Реализация службы поиска в информационных источниках портала РАН

© А.В. Босов, Р.Б.Чавтараев

Институт проблем информатики Российской академии наук
AVBosov@ipiran.ru, RCh@ipiran.ru

Аннотация

Обсуждается решение, реализованное в составе действующего Информационного web-портала РАН с целью расширения его функциональности услугами гибкой поисковой системы. Рассмотрена архитектура и программная реализация решения.

1 Введение

Поиск информации – одна из первых задач, которую пытается решать пользователь, подключившись к Интернет. Другая заинтересованная сторона предоставляет информацию, стараясь обеспечить ее легкую доступность. Таким образом, 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. Далее данных стало больше, они перестали храниться в статике, а размещались, как правило, в базах данных, работать с которыми поисковые машины уже не могли. Можно считать, что так контент Интернет поделился на «поверхностный» (Surface Web) и «глубинный» (Deep Web). Любопытно, что оценки, характеризующие объемы этих частей (например, в [1] 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 [2,3], 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, поэтому охарактеризовать тематику можно было бы как «научная информация», но, очевидно, что на самом деле такое «ограничение» мало что проясняет. Во-вторых, предъявляются очень суровые 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 облегчить жизнь разработчика (администратора или даже модератора), предложив более гибкие средства интеграции ресурсов для целей поиска.

В данной работе рассмотрено решение, реализованное в действующем Информационном web-портале РАН www.ras.ru [4,5], с целью усовершен-

Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2006, Суздаль, Россия, 2006.

ствования 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.

2 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 простоту, 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 поразному 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 при-

вести 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.

3 Определение используемых технологий

Итак, существует портальное решение www.ras.ru, которое можно использовать для реализации выбранной стратегии. Элементы программной инфраструктуры портала, привлекаемой для нужд поиска, рассмотрим далее (собственно все, что создавалось в портале на эти цели и ориентировалось). Задача портала состоит в интеграции разнородной информации и обеспечении на разных уровнях интероперабельности между системами в целях выполнения информационного поиска в распределенной среде web-портала.

В начале определим требования к а) формату данных, используемых для обмена информацией; б) технологии обмена данными.

Предъявляемые в рамках портала требования к формату обмена данными таковы:

- *универсальные выразительные возможности*. Формат обмена должен позволять представлять любой вид данных, поскольку нельзя учесть интересы всех потенциальных систем;
- *синтаксическая интероперабельность*. Формат обмена должен позволять легко создавать или получать синтаксические анализаторы, прикладные интерфейсы, необходимые для манипулирования данными. Приложения должны иметь возможность читать данные и получать их представление, с которыми они смогут работать;
- *семантическая интероперабельность*. Это наиболее важное и сложное требование к формату обмена состоит в том, что данные должны быть понятными. Это связано с установлением соответствия между терминами, используемыми в данных, что требует анализа содержимого.

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.

Удовлетворять перечисленные требования предлагается 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. Добавляет семантическую составляющую к XML.

Далее, 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 службой поиска осуществляется передача поисковых запросов и получение данных федеративных информационных источников. В случае электронных библиотек основными компонентами промежуточного слоя являются посредники/медиаторы (mediator) и адаптеры/оболочки (wrapper). Адаптер используется для организации доступа к источнику информации, поддерживая интерфейсы доступа к данным (для каждого источника – свой адаптер, владеющей информацией о информационном источник и способах взаимодей-

ствия с ним). 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. Посредник осуществляет интеграцию данных различных источников, предоставляемых адаптерами, и может взаимодействовать как с адаптерами, так и с другими посредниками, что позволяет строить сложные сети посредников, взаимодействующих между собой.

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 включать в себя массу условий. При этом, несмотря на то, что каждая система хранения данных оперирует запросами для выборки данных, 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.

Из этого следует, что необходим механизм, который преобразует данные формы в запрос и передает этот запрос на выполнение информационному источнику. Как и в электронных библиотеках, в нашем случае задачу взаимодействия с информационным источником решает адаптер. Единственным существенным отличием портальных адаптеров является конкретизация требований к предоставляемым ими интерфейсам за счет сужения решаемой задачи.

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 портала отвечает за реализацию набора типов (подсхемы), которые поддерживает информационный источник. Задача адаптера – 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.

4 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 каком-либо цифровом виде, обеспечена в существующей реализации портала [10] 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.

