Технология поиска данных в информационных источниках web-портала

А.В. Босов, Р.Б. Чавтараев

Аннотация. Рассматривается программное решение, вошедшей в состав служб действующей версии Информационного web-портала с целью расширения его функциональности услугами поисковой системы. Основу решения составляет применение функционала портального сервера интеграции и доступа [1] и инфраструктуры портала [2] в интересах поиска информации во внешних информационных источниках. Рассмотрена архитектура и программная реализация решения, приведены примеры.

Введение

Поиск информации – одна из первых задач, которую пытается решать пользователь, подключившись к Интернет. Другая сторона предоставляет информацию, стараясь обеспечить ее легкую доступность, т.е. заинтересованный разработчик информационной системы для Интернет должен предложить пользователю удобные инструменты для поиска предоставляемых ресурсов. В современном web-пространстве имеется немало вариантов организации поисковых систем. Но и по настоящее время фактами остается то, что, во-первых, для успешного поиска нужной информации пользователю требуется иметь довольно высокую квалификацию, во-вторых, подавляющее большинство запросов остаются с нерелевантными ответами: информационного «мусора» больше, чем полезных ссылок.

Такое положение сформировалось постепенно. Когда Web только появился, было мало ресурсов, они были статичными, и простейшие поисковые машины, индексируя тексты, решали проблемы немногочисленных пользователей. Затем данных стало больше, они перестали храниться в статике, а размещались, как правило, в базах данных, работать с которыми поисковые машины уже не смогли. Можно считать, что так контент Интернет поделился на «поверхностный» (Surface Web) и «глубинный» (Deep Web). Любопытно, что оценки, характеризующие объемы этих частей (например, в [3] дается оценка отличия объемов в 400 – 550 раз), практически однозначно говорят о том, что искать нужную информацию в «поверхностном» Web, индексируемом глобальными поисковыми машинами, бесполезно.

При этом подавляющее большинство ресурсов предоставляет собственные средства поиска «внутри» себя, которые дают заведомо релевантные результаты. Таким образом, можно с успехом создавать централизованные системы, объединяющие всю информацию в рамках некоторой достаточно широкой темы. Известны положительные результаты такого подхода [4,5], но все-таки очевидно, что он не универсален, как из-за объемов информации, так и из-за сложности поддержки контента. Таким образом, для совершенствования поиска в Интернет нужны поисковые службы, обладающие некоторыми интеллектуальными возможностями.

Частично улучшают ситуацию web-порталы, но в основном не за счет особо удобных систем поиска, а за счет ограничения на индексируемые ресурсы, построение неких тематических каталогов, «вертикального» объединения пользователей

и т.п. Эффективность при этом обеспечена тем, что найдя нужный портал, пользователь дальше его средствами ищешь только в той предметной области, которую действительно отражают ресурсы портала. Но и это решение нельзя признать окончательным. Во-первых, есть области, где искусственно ограничить тематику довольно трудно. Например, в данной работе обсуждается портал Российской академии наук, для которого охарактеризовать тематику можно было бы как «научная информация», но, очевидно, что на самом деле такое «ограничение» мало что проясняет. Во-вторых, предъявляются очень жесткие требования и ограничения к ресурсам, подключаемым к порталам. Обычно предлагается ограниченный перечень адаптеров к ограниченному же набору распространенных систем, расширение же списка – крайне трудоемкая работа, связанная с программированием именно для того портального решения, что используется. Конечно, полностью отказаться от разработки нельзя, но кое-что усовершенствовать в технологии поиска в web-портале можно. Именно, можно минимизировать усилия разработчика (администратора или даже модератора), предложив более гибкие средства интеграции ресурсов для целей поиска.

В данной работе рассмотрено решение, реализованное в рамках проекта создания Информационного web-портала, используемого в настоящее время в качестве официального представителя РАН в сети Интернет (www.ras.ru), с целью усовершенствования механизма поиска информации в научном информационном пространстве РАН. Для этого портала есть и упомянутая широкая область тематики, и внедренное и эксплуатируемое решение, к которому периодически подключаются новые ресурсы, в т.ч. и в большей степени для поиска по ним, что делает поставленную задачу весьма актуальной.

1. Варианты организации поиска в web-системах

Для определения места обсуждаемой далее разработки приведем некоторые общие сведения о информационно-поисковых системах. Первый самый известный представитель таких систем – глобальная поисковая машина. С ее помощью для множества сайтов, подключенных к соответствующему сервису индексации, могут выполняться полнотекстовые запросы. Функционирование глобальной поисковой машины обеспечивают роботы (crawlers), которые выполняют просмотр ссылок для зарегистрированных сайтов и индексируют текстовое содержание страниц, возможно с учетом простейших метаданных. Обработку полнотекстовых запросов выполняет поисковый сервер, использующий построенный при сканировании индекс. Алгоритм работы поискового сервера может основываться на разнообразных методах информационного поиска [6-8], в т.ч. могут учитываться особенности терминологического состава проиндексированных документов, структуры хранилища, тематические каталоги, различные статистические методы и др. В любом случае пользователь будет пытаться общаться с поисковой машиной на естественном языке той предметной области, которую он имеет в виду, но о которой вряд ли что-то «знает» поисковая машина. Хотя семантическому анализу запросов и ресурсов посвящено множество исследований, но реальных решений, доведенных до практического использования в Интернет, пока нет. Кроме того, поиск изначально осуществляется только в «поверхностном» Web.

