиску дублирующихся записей, работе с различными версиями, обработке значений похожих онтологий и другим приемам анализа.

Название системы: Watson	URL: watson.kmi.open. ac.uk
Язык интерфейса	Английский
Тип поиска	Поиск по Semantic Web
Удобство использования	Удовлетворительно
Качество поиска	Удовлетворительно
Разработчик	Knowledge Media Institute / The Open University

Поскольку количество данных в семантических форматах в Сети растет, все более заметным становится разница в качестве таких ресурсов. Если рассмотренные ранее поисковики стремятся охватить как можно больше документов, то целью Watson являются в первую очередь наиболее качественно представленные ресурсы. Это проявляется в наличии системы рейтингов, которые выставляются

системой тем или иным ресурсам в ходе анализа проиндексированных источников данных. Предполагается, что это позволит улучшить релевантность поиска. Нельзя не отметить, что подобный подход вполне соответствует классическим теоретическим подходам к созданию семантического веба. Пусть это еще не знаменитая кнопка «Да ну?» Бернеса-Ли, но определенно шаг в нужном направлении.

Предполагается, что данной системой будут пользоваться как программы-агенты, так и люди. Для подключения первых предлагается собственный прикладной программный интерфейс, для вторых — довольно простой, но любопытный визуальный интерфейс. Он состоит из поля поиска и небольшого набора дополнительных фильтров, вызываемого кнопкой Search Options. Здесь предлагают включить поиск по фразе, а также некоторые специфические фильтры. Так, например, фильтр Entities позволяет искать только среди классов или свойств объектов. Фильтр Score дает возможность ограничить выдачу значениями конкретных свойств объекта (URI, заголовок, комментарий).

Заметим, что список выдачи у системы Watson меньше по объему, чем у других поисковиков для семантического веба, однако это компенсируется обязательными ссылками на соответствующие онтологии (рис. 10.2). Каждая найденная ссылка ведет к своеобразному «паспорту» исходного ресурса, который содержит сведения о количестве элементов в онтологии, использованном для описания языке, а также комментарии пользователей.

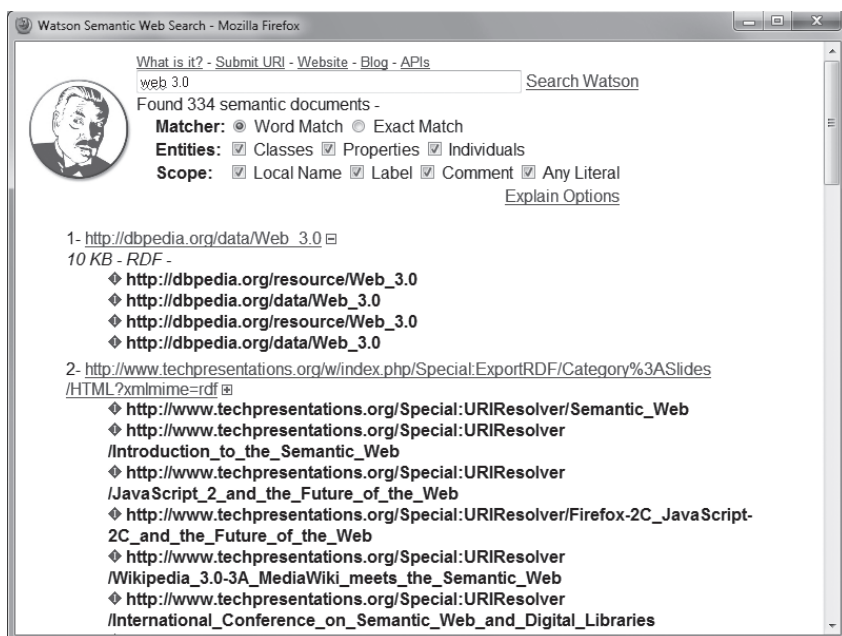


Рис. 10.2. Страница выдачи поисковой системы Watson

Кроме того, предлагается ссылка для загрузки файла онтологии. Очень интересна возможность сразу же составить и отправить запрос к найденному ресурсу на языке SPARQL.