Другой 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.

5 Архитектура и реализация поисковой службы

Служба поиска 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, формирования поискового запроса и его передача 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 используется виртуальная схема, 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 три части: *репозиторий типов, систему реализации типов* и *описания форм*.

Функционально служба поиска 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;
- *процессором запросов.* Этот компонент осуществляет взаимодействие с адаптерами информационных источников с целью выполнения поискового запроса и формирования общего результата поисковой операции. 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.

5.1 Репозиторий типов

Для описания данных, по которым могут производиться поисковые операции, используется *репозиторий типов.* Описание типов представляет собой структуру данных, общую для всех информационных источников, которые участвуют в поиске. Репозиторий типов является интерфейсной частью виртуальной схемы портала, которая отображает общую структуру данных, абстрагированную от деталей реализации доступа и преобразования форматов. Репозиторий типов основывается на трех основных элементах – тип, связь, мультиязычное описание.

Тип представляет собой совокупность атрибутов, которые могут быть представлены в формате XML-Schema. Каждому типу соответствует мультиязычное описание, которое используется для представления пользователю. Так, например, типу, который описывает личность соответствует описание, содержащее выводимое имя на трех языках – «персона» (русский), «person» (английский), «personne» (французский) и т.п. Каждый тип однозначно идентифицируется свойством «name» и, соответственно, существование двух типов с одинаковым свойством «name» невозможно.

Репозиторий типов реализует множественное наследование. Каждый тип может иметь один или несколько базовых типов: наследование означает включение в состав атрибутов данного типа всех атрибутов базовых типов. Если атрибуты базовых типов пересекаются по названиям, результирующий тип будет содержать один унаследованный атрибут с данным именем. Ниже приведен пример описания двух типов с использованием наследования.

```
   <Type name="named">
   <Field name="Id" entity="id"/>
   <Field name="Name" entity="name"/>
   <Field name="Description"
entity="description"/>
   <Presentation type="list"/>
   </Type>
   <Type name="file" entity="file">
   <Parent name="named"/>
   <Field name="FSize" entity="size"/>
   <Field name="UploadTime"
entity="createtime"/>
   <Field name="MaterialDate"
entity="date"/>
   <Field name="FType"
entity="filetype"/>
   <Field name="Copyright"
entity="copyright"/>
   <Field name="Source"
entity="source"/>
   </Type>
```

В приведенном фрагменте тип «file» будет содержать атрибуты Id, Name, Description родительского типа и собственные FSize, UploadTime, MaterialDate, FType, Copyright и Source.

Мультиязычные описания можно поставить в соответствие любому типу и любому атрибуту типа. Несколько типов или атрибутов могут быть соотнесены к одному описанию. Описание определяется атрибутом «Entity» в представлении типа или атрибута. Ниже приведен пример описания для типа «file» и некоторых его атрибутов.

```
  <Entity   name="file">
  <DisplayName
lang="ru">Файл</DisplayName>
  <DisplayName
lang="en">File</DisplayName>
  </Entity>
  <Entity   name="filetype">
  <DisplayName lang="ru">Тип
файла</DisplayName>
  <DisplayName lang="en">File
type</DisplayName>
  </Entity>
  <Entity   name="copyright">
  <DisplayName lang="ru">Авторское
право</DisplayName>
  <DisplayName
lang="en">Copyright</DisplayName>
  </Entity>
```

Репозиторий типов поддерживает определение связей между типами. Связь определяется как набор типов (два или более), участвующих в соединении. Таким образом, можно задать стандартные виды связей – «один к одному», «один ко многим», «многие ко многим». Ниже приведен пример описания связи для типов «file» и «Category», данная связь имеет вид «многие ко многим».

```
  <Link name="file_category">
  <Type name="file"
occurs="multiple"/>
  <Type name="category"
occurs="multiple"/>
  </Link>
```

5.2 Система реализации типов

Система реализации типов предназначена для поддержки взаимодействия с информационными источниками. Как уже говорилось, в контексте поиска это взаимодействие состоит в формировании поисковых запросов к источникам, получении данных и форматировании результатов, и выполняется адаптером источника. В виртуальной схеме портала для каждого источника отводится отдельная секция, которая содержит информацию двух видов. Во-первых, в секции источника содержатся все необходимые данные для создания среды взаимодействия с ним. Например, для взаимодействия с реляционной базой данных в эту секцию включается информация для создания сессии (настройки источника ODBC), для web-сервиса – WSDL-описание или URL, по которому его можно получить. Во-вторых, в секции источника содержится описание отображения части типов и связей из репозитория типов на данный источник (метаданные адаптера). В этом отображении присутствуют только те типы и связи, которые реализуются данным источником (его подсхема). При помощи отображения типов реализуются операции над экземплярами типов (под экземплярами в данном контексте понимается структура данных, соответствующая описанию типа).