Из сказанного вовсе не следует, что глобальные поисковые машины не нужны. Несмотря на кажущуюся ограниченность, их место в Интернет прочно и надолго закреплено, по крайней мере, в качестве начальной точки при поиске ресурсов. Кроме того, для большинства известных глобальных поисковых машин наблюдается и определенный прогресс. Именно, многие сайты, изначально созданные для предоставления услуг по поиску в Интернет, например, Yahoo, AltaVista, Yandex, Rambler, постепенно превращаются, а то и уже оформились, в качестве полноценных «горизонтальных» порталов. В этих порталах объединены функции уже не одной поисковой службы, а нескольких сервисов, которые в т.ч. позволяют более точно определиться с предметной областью, интересующей пользователя, и уточнить запрос (даже если это происходит и неявно для пользователя). Конечно, этого все еще мало. Релевантность в любом случае мала,

и, к сожалению, чем более подготовлен пользователь, чем точнее он готов ограничить условия поиска, тем сложнее ему найти инструмент, удовлетворяющий его потребностям.

Другой класс систем, обеспечивающих пользователей услугами поиска – специализированные предметно-ориентированные системы. Узкая специализация и возможность доступа к «глубинному» web-контенту, охватывающему всю специализацию, в рамках одной системы, позволяют обеспечить значительно более высокое качество результатов поиска. При этом и пользовательский интерфейс (как при формировании поискового запроса, так и при выводе результатов поиска) имеет максимально удобный вид. В рамках одной системы, как правило, легко структурировать тематическую информацию, сопроводив документы (объекты) формальными описаниями связанных с ними терминов предметной области, в т.ч. с использованием словарей и классификаторов.

В итоге появляется возможность сформировать так называемый атрибутный (атрибутно-полнотекстовой) запрос, максимально детализирующий потребности пользователя. Проблемой остается обнаружение самой специализированной предметно-ориентированной системы и ее взаимодействие с другими системами, ориентированными на ту же самую или тематически близкую предметную область и, возможно, не менее полезными. Если для преодоления первой проблемы можно создавать некие глобальные предметные каталоги (это, конечно, несложно, но вот эффективность будет с ростом числа систем стремится к нулю), то вторая проблема требует уже создания некоторой распределенной среды, заставить в которой работать «старые» узкоспециальные системы непросто.

Таким образом, нужны другие технологии. Ключевая идея в этой связи одна, хотя ее можно и по-разному интерпретировать. Нужен некий представитель, хорошо разбирающий в узкой предметной области, и каталог (рубрикатор, онтология, база знаний и т.п.) «известных» предметов. В этой идее нет ничего революционного, она уже давно и успешно используется в электронных библиотеках, равно как и рассуждения о возможности переноса идеи адаптеров в Интернет [9]. Электронные (цифровые) библиотеки – важнейший класс информационно-поисковых систем. Реализуемые в них технологии основаны на понятии метаданных, позволяющих создавать коллекции распределенных ресурсов, поддерживаемых самостоятельными системами, вошедшими в коллекцию на федеративной основе (по принципу разделения ответственности за информацию между организациями, входящими в систему): коллекции распределенных в Интернет ресурсов обмениваются описательной информацией, формируя репозиторий метаданных, над которым работает индексная служба и служба поиска. Для организации такой распределенной среды применяются хорошо известные и распространенные протоколы, такие как HTTP, Z39.50, LDAP, CORBA.

В качестве стандартного примера метаданных репозиториев электронных библиотек можно привести библиографические атрибуты документов: название, аннотация, рубрики, ключевые слова, авторы и т.д. Экстрагированные метаданные, описывающие документ и его местонахождение, хранятся на поисковом сервере библиотеки и используются пользователем для поиска информации: на основе метаданных строятся индексы (в т.ч. и полнотекстовые) и выполняется атрибутный поиск. Аналогично поиску в специализированных системах за счет уточнения семантики искомых термов можно задавать условия на атрибуты. Результат такого поиска будет максимально удовлетворять сформулированному пользователем запросу, но при этом все-таки выполняться запрос будет над некоторым централизованным (быть может, виртуальным) репозиторием метаданных, сами же данные задействовать в поиске, вообще говоря, не предполагается. Кроме того, входящие в библиотеку системы должны следовать некоторой общей схеме репозитория (метарепозитория), по крайней мере поддерживать собственную подсхему, согласующуюся с метарепозиторием. Таким образом, сложности есть и здесь.

