СЕМАНТИЧЕСКИЙ ВЕБ И АСТРОНОМИЯ

О.С. Бартунов , С.В. Карпов

Семантический Веб - следующая стадия развития всемирной паутины. Основной идеей его является переход от представления информации в виде, оптимальном для человека, к более машино-читаемым форматам. Целью является организация возможности полностью автоматического извлечения знаний из хранящихся в Сети огромных массивов данных - очевидно, что ручной ее поиск, даже с помощью современных поисковых машин, становится все более сложным.

К сожалению, большая часть существующего в современной Сети контента не предназначена для автоматической обработки - с машинной точки зрения, она представляет собой просто текст. Выходом может служить добавление к нему - фактически, к обычной веб-странице - специальной разметки, понятной программам и отмечающей части текста, несущие определенную семантическую информацию — так называемые микроформаты. Примером может служить сервис Wikipedia — свободно пополняемая онлайн-энциклопедия, содержащая уже более трех миллионов словарных статей. Ее тексты содержат микроформаты, которые могут выделять в словарных статьях различную информацию - имена и фамилии упомянутых людей, даты, географические координаты и так далее. Извлечение ее (этим занимается, к примеру, проект WikiDB) и представление в машиночитаемом виде позволяет использовать накопленные энциклопедией знания, к примеру, для организации семантического поиска - поиска, "понимающего" суть запроса. Некоторые современные поисковые системы (Goole, WolframAlpha) уже реализуют подобную возможность - при запросе "supernova type of sn 1987a" вы получите ответ "SN 1987A -Supernova Type: Type II-P (Unusual), According to http://en.wikipedia.org/wiki/SN_1987A", а не просто ссылки на тексты, содержащие эту фразу.

Основным форматом хранения информации в рамках семантического веба является набор спецификаций W3C под названием Resource Description Framework - RDF. Он базируется на представлении знаний в виде элементарных утверждений-триплетов вида "субъект - предикат - объект". На основании подобных утверждений могут быть построены произвольно сложные информационные структуры, такие, к примеру, как онтологии предметных областей. В астрономии, в рамках проекта "Виртуальная Обсерватория", разрабатываемая онтология (http://www.ivoa.net/Documents/latest/AstrObjectOntology.html) описывает структуру и взаимосвязи различных классов объектов и понятий, основываясь на элементарных утверждениях вида "Планетарная Туманность содержит Центральную Звезду" либо "Фотосфера является частью Звезды". Такого рода онтологии могут использоваться как для базовой классификации понятий, так и для комплексных процедур выведения зависимостей между ними и обеспечения непротиворечивости описания предметной области, что особенно важно при взаимодействии разнородных программных агентов.

Другим важным понятием семантического веба является Uniform Resource Identifier, URI - уникальный идентификатор ресурса, которым может быть документ, изображение, сетевой сервис или физический объект.

В идеале, процесс получения наблюдательных данных должен сопровождаться и созданием соответствующей мета-информации, их описывающей. Реализация полноценной обработки накопленного материала и публикации исключительно "science-ready", пригодной для непосредственного анализа информации в большинстве случаев оказывается очень сложной. Очевидно, что поэтому для обеспечения последующего использования этих данных, они должны сопровождаться полноценным описанием как процесса их получения, так и методов анализа, к ним применимых. В случае каталогов наблюдательных данных это сводится, в частности, к описанию семантического значения их отдельных полей - например, необходимо указать, какие поля описывают положение объекта - его координаты, в какой именно системе, и как они могут быть впоследствии преобразованы. Неоднозначность представления одной и той же информации в различных каталогах приводит к задаче построения онтологий физических величин, используемых в данной области знания. При наличии таких онтологий возможна уже организация полноценного автоматического извлечения информации их разнородных источников. Основная идея может быть продемонстрирована следующим примером.

Пусть необходимо найти изображения галактики до взрыва Сверхновой 1972e. Типичные действия астронома - из каталога Сверхновых звезд извлечь координаты Сверхновой 1972e и найти все снимки области неба с центром, имеющий эти координаты, в избранных обзорах неба. Это рутинная задача требует много времени и вполне могла быть автоматизирована, если только компьютер смог:

1. Узнать, что данные о Сверхновых находятся в каталоге Сверхновых звезд
2. Найти адрес ресурса, который содержит каталог Сверхновых звезд, желательно его актуальную и аутентичную версию
3. Обратиться к ресурсу, сформулировав запрос согласно его API
4. Понять какие колонки в результирующей строке являются координатами
5. Найти адрес ресурса, который содержит информацию об обзорах
6. Обратиться к ресурсу, задав координаты центра поля, сформулировав запрос согласно его API

7. Полученные адреса изображений передать астроному

Проблемой, стоящей на пути прямого приложения понятий семантического веба к современной науке, в частности - астрономии, является большой объем уже накопленной информации, часто представленной в свободном текстовом виде, для которой отсутствуют какие-лобо метаданные. Знания, содержащиеся в огромном корпусе статей и препринтов, не могут быть впрямую использованы программами. Извлечение семантически значимой информации из текстов может проводиться методами автоматической их классификации - выделения ключевых слов и понятий предметной области и построения их взаимосвязей на основании совместной встречаемости в отдельных текстах либо в работах отдельных авторских коллективах. Это позволяет строить актуальную (меняющуюся со временем в процессе развития и накопления знаний) семантическую карту - онтологию - предметной области. Актуализация такой онтологии, переход от общих утверждений вида "Планетарная Туманность имеет Центральную Звезду" к частным, объединяющим известные астрономические объекты "Остаток Сверхновой M1 имеет в центре Нейтронную Звезду PSR B0531+21", решает задачу извлечения всей доступной информации о заданной области неба (дата-майнинг) путем построения иерархической структуры всех попадающих в нее объектов и их составных частей.

Работа выполнена при поддержке грантов РФФИ 09-07-00499-а, 09-07-00471-а.