Взаимодействие с информационным источником строится следующим образом. Процессор запросов генерирует текстовую команду, которая соответствует операции над экземплярами типа – в контексте поиска это операция выборки данных по условиям. Текстовая команда строится с учетом набора условий (conditions), набора возвращаемых полей, языка и т.п. Команда может быть как статической, так и динамической – генерироваться каждый раз при выполнении запроса. Службой поиска не определяются какие-либо ограничения на формат и содержимое команд, за исключением набора зарезервированных слов. После того, как команда сгенерирована, она передается адаптеру, который выполняет команду и возвращает результирующий набор данных.

Система реализации типов позволяет определять статические и динамические команды, а также модули исполнения с использованием языков программирования, поддерживающих CLR [11]. Ниже приводится пример описания реализации для типа «file» в источнике данных.

```
  <Type name="file">
  <Script name="Search">
  <![CDATA[
  SELECT %fields% FROM FSFiles
  <%if (Conditions != null &&
Conditions.Count > 0) {
  if (Conditions["category"] !=
null)%>
  INNER JOIN FSBelongs ON FSFiles.Id
= FSBelongs.FileId
  <%}%>
  WHERE %conditions%
  ]]>
  </Script>
  </Type>
```

В приведенном примере реализации типа «file» с использованием динамической команды применен язык программирования C#, фрагменты программного кода заключены в маркеры «<%» и «%>». Команда формируется из строк текста, которые не являются кодом. Код предназначен для конструирования команды из фрагментов текста, размещенным между строками кода. В приведенном выше примере фрагмент команды «INNER JOIN FSBelongs ON FSFiles.Id = FSBelongs.FileId» будет включен в ре-

зультирующий текст только при выполнении условия оператора «if (Conditions["category"] != null)». Таким образом, соединение с таблицей FSBelongs в команде генерируется только при наличии условия поиска в этой таблице.

В коде могут быть использованы переменные «Conditions», «Fields», «Lang» и «User». Переменная «Conditions» представляет собой коллекцию условий. Каждое условие выражается тройкой «полезначение-операция» и в контексте поисковых задач выражает критерии отбора данных. В переменной «Fields» содержится коллекция имен терминов, которые должны быть возвращены. Параметр «Lang» определяет язык, который используется в процессе поиска, с его помощью реализуются мультиязычные свойства портала. Параметр «User» используется для интеграции с подсистемой безопасности портала. Он представляет собой объект, реализующий интерфейс IPrincipal, через который доступна информация о вызывающем пользователе.

Для использования в коде текста формируемой команды можно использовать предопределенную переменную «_result» строкового типа, которая содержит сформированную часть команды.

Существует также еще одно средство для конструирования команд – это шаблоны подстановки %fields% и %conditions%. При генерации текста команды эти идентификаторы будут заменены на строки, которые формируются, соответственно, из коллекций «fields» и «conditions» с применением шаблона подстановки. Шаблон подстановки представляет собой строку, которая отражает способ преобразования массива элементов в строковый вид с использованием разделителей. Например, если коллекция «fields» содержит элементы «id» и «name», то, применив шаблон подстановки «{0[]:*:, }», получим строку «id, name». Шаблоны подстановки могут быть определены для источника данных, для реализации типа и для конкретной команды.

После того, как команда сгенерирована, она должна быть передана адаптеру для выполнения и получения результирующего набора данных. Для этих целей служит секция источника, которая называется «Exec». В этой секции определяется код для манипулирования с данными, а именно код, который обрабатывает команду, созданную на предыдущем этапе, и формирует результирующий набор данных. В качестве параметров передается сама команда и описанные выше параметры «fields», «conditions», «lang», «user».

Для некоторых видов информационных источников реализованы встроенные функции обработки команд. К ним относятся источники данных ODBC, OLEDB, реализации для SQL Server, Oracle, XML, а также сервер интеграции и доступа, входящий в состав существующей реализации портала [10].