Из сказанного можно сделать вывод о наиболее перспективном направлении в развитии информационно-поисковых систем. Есть две хороши идеи: порталы и электронные библиотеки. Объединение и развитие этих концепций

непосредственно в интересах поиска должно дать выигрыш в доступности нужных данных для конечного пользователя.

2. Определение используемых технологий

Служба поиска Информационного web-портала, которой посвящена данная работа, естественно формировалась на основе других решений, реализованных в портале. Подробнее эти решения описаны в [1, 2], здесь же просто кратко их перечислим.

2.1. Форматы и технологии обмена данными

Предъявляемые в рамках портала требования к формату обмена данными таковы:

- универсальные выразительные возможности;
- синтаксическая интероперабельность;
- семантическая интероперабельность.

Требования к технологии обмена данными таковы:

- технологические решения должны максимально опираться на открытые стандарты манипулирования с данными;
- технологические решения должны способствовать распространению и использованию других технологий и служб, уже реализованных в портале (таких как подсистема безопасности, поддержка мультиязычности и др.);
- предъявляемые требования и влияние на подключаемые системы (информационные источники) должны быть минимизированы.

Удовлетворяются перечисленные требования за счет использования следующих решений:

- для обмена данными с информационными источниками используются формат XML (для неструктурированных данных), форматы реляционных данных (для структурированных данных). Язык XML обеспечивает все предъявленные требования за исключением семантической составляющей. Этот же формат используется для обмена неструктурированными данными между внутренними службами и подсистемами портала;
- технология обмена данными с информационными источниками – распространенные стандарты доступа к данным (ODBC, OLEDB, Web-services и т.д.);
- унифицированный формат для обмена структурированными данными между внутренними службами и подсистемами портала основан на реализации реляционной модели данных в используемой платформе .Net (класс DataSet);
- формат описания структур данных – XML-Schema [10]. Добавляет семантическую составляющую к XML.

2.2. Средства интеграции портала

Поскольку речь идет о функционале для распределенной системы и о взаимодействии с неоднородными источниками информации, то по аналогии с электронными библиотеками, делается вывод о необходимости слоя промежуточного программного обеспечения (mediator middleware). Через этот слой сервер интеграции и доступа [1] осуществляет передачу запросов и получение данных федеративных информационных источников. Как и в случае электронных библиотек с этой целью используются адаптеры: при получении запроса адаптер обращается к источнику через предоставляемый источником интерфейс, получает от источника данные и конвертирует их в некий общий для системы целевой формат.

Задача поиска в распределенной среде накладывает свою специфику, в т.ч. и на взаимодействие с информационными источниками. Во-первых, присутствуют только запросы на получение данных, и отсутствуют запросы на модификацию. Во-вторых, спектр запросов на получение данных может быть достаточно широк и включать в себя массу условий. При этом, несмотря на то, что каждая система хранения данных оперирует запросами для выборки данных, и поддерживает, как правило, некоторый язык запросов, пользователю надо предоставить все же интерфейсную форму с элементами управления для определения значений поиска, условий и т.п. Естественно, что ни о каком языковом выражении запроса (даже с учетом того, запрос может быть выражен либо в текстовом представлении, т.е. в виде команды на каком-либо языке, либо в виде программной функции с параметрами) речи быть не может, и допустимый максимум в данном случае – это использование условных операторов «AND», «OR» и т.п. в поисковом поле.

В портале реализован и другой традиционный элемент электронных библиотек – виртуальная (каноническая) схема. Виртуальная схема по сути состоит их метаданных, описывающие все сущности, с которыми оперирует служба поиска наряду с другими подсистемами портала. Она представляет собой набор типов (описаний объектов), их мультиязыковые представления, связи и различные технологические атрибуты. Пользователю предлагается формулировать запросы в терминах этой канонической схемы. Адаптер информационного источника отвечает за реализацию набора типов (подсхемы), которые поддерживает источник. Его задача – выполнить динамическую генерацию команды и списка параметров со значениями на основе данных, переданных поисковой формой. Далее сформированная команда выполняется адаптером в терминах информационного источника, используя протоколы доступа к данным, реализуемые источником.

Данные, полученные из адаптера, хотя и имеют унифицированный формат, однако нуждаются в приведении к терминам виртуальной схемы. В задачу адаптера, таким образом, входит преобразование полученных данных. Например, для поисковой web-системы результатом поиска должен являться URL объекта поиска. В результирующем наборе, переданном адаптером, может в явном виде не содержаться поля URL, однако его можно сформировать из других полей. Адаптер решает и эту задачу.

2.3. Место службы поиска в инфраструктуре портала

Информационный web-портал изначально создавался для аккумуляции научных ресурсов и, что самое важное, предоставления к ним эффективного доступа всем категориям заинтересованных пользователей. Технология доступа к научным ресурсам, которые не были представлены в web-пространстве, или даже в какомлибо цифровом виде, обеспечена в существующей реализации портала средствами, не рассматриваемыми здесь. На данный момент портальная инфраструктура позволяет создавать достаточно богатый контент и манипулировать различной информацией, а также создавать и управлять ее представлением. При этом поисковый сервис (как контекстный, так и атрибутный) по этим данным (внутренним ресурсам) реализован в полном объеме.