Falcons

Интерес к семантическому поиску проявляют не только американские и европейские разработчики. Проект Falcons — это пример азиатского ресурса. Название поисковика представляет собой сокращение слов «Finding, Aligning and Learning ONtologies». Разработан поисковик в Юго-Восточном университете Китая (точнее, в институте Institute of Web Science). Любопытно, что даже сайт-«визитка» самого института широко использует RDF-разметку.

Название системы: Falcons	URL: iws.seu.edu.cn/ services/falcons/
Язык интерфейса	Английский
Тип поиска	Поиск по Semantic Web
Удобство использования	Хорошо
Качество поиска	Удовлетворительно
Разработчик	Institute of Web Science / Southeast University P. R. China

Поисковик в настоящее время умеет работать только с документами, использующими формат RDF/XML. Технологическая база Falcons полностью построена на программном обеспечении с открытым исходным кодом. Задействованы технологии MySQL, Apache Lucene и Jena — специализированная платформа для Java, предназначенная для работы с данными в семантических форматах. Для взаимодей-

ствия внешних приложений с Falcons предлагается достаточно простой прикладной программный интерфейс REST. Все соответствующие рекомендации и примеры уместились всего на одну веб-страницу справки Falcons.

Пользовательский интерфейс Falcons выдержан в строгом стиле. Начинать поиск можно только с помощью ключевых слов. Собственного языка запросов система не предлагает, ограничиваясь поддержкой поиска по взятой в кавычки точной фразе, а также стандартных логических операторов. Доступны несколько режимов поиска: объектов (Object), классов и свойств (Concepts), а также всех проиндексированных документов с семантической разметкой (Document).

Найденные объекты демонстрируются в виде простого списка сниппетов. Каждый сниппет содержит заголовок, URI, присвоенный системой тематический ярлык и краткую аннотацию, которая генерируется на основе

RDF-описания найденного объекта. Щелкая на ссылках, можно перейти к более подробным описаниям.

Интересная находка Falcons — боковая панель с иерархическим деревом классов, к которым относятся найденные объекты. Она позволяет быстро уточнить запрос и упрощает навигацию в больших списках результатов. Данный инструмент напоминает панели кластерных интернет-поисковиков, однако шума здесь гораздо меньше, а структурирование информации выглядит весьма качественно. Достигается это за счет анализа семантической разметки документов.

Semantic Web Search

Коммерческих предложений в области поиска в семантическом вебе пока немного. Упоминания заслуживает проект Semantic Web Search. Основной бизнес фирмы-разработчика Semantic Web Search — внедрение собственного коммерческого программного обеспечения для хранения и обработки данных в формате RDF. Свободно доступный сервис Semantic Web Search носит демонстрационно-рекламный характер.

Название системы: Semantic Web Search	URL: www. semanticwebsearch. com
Язык интерфейса	Английский
Тип поиска	Поиск по Semantic Web
Удобство использования	Удовлетворительно
Качество поиска	Удовлетворительно
Разработчик	Intellidimension

Сведений о размере индексной базы проекта на сайте Semantic Web Search не приводится, однако если судить по результатам тестовых запросов, она заметно уступает индексам некоммерческих академических поисковиков, о которых шла речь ранее.

В отличие от обычных поисковиков, на данном ресурсе можно составлять запросы, позволяющие,

к примеру, найти все статьи человека с определенным именем и фамилией. Для создания сложных запросов применяется собственный язык запросов, использующий FOAF-разметку ресурсов.

Данный проект интересен в первую очередь своими дополнительными инструментами. Начнем с того, что на сайте есть аналог «расширенного поиска» — приложение Search Agent, которое помогает в составлении сложных запросов. Более того, на сайте предлагается простой онлайн-редактор Agent Developer, с помощью которого можно создать собственную программу-агента. Для этого понадобятся только самые общие представления о составлении поисковых запросов, остальные операции