Ниже приведен пример определения обработки команды в секции «Exec» для базы данных SQL Server.

```
  <Exec>
  _result =
Utils.ExecSqlSource("workstation
id=localhost;packet
size=4096;integrated
security=SSPI;data
source=localhost;persist security
info=False;initial
catalog=MArchive2", Script,
Conditions, Lang, User);
  </Exec>
```

Таким образом, предложенная архитектура позволяет реализовывать функции поиска практически к любым источникам данных, даже не поддерживающим какой-либо язык запросов.

5.3 Процессор запросов

Реализацию цепочки выполнения запроса производит *процессор запросов*. Описанная выше функциональность реализуется с помощью динамического построения и использования сборок, которая использует технологии генерации кодов CodeDom и позднего связывания Reflection.

Как упоминалось выше, метаданные, используемые службой поиска, содержат фрагменты кода, определяющие функциональность при взаимодействии с информационными источниками. Эти фрагменты сосредоточены в двух местах для каждого источника – в описании реализации типа в системе реализации типов и в процедуре исполнения команды «Exec». Для взаимодействия с информационным источником процессор запросов предварительно компилирует этот код в исполняемый модуль – сборку (assembly) в терминах .NET, а при выполнении запроса вызывает методы объектов сборки по заданным правилам.

Рассмотрим принципы генерации сборок. Для каждого источника данных генерируется своя сборка, содержащая определения классов в соответствии с типами, реализуемыми источником, и предопределенный класс Global. Класс Global реализует единственный метод – «Execute» – который содержит код, содержащийся в секции «Exec» источника. Каждый класс, соответствующий реализуемому типу, содержит методы формирования команд. Каждой команде соответствует свой метод. Метод формируется из фрагментов кода и фрагментов команды, описанных в секции «script» для реализуемого типа. Возвращаемое значение такого метода – строка, которая является сгенерированной командой. Ниже приведен пример кода метода, сгенерированного по описанию типа «file».

```
  public class file {
  public virtual string
Search(PortalSearch.PSFieldsEnumerati
on Fields, PortalSearch.PSConditions
Conditions, string Lang, string User)
{
```

```
   System.Text.StringBuilder _result =
new System.Text.StringBuilder();
   _result.Append("\r\n\t\t\t\t\t\tSEL
ECT %fields% FROM
FSFiles\r\n\t\t\t\t\t\t");

   if (Conditions != null &&
Conditions.Count > 0)
   { if (Conditions["file_category"]
!= null)
   _result.Append("\r\n\t\t\t\t\t\t\t\
t\tINNER JOIN FSBelongs ON FSFiles.Id
=
FSBelongs.FileId\r\n\t\t\t\t\t\tWHERE
%conditions%\r\n\t\t\t\t\t\t");
   } return _result.ToString(); } }
```

Существует также возможность использовать другие сборки, в которых могут быть реализованы функции доступа к данным, формирования команд и др. Для этих целей используется секция источника «assembly», которая содержит путь к используемой сборке, ее полное имя и другие необходимые данные. После построения сборок становится возможным обслуживание запросов доступа к данным источников. Общая последовательность поисковой операции представлена на рисунке.

Процессор запросов в качестве исходных параметров принимает имя типа, по которому производится поиск, набор условий поиска, условия сортировки, набор возвращаемых полей, язык, на котором должны быть выданы результаты, и учетные данные вызывающего пользователя. Далее определяется список информационных источников, которые реализуют данный тип, с помощью системы реализации типов виртуальной схемы. Для каждого типа выполняется операция построения и исполнения команды, результатом которой является набор данных DataSet. Результирующие наборы данных по всем участвующим в запросе информационным источникам объединяются в один и производится сортировка, после чего этот результирующий набор данных выдается для форматирования и вывода пользователю.