Значительный пласт представленной на портале информации обеспечивают внешние системы, например, созданные ранее web-сайты и базы данных. Для этих данных и требуется совершенствовать поисковые механизмы. Базовой задачей службы поиска портала является включение этих данных в распределенную web-среду портала с целью предоставления единого поискового инструмента, который обеспечивает эффективный информационный поиск с учетом предметной релевантности.

В составе Информационного web-портала служба поиска оформлена как самостоятельная подсистема, реализующая взаимодействие с различными информационными источниками для выдачи поисковых запросов, получения, преобразования и форматирования совокупных данных с целью предоставления их пользователю через сайт портала. Очевидно, что служба поиска интегрирована с другими важнейшими службами портала: подсистемой безопасности, системой управления содержанием и пр.

3. Архитектура и реализация поисковой службы

Служба поиска Информационного web-портала обеспечивает поиск данных по всем подключенным к порталу информационным источникам. Основной подход при выполнении операций поиска заключается в том, что службой поиска предоставляется web-интерфейс для ввода критериев поиска, формирование поискового запроса и его передача серверу интеграции и доступа, который в свою очередь посредством адаптера передает запрос системе-владельцу данных, а система-владелец данных должна сама обеспечивать поисковый сервис над своими данными. Служба поиска выступает как получатель массивов найденной информации: полученные данные форматируются, объединяются, и, если необходимо, сортируются, после чего выдаются пользователю.

Службой поиска также используется виртуальная схема, поддерживаемая сервером интеграции, описывающая структуры данных и

расширенная метаданными, предоставленными информационными источниками, содержащими описания способов доступа к информации. Вся совокупность сформированных таким образом метаданных размещается в файлах XML и условно делится на репозиторий типов и систему реализации типов (подробнее, в [1]).

Собственно же служба поиска функционально реализуется следующими компонентами:

- *шаблоном поисковых форм*. При помощи этого шаблона формируется web-интерфейс для поиска. Шаблон поисковых форм представляет собой APSX-страницу, обеспечивающую формирование страницы поиска на основе *репозитория типов* и с использованием *процессора форм*, и позволяет передать заданные пользователем критерии поиска *процессору запросов* для обработки;

- *шаблонами вывода результатов,* представляющими собой APSX-страницу, которая осуществляет форматирование и вывод результатов поиска. Этот шаблон получает результирующий набор данных от *процессора запросов,* форматирует и выдает результат в виде web-страницы;

- *процессором форм*. Этот компонент на основе *описания форм* обеспечивает функциональность для элементов поисковой формы (элементов управления). Может использовать *процессор запросов*, например, для построения списков выбора значений или других целей.

Используемый службой поиска процессор запросов, входящий в состав сервера интеграции и доступа, осуществляет взаимодействие с адаптерами информационных источников с целью выполнения поискового запроса и формирования общего результата поисковой операции. Процессор запросов использует систему реализации типов для построения поисковых запросов и их выполнения.

Последовательность выполнения поисковой операции выглядит следующим образом. Шаблон поисковых форм, используя репозиторий типов из виртуальной схемы портала и процессор форм, формирует web-форму для пользователя. Пользователь при помощи элементов управления задает критерии поиска. Шаблон поисковых форм при помощи процессора форм строит набор поисковых условий (поисковый запрос) и передает процессору запросов для обработки. Процессор запросов взаимодействует с адаптерами с целью получения данных и консолидирует результаты. Итогом работы процессора запросов является результирующий набор данных, который форматируется и выводится при помощи шаблона вывода результатов.

Далее подробнее рассмотрены работа перечисленных компонентов и используемые службой поиска метаописания.

3.1. Репозиторий типов

Репозиторий типов является той частью виртуальной схемы портала, которая отображает общую структуру данных, абстрагированную от деталей реализации доступа и преобразования форматов. Собственно *тип* представляет собой совокупность атрибутов, которые могут быть представлены в формате XML-Schema, и процедур для манипулирования данными. В портале поисковая операция всегда производится над данными, описанными в схеме одним из типов, в связи с этим отметим два момента:

- описание типа в репозитории содержит описания полей (включая описания типов данных, комментарии и выводимые имена на разных языках), по которым производится атрибутный поиск, а также формируется результат;

- описание типа в репозитории содержит описание процедур, которые используются для получения данных по поисковым критериям.

Описания полей необходимы для реализации формы поиска, в частности сбора и проверки поисковых значений, а также для форматирования и вывода результатов. Мультиязыковые описания используются для оформления элементов управления и ввода значений полей.

