# О семантическом поиске в электронной библиотеке вуза

## О. Ю. Насадкина, А. Г. Марахтанов

Петрозаводский государственный университет

Петрозаводск

onasad@petrsu.ru, marahtanov@petrsu.ru

В настоящее время наблюдается значительный интерес со стороны научного сообщества к проблемам семантического поиска информации. Действительно, данный вид поиска должен обеспечивать лучшую релевантность результатов запросу пользователя, поскольку осуществляется не только на основе статистических наблюдений о встречаемости терминов в документах, но и с учетом семантики, смысла текста. Кроме того, подобные системы должны не только выдавать список ресурсов, содержащих ответ на запрос, но и указывать непосредственно ответ на заданный пользователем вопрос.

Большое развитие получили модели представления знаний для семантического поиска ресурсов в сети Интернет, разработаны языки описания онтологий, такие как RDF и OWL. Страницы в Wikipedia описываются на основе модели представления знаний semantic wiki. Семантически размеченный интернет уже получил название Web 3.0.

Для осуществления глобального семантического поиска в сети Интернет существуют такие поисковые системы, как зарубежные SearchMonkey от Yahoo, Powerset, Freebase, отечественная AskNet. Для семантического поиска в коллекции данных разработаны информационные системы, такие как BRICK (средство поиска в цифровых ресурсах, содержащих знания о культуре), SIMILE и FEDORA (поиск в цифровых ресурсах различного типа). Из российских разработок можно отметить проект Семантической библиотеки Свердловской области (http://semantic.uraic.ru), который в настоящее время находится в стадии разработки. Кроме того, построены и доступны для использования различные онтологии верхнего уровня, такие как OpenCyc, DOLCE, SUMO, содержащие универсальные базовые понятия и отношения.

В то же время многие из существующих на данный момент систем семантического поиска на практике не дают результаты лучше, чем классические средства. Это обусловлено тем, что для осуществления семантического поиска часто используют средства ввода и обработки запросов, пришедшие из классических систем информационного поиска, а также тем, что поиск осуществляется по коллекции, относящейся к большому числу различных предметных областей, в которых есть пересечения используемых терминов и понятий, осложняющие автоматическое индексирование и распознавание семантических конструкций в текстах.

Таким образом, повышению качества поиска (его точности или полноты) в семантических базах знаний будет способствовать:

* применение специально разработанных для этого интерфейсов, отличных от принятого во многих системах способа – ввода запроса на естественном языке в строку. При этом серьезной задачей является разработка таких интерфейсов, которые, с одной стороны, позволяли бы указывать дополнительные параметры в запросе (значимые свойства объектов, виды отношений и т. п.), с другой стороны, были бы понятны пользователю и не требовали значительных дополнительных временных затрат на ввод данных;
* сужение предметной области и использование специально разработанных для выбранной предметной области концептуальных моделей знаний (онтологий).

С учетом обозначенных выше положений предполагается разработка системы семантического поиска информации в коллекции текстовых ресурсов схожей тематики. В качестве предметной области выбрана краеведческая информация о регионе России (Карелии). В качестве текстовых документов – источников информации выбран корпус текстов о Республике Карелия из фондов Электронной библиотеки Республики Карелии (более 1500 ресурсов соответствующей тематики).

Данная работа будет состоять из следующих этапов:

1. Построение концептуальной модели знаний о предметной области, содержащей основные типы объектов, свойств объектов и отношений базы знаний.
2. Разработка и применение методов семантического индексирования, в том числе методов лексико-морфологического и синтаксического анализа текстов (что позволит из множества терминов сформировать массив слов и словосочетаний в начальной форме и их характеристик) и методов семантического анализа и разметки текстов (на основе концептуальной модели).

На этих этапах предполагается использование специально отобранных и подготовленных специалистами словарей, как общих (словари синонимов, омонимов, аналогий), так и специфических для выбранной предметной области словарей наименований географических, административно-территори-альных объектов, учреждений, организаций и ведомств региона и т.п.).

1. Реализация методов и моделей информационного поиска, учитывающих семантику и зависимость между терминами.
2. Использование методов построения баз данных и информационных систем, методов реализации веб-интерфейсов, методов программной реализации поисковых моделей. Основными требованиями к веб-интерфейсам, реализующим взаимодействие осуществляющих запросы пользователей с базой знаний, должны являться: скорость обработки запроса, наглядность представления данных, простота использования, возможность объяснения полученных результатов, возможность отсылки к источнику информации – текстовому документу коллекции.

Представляется необходимым при разработке обозначенной системы принимать во внимание следующие обстоятельства, которые смогут повысить качество выполняемой работы и качество семантического поиска, осуществляемого посредством системы, в целом:

* тщательно анализировать поисковые потребности пользователей и учитывать их при разработке концептуальной модели предметной области;
* учитывать опыт построения других систем семантического поиска;
* использовать существующие стандарты в данной области (языки представления онтологий RDF или OWL).

Созданная в соответствии с обозначенными в данной работе принципами система семантического поиска станет важным элементом Электронной библиотеки Республики Карелия (http://elibrary.karelia.ru), разрабатываемой сотрудниками Регионального центра новых информационных технологий Петрозаводского государственного университета. Она позволит эффективнее решать задачи поиска, возникающие в процессе обучения или научного исследования, связанного с использованием ресурсов коллекции [1, 2].

Предполагается, что содержательное и технологическое развитие Электронной библиотеки Республики Карелия будет способствовать повышению эффективности ее использования в учебной и научной деятельности, росту числа посещений библиотеки, росту числа научных исследований, проводимых с использованием документов, представленных в библиотеке.

**Библиографичекий список**

1. Марахтанов А. Г. Совершенствование Электронной библиотеки Республики Карелия за счет внедрения новых поисковых и навигационных сервисов // Информатизация образования и науки. 2010. ¹ 4 (8). С. 46–55
2. Марахтанов А. Г.  Совершенствование системы поиска в Электронной библиотеке Республики Карелия // Материалы XII всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2010. Казань, 2010. С. 465–468.