Рассмотрим подробнее процесс построения и выполнения команды для источника. Как упоминалось выше, каждому типу, реализуемому информационным источником, соответствует сгенерированный класс. Процессор запросов в первую очередь создает экземпляр класса, соответствующий типу, и вызывает для построения команды метод, который соответствует операции поиска. Этот метод возвращает текст сконструированной команды. Текст команды на данном этапе может содержать идентификаторы подстановки (%fields% и %conditions%). Используя шаблоны подстановки для коллекций «fields» и «conditions», процессор запросов строит их строковые отображения и осуществляет подмену идентификаторов этими строковыми отображениями. На этом этапе формирование команды завершено. После формирования команды процессор запросов, используя экземпляр класса Global сборки адаптера источника, вызывает метод «Execute», в который помимо прочих параметров передается сгенерированная команда. Метод «Execute» обрабатывает команду и возвращает результат в виде набора данных.

5.4 Шаблон поисковых форм

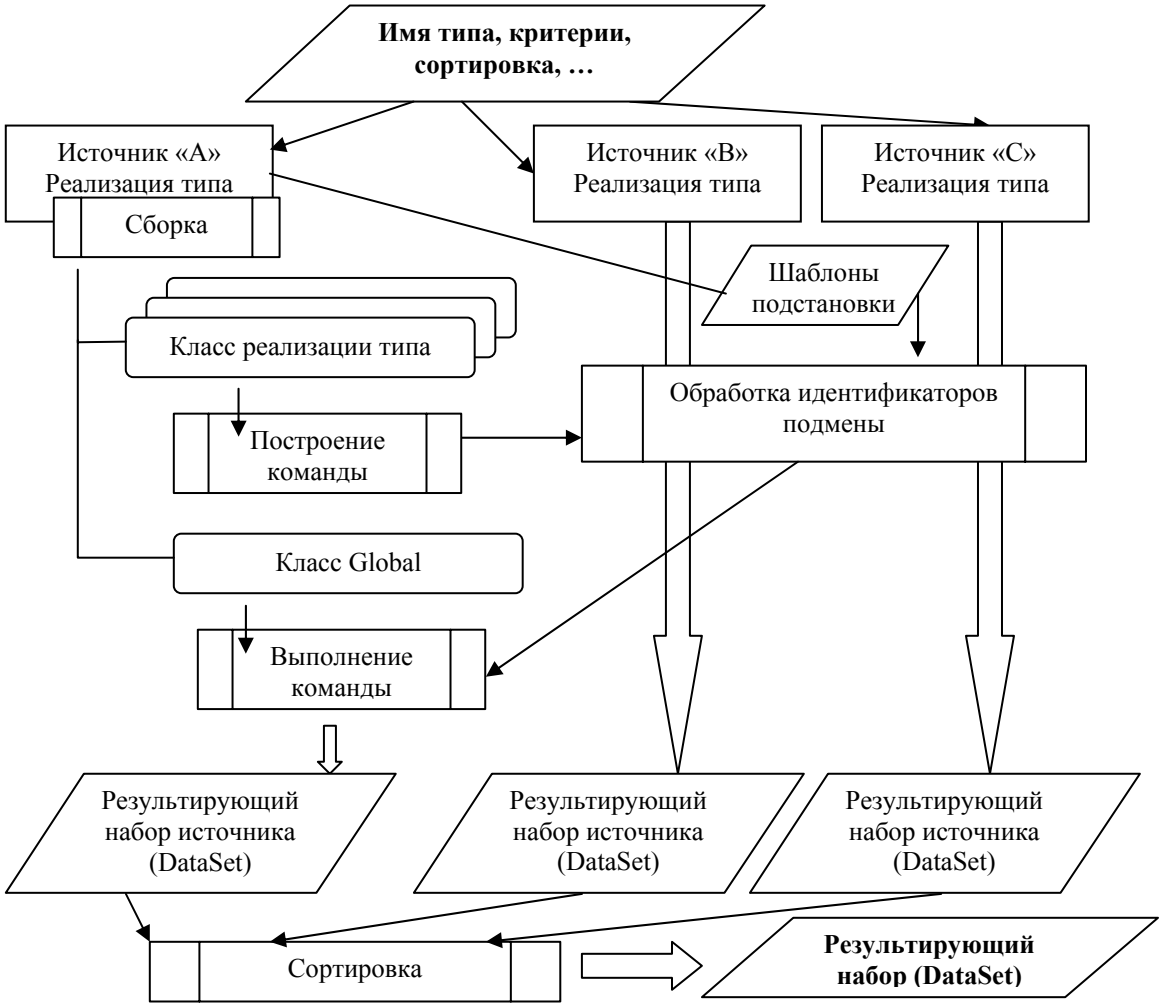
На основе шаблона поисковых форм строятся интерфейсы для поиска данных различных источников. Шаблон поисковых форм содержит компоненты, единые для всех страниц портала, такие как «шапка» и «подвал», меню и т.п. Содержанием конкретной поисковой формы является набор элементов управления для ввода значений 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.