В портале реализуется как атрибутный, так и полнотекстовый поиск, поэтому каждый тип содержит описание, как минимум, двух процедур – для полнотекстового и для атрибутного поиска. В описание процедуры полнотекстового поиска входит всего один строковый параметр – текст, по которому будет выполнен полнотекстовый запрос. Процедура, используемая для атрибутивного поиска, имеет параметры, которые соответствуют набору полей, участвующих в поиске.

Приведем пример описания типа для объекта «Персона».

```
<Type name="Person" entity="Person">
        <Field name="Id" entity="Id"/>
        <Field name="Name" entity="Name"/>
        <Field name="Surname" entity="Surname"/>
        <Field name="BornDate" entity="BornDate"/>
        <Field name="Address" entity="Address"/>
        <Script name="Get">
                <Parameter name="Id" entity="Id" optional="true"/>
        <Condition name="Name" entity="Name">
                        <Operation name="IN"/>
                        <Operation name="equal"/>
                </Condition>
                <Condition name="BornDate">
                        <Operation name="equal"/>
                        <Operation name="greater"/>
                </Condition>
                ...
        </Script>

        <Script name="FullTextSearch">
        <Parameter name="QueryString" entity=" QueryString " optional="false"/>
        </Script>
        …
</Type>

…

<Entity name="Person">
        <DisplayName lang="ru">Персона</DisplayName>
<DisplayName lang="en">Person</DisplayName>
        <DisplayName lang="fr">Personne</DisplayName>
</Entity>
```

Приведенный пример описывает тип с набором полей и двумя процедурами. Процедура «Get» в данном описании предназначена для атрибутного поиска, «FullTextSearch» – полнотекстового.

Как видно, реализация типов содержит только описательную часть – процедуры только декларируются. Для получения данных из источника этого недостаточно – нужно соединиться с источником данных и передать поисковый запрос, оформив его в терминах источника. Для этих целей виртуальная схема портала имеет секцию реализации типов, которая описывает необходимые процедуры. Эти процедуры выполняются процессором запросов на стадии взаимодействия с источником данных. Подробно система реализации типов описана в [1] .

3.2. Шаблон поисковых форм

На основе шаблона поисковых форм строятся интерфейсы для поиска данных различных источников. Шаблон поисковых форм содержит компоненты, единые для всех страниц портала, такие как «шапка» и «подвал», меню и т.п. Содержанием конкретной поисковой формы является набор элементов управления для ввода значений и критериев поиска. Шаблон поисковых форм может реализовывать функциональность web-визарда в случаях, когда по эргономическим соображениям элементы управления нужно разделить на несколько страниц. Также определяется состав и элементы управления для задания критериев поиска (имеются в виду логические операции над всеми введенными пользователем значениями полей, но не сами значения). Шаблон поисковых форм не определяет состав и типы элементов управления для самих полей.

Каждому полю формы соответствует элемент управления, который позволяет задать поисковое значение для соответствующего атрибута типа, а также условие для проверки введенного значения. Элемент управления может быть любой. Единственным условием применения является то, что задаваемое значение должно преобразовываться в строковый вид. Тривиальным элементом управления для поиска является поле ввода. Для построения элементов управления, их инициализации и получения значений шаблон поисковых форм использует *процессор форм*. При помощи коллекции управляющих объектов строится содержательная часть поискового интерфейса: создаются заголовки, элементы управления и компоненты для проверки значений (валидаторы), после чего готовая поисковая форма выдается пользователю. Данные, введенные пользователем, должны быть обработаны и переданы *процессору запросов*.

После выполнения поискового запроса результат возвращается в виде объекта стандартного типа DataSet и выводится *шаблоном вывода результатов* в виде таблицы, которая

содержит как некоторые атрибуты найденных ресурсов, так и ссылки на них.

3.3. Процессор форм и описание форм

Как было сказано выше, первая задача, которую решает процессор форм – построение содержательной части поисковой формы – набора элементов управления для ввода и проверки поисковых значений. Метаданные для построения этого набора содержаться в специальном разделе схемы. Фактически эти метаданные являются описанием форм, т.е. элементов управления и их расположения.

Второй задачей процессора форм является построение поискового запроса и передача его процессору запросов для исполнения. Запрос в терминах виртуальной схемы включает в себя тип, процедуру, набор условий и параметров. Поисковый запрос выражается в терминах виртуальной схемы и состоит из набора условий, которые объединяются в логическое выражение. Каждое условие представляет собой тройку «поле» - «операция» - «значение». Поисковое выражение формируется кодом формы. Соответственно, для построения поискового запроса метаданные схемы содержат имя типа и процедуры. Условия формируются из значений, введенных пользователем в элементы управления.

Процессор форм в порядке инициализации компилирует сборку для каждой формы на основе ее метаописания. Каждая сборка содержит общий класс для формы и по одному классу управления для каждого поля формы. Каждый такой класс наследуется от базового (PSControlManager), в котором определены следующие функции:

- Init – для инициализации;
- CreateControl – используется шаблоном поисковых форм для создания элемента управления;
- GetValue – возвращает условие (поле, значение и операцию);
- DataBinding – используется для отображения данных в сложных элементах управления.

Такой класс выступает в качестве «обвязки» каждого элемента управления и устанавливает интерфейсную связь между шаблоном поиска и процессором форм при генерации формы и формировании поискового запроса. Класс «PSControlManager» реализует при помощи этих функций работу с элементом управления «TextBox». В процессор форм встроены также еще несколько классов для работы с другими элементами управления, такими как «ListBox», «DropDownList» и т.д.

Для задания собственной функциональности взаимодействия с элементом управления посредством методов «CreateControl», «GetValue» и «DataBinding» можно переопределить методы базового класса, поместив код метода в соответствующую секцию метаописания поля формы. По умолчанию используемым классом управления является класс «PSControlManager», но можно использовать и другой встроенный базовый класс, указав его псевдоним в теге описания поля атрибутом «controlpattern».

Для отображения элементов управления шаблон поисковых форм получает коллекцию объектов управления полями поиска. Каждый такой объект реализуется процессором форм. Он содержит методы, позволяющие:

- получить заголовок для отображения поля;
- создать элемент управления для поля;
- проинициализировать созданный элемент управления, в т.ч. с использованием запросов к данным;
- получить критерий поиска по полю из значения элемента управления и условия сравнения (>, <, = и т.п.).

При построении коллекции элементов управления для ввода значений условий шаблон поисковых форм для каждого поля формы выполняет следующие действия:

- создает объект вышеописанного класса;
- инициализирует его функцией «Init»;
- создает элемент управления при помощи «CreateControl»;
- связывает с ним данные при помощи «DataBind».

После этих действий форма отображается пользователю, он вводит значения. Для обработки значений и генерации набора условий для каждого элемента управления вызывается функция «GetValue», которая возвращает объект-условие. После этого сгенерированный таким образом набор условий, имя типа и процедуры образуют запрос в терминах виртуальной схемы, который передается процессору запросов для выборки данных.

Ниже приведен пример описания формы, который показывает как простое описание поля формы, так и описание с переопределением методов.

В данном примере форма предназначена для атрибутного поиска по типу «Персона». В описании формы перечислены условия процедуры поиска, которые используются формой, а также набор полей для вывода результатов. Предполагается, что данный тип связан с типом «Профессия» («Occupation») (в данном примере не указан), который является справочником профессий. Поле формы, соответствующее условию принадлежности к профессии («Occupation»), оформлено как выпадающий список, содержащий названия профессий. Для этого в описании формы для условия «Occupation» переопределены методы «CreateControl», «GetValue» и «DataBinding».

На Рис.1 приведен пример поисковой формы портала www.ras.ru.

После инициализации и построения сборок для форм процессор форм используется шаблоном поисковых форм для создания интерфейса для задания поисковых критериев и инициализации элементов управления.

3.4. Шаблон вывода результатов

Шаблон вывода результатов предназначен для формирования web-представления результирующего набора данных. После выполнения запроса процессором форм результирующий набор данных представляется объектом DataSet. Для форматирования данных, содержащихся в наборе, шаблон поисковых форм использует метаописания поисковой формы, содержащиеся в специальной секции виртуальной схемы «output». На основе них строится результирующая таблица, в которой каждое поле форматируется в соответствии с заданными правилами, а вся таблица разбивается на страницы.

Пример поиска по одному из внешних информационных источников портала РАН – электронному архиву академика В.И.Вернадского – приведен на Рис. 2.

3.5. Полнотекстовый поиск

Совершенно аналогично атрибутному поиску, рассмотренному выше, может быть организован и полнотекстовый поиск. Для этого достаточно, чтобы набор методов, поддерживаемых типами, содержал специальный метод для полнотекстового поиска, например, так, как с типом «Персона». Этот метод имеет один параметр – строку с выражением для полнотекстового поиска. Например, полнотекстовой запрос может касаться фамилии, имени, отчества и местожительства.

```
<Form name="PersonForm" entity="_Person" type="Person">
<Script name="GetPerson">
<Condition name="LastName" operation="in" />
<Condition name="FirstName" operation="in" />
<Condition name="PatronymicName" operation="in" />
<Condition name="FullName" operation="in" />
<Condition name="BornPlace" operation="in" />
<Condition name="Occupation" operation="equal" >
        <CreateControl controltype="System.Web.UI.WebControls.DropDownList, System.Web,
        Version=2.0.0.0, Culture=neutral, PublicKeyToken=b03f5f7f11d50a3a"/>
        <GetValue valuefieldprop="Text">
                _result = new PSCondition(FieldInfo.Name, "=",
                (DropDownList)ControlObj.SelectedItem.Value);
        </GetValue>
        <DataBinding sourcetype="link" sourcename="Occupations"
        textcolumn="Name" datacolumn="Id"/>
</Condition>
<Field name="id" href="true" />
<Field name="LastName" />
<Field name="FirstName" />
<Field name="PatronymicName" />
<Field name="FullName" />
<Field name="BornDate" />
<Field name="BornPlace" />
<Field name="INN" />
</Script>
</Form>
```