5.5 Процессор форм и описание форм

Задача процессора форм состоит в построении коллекции элементов управления для ввода значений атрибутов, их инициализации, проверке введенных значений и преобразовании их в унифицированный формат. 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.).

Подобно процессору запросов процессор форм в порядке инициализации компилирует сборку для каждой формы на основе ее метаописания. Классы этой сборки реализуют вышеперечисленные функции. Каждая сборка содержит общий класс для формы и по одному классу управления для каждого поля формы. Все классы управления по умолчанию наследуются от встроенного класса «PSControlManager», у которого определены три виртуальные функции – «CreateControl», «GetValue» и «DataBinding». 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».

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

```
  <Form name="ArchiveSearch"
type="act" >
  <DisplayName lang="ru">Поиск по
архивам</DisplayName>
  <DisplayName lang="ru">Archive
search</DisplayName>
  <Field name="Name"/>
  <Field name="Inventory">
  <CreateControl
controltype="System.Web.UI.WebControl
s.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="Inventory"
textcolumn="Name" datacolumn="Id"/>
   </Field>
   <Source name="MArchive"/>
   <Source name="VArchive"/>
   </Form>
```

В данном примере поле с именем «Name» описано в простом виде. В этом случае для реализации класса управления будет использован предопределенный класс «PSControlManager» и для него будет сгенерирован элемент управления «TextBox». Поле «Inventory» описано таким образом, что будет сгенерирован класс-наследник, который оперирует с элементом управления «DropDownList» и переопределяет метод «GetValue» базового класса. Также в приведенном описании формы видны источники, по которым будет производиться поиск. Если источники не заданы, поиск производится по тем источникам, где присутствует реализация класса «Act».

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

5.6 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 секции «output». На основе них строится результирующая таблица, в которой каждое поле форматируется в соответствии с заданными правилами, а вся таблица разбивается на страницы.

5.7 Средства администрирования

Вся необходимая информация, используемая службой поиска, физически размещается в файлах XML. Несмотря на то, что эти файлы могут быть сформированы и отредактированы вручную при помощи любого текстового редактора, средства администрирования службы поиска существенно облегчают управление службой, позволяя пользователю не вдаваться в детали реализации виртуальной схемы.

Средства администрирования представляют набор пользовательских интерфейсов, которые позволяют:

- управлять репозиторием типов;
- выполнять подключение информационных источников и управлять настройками доступа к ним (настраивать адаптеры);
- создавать и редактировать формы атрибутного поиска;
- создавать и проверять фрагменты кода для нужд службы поиска на разных языках платформы .NET.

Литература

[1] Deep Web – Bright Planet's white paper // http://www.completeplanet.com/tutorials/deepweb/.

[2] European Research Gateways Online // http://www.cordis.lu/ergo.

[3] Laitinen, Sauli; Sutela Pirjo & Tirronen, Kerttu, Development of Current Research Information Systems in Finland // Proceeding of CRIS-2000.

[4] Соколов И.А., Босов А.В., Бездушный А.Н. О Информационном Web-портале Российской академии наук // Системы и средства информатики. Вып. 13. – М.: Наука, 2003. С. 119-138.

[5] Россия. Веб-портал РАН обеспечивает эффективный доступ научного сообщества к актуальной информации. Информационный бюллетень Microsoft. Выпуск № 26. Ноябрь 2004 года.

[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] Босов А.В, Иванов А.В., Полухин А.Н., Чавтараев Р.Б. Управление сайтом информационного web-портала // Системы и средства информатики. Вып. 15. – М.: Наука, 2005. С.233-259.

[11] Рихтер Дж. Программирование на платформе Microsoft .NET Framework. Мастер класс. / Пер. с англ. – 3-е изд. – М.: Издательско-торговый дом «Русская редакция»; СПб.: Питер, 2005.

Search Service Realization for RAS Portal Information Sources

A.V.Bosov, R.B.Chavtaraev

The solution realized in RAS Informational web-portal (*www.ras.ru*) system for search purposes is consider. Aspects of architecture and program realization for search service solution are discussed.