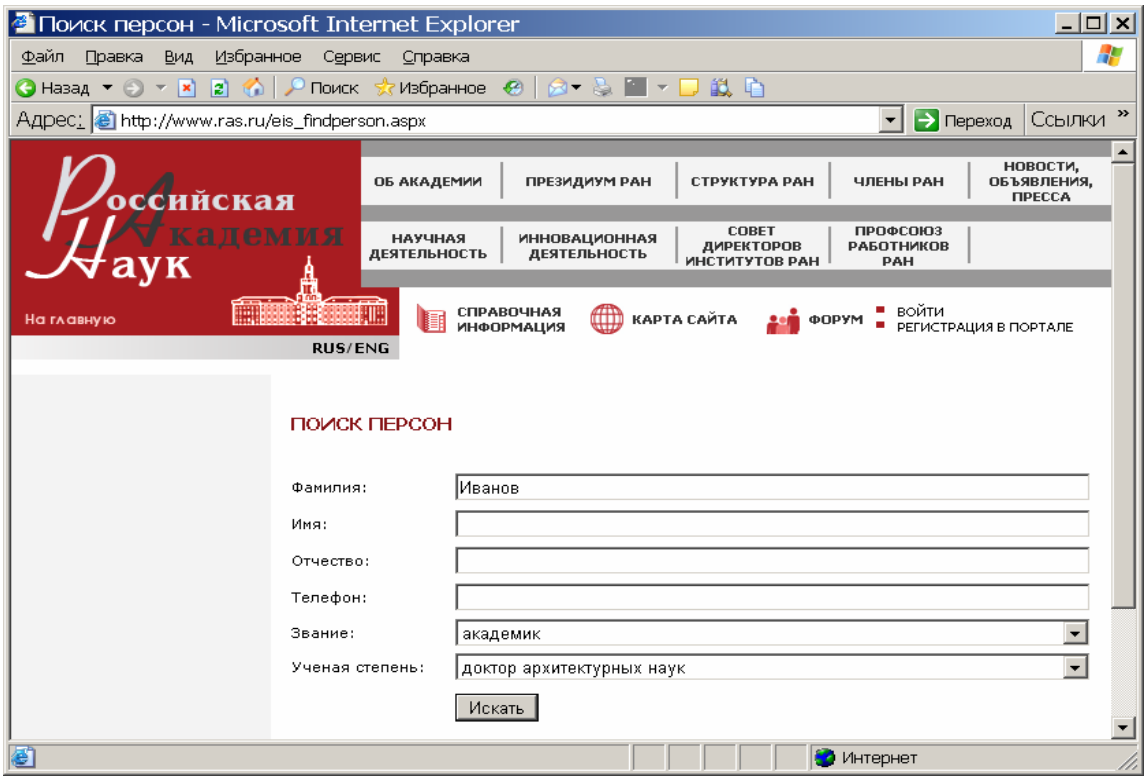

Рис.1 Пример поисковой формы

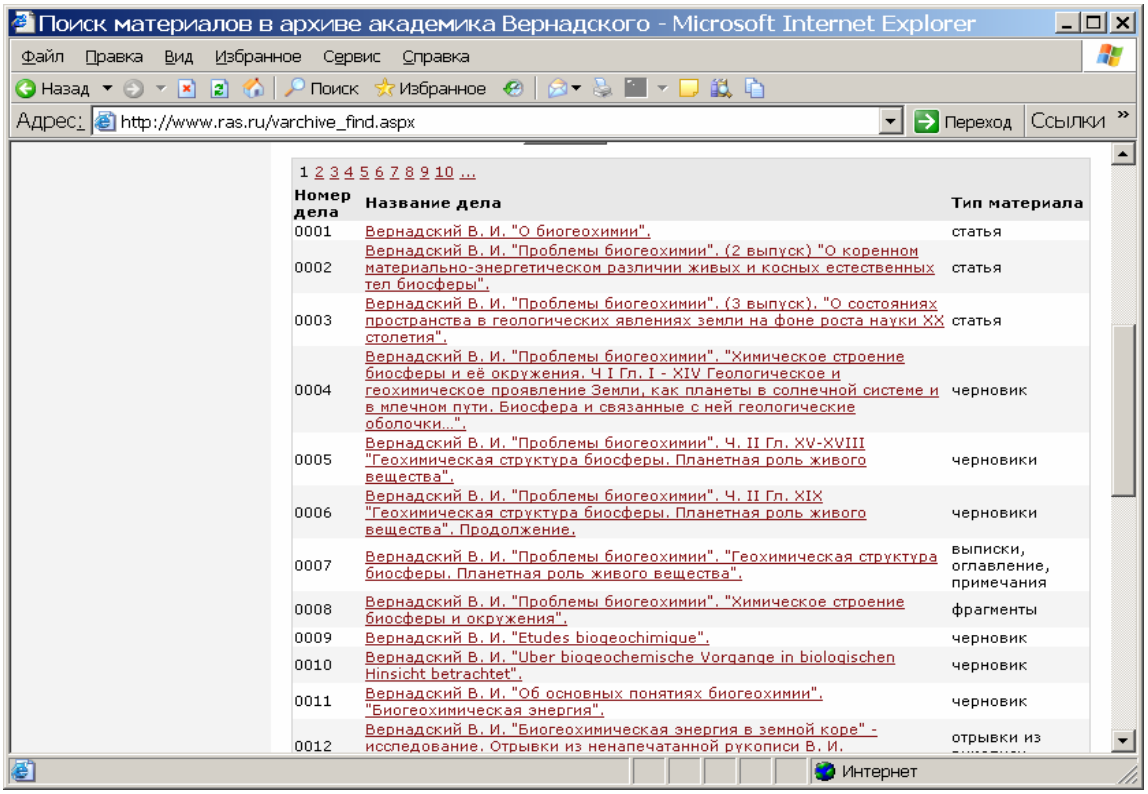

Рис. 2. Результаты атрибутного поиска по архиву В.И.Вернадского

```
    …
<Script name="FullTextSearch">
<Body><
        ![CDATA[
        SELECT %fields% FROM PERSON INNER JOIN CONTAINSTABLE(PERSON,
         *, @QueryString) AS Ranks ON Ranks.[Key] = PERSON.[OWNER_ID]
        ORDER BY Ranks.[Rank] DESC
        ]]>
</Body>
</Script>
…

…
<Script name="FullTextSearch">
<Body><
        ![CDATA[
        <?xml version="1.0" encoding="utf-8"?>
        <soap:Envelope
                xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
                xmlns:xsd=http://www.w3.org/2001/XMLSchema
                xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
                <soap:Body>
                        < FullTextSearch xmlns="http://tempuri.org/">
                                %parameters%
                        </ FullTextSearch >
                </soap:Body>
        </soap:Envelope>
        ]]>
</Body>
</Script>
```

Для каждого источника реализация процедуры полнотекстового поиска будет выглядеть по-своему. Выше приведены два примера описания процедуры полнотекстового поиска для реляционной базы данных SQL-сервера и web-сервиса.

Заключение

В настоящей статье рассмотрены основные принципы и архитектура одного из важнейших компонентов Информационного web-портала – программной системы, обеспечивающей официальное представление Российской академии наук в сети Интернет. Реализация описанной службы поиска в составе действующего академического портала обеспечила расширение его функциональности в части наиболее значимых потребностей пользовательской аудитории и позволила расширить спектр ресурсов, имеющих перспективу интеграции в Единое научное информационное пространство РАН.

Литература

1. Босов А.В., Чавтараев Р.Б. Технология доступа к данным в Информационном web-портале // Информационные технологии и вычислительные системы. №1– М.: 2007.. С.35-48.
2. Босов А.В., Иванов А.В. О реализации системы управления содержанием информационного Web-портала // Информационные технологии и вычислительные системы. №4. – М. С.85-103.
3. Bergman M.K. The Deep Web: Surfacing Hidden Value // http://www.press.umich.edu/jep/07-01/bergman.html.
4. European Research Gateways Online // http://www.cordis.lu/ergo.
5. Laitinen, Sauli; Sutela Pirjo & Tirronen, Kerttu, Development of Current Research Information Systems in Finland // Proceeding of CRIS-2000.
6. Некрестьянов И.С., Пантелеева Н. Системы текстового поиска для Веб // Программирование, № 4. 2002. С. 33-57.
7. Гаскаров Д.В. Интеллектуальные информационные системы. М.: Изд-во «Высшая школа», 2003.
8. Гринберг И., Гарбер Ли. Разработка новых технологий информационного поиска. Открытые системы. №№9-10. 1999.
9. Влад Жигалов. Как нам обустроить поиск в Сети? //Журнал «Открытые системы», №2, 2000. Издательство «Открытые Системы» http://www.osp.ru/os/2000/12/053.htm.
10. XML Schema Part 2: DataTypes, Primitive DataTypes // http://www.w3.org/TR/xmlschema-2/.

Босов Алексей Вячеславович. Заведующий сектором Института проблем информатики РАН. Окончил Московский государственный авиационный институт в 1993 году. Кандидат физико-математических наук, доцент, автор более 50 научных работ. Специалист в области прикладной математики и информатики. Область научных интересов – информационные технологии, аналитические и обучающие системы, Интернет-технологии и порталы.

Чавтараев Рустам Баширович. Старший научный сотрудник Института проблем информатики РАН. Окончил Московский институт радиоэлектроники и автоматики в 1995 году. Автор 12 научных работ. Специалист в области автоматизированных систем обработки информации. Область научных интересов – информационные технологии, Интернет-технологии и порталы.