Оптимизация обработки поисковых запросов на WEB-портале ВГБИЛ

Optimization of Search Query Processing on the Library's for Foreign Literature Web-portal

Оптимізація обробки запитів на WEB-порталі ВДБІЛ

Колосов К. А.
Всероссийская государственная библиотека иностранной литературы им. М.И.Рудомино, Москва, Россия

Kirill Kolosov
M. I. Rudomino All-Russian State Library for Foreign Literature, Moscow, Russia

Колосов К. А.
Всеросійська державна бібліотека іноземної літератури ім. М. І. Рудоміно, Москва, Росія

Рассматриваются вопросы использования технологии «сервер Z39.50 – шлюз HTTP-Z39.50» при разработке библиотечного портала ВГБИЛ. Описаны использованные технические решения, в том числе, сервер Z64 и разработанные провайдеры данных. Рассмотрен алгоритм последовательного поиска по локальным, региональным и расширенным информационным источникам с использованием шлюза HTTP-Z39.50.

Using the technology «Z39.50 – HTTP Gateway-Z39.50» for building the library web-portal of the All-Russian State Library for Foreign Literature is discussed. Technological solutions are described, among them, Z64-server and designed data providers. The algorithm of sequential search in local, regional and extended information resources using HTTP-Z39.50 gateway is examined.

Розглядаються питання використання технології «сервер Z39.50 – шлюз HTTP-Z39.50» при розробці бібліотечного порталу ВДБІЛ. Описані використані технічні рішення, в тому числі, сервер Z64 і розроблені провайдери даних. Розглянуто алгоритм послідовного пошуку за локальним, регіональним та росширеним інформаційними джерелами з використанням шлюзу HTTP-Z39.50.

Современные поисковые Интернет-порталы (Google, Yandex, Yahoo, Rambler и т.п.) предлагают пользователям простые и в тоже время эффективные варианты ввода поисковых запросов: простой поиск, при котором вводится произвольное поисковое выражение (фраза) и расширенный поиск, в режиме которого пользователь может указать поисковые атрибуты, уточнить критерии и ограничить список выдаваемых результатов. В результате поиска пользователь получает список гиперссылок и краткие аннотации к ним. Для пользователей это дает возможность быстро и без лишних усилий находить тот или иной ресурс.

Если сравнить этот подход с организацией поиска на традиционных библиотечных сайтах, то становится видно, что, несмотря на наличие большого числа информационных источников и электронных ресурсов, найти нужную информацию пользователю бывает непросто. Как правило, имеется возможность поиска по сайту, в основе которого лежит поиск по индексированному массиву документов, хранящихся на сайте в наиболее распространенных форматах (HTML, WORD, PDF). Кроме этого, как правило, существует отдельный раздел поиска по каталогам, в котором пользователь может переходить от одного каталога к другому, повторяя в каждом каталоге свой поисковый запрос. В более продвинутых интерфейсах пользователь вводит поисковый запрос в экранной форме главной страницы сайта, а затем указывает, какой каталог использовать для поиска (например, искать в электронном каталоге книг, искать в каталоге авторефератов, искать в электронной библиотеке и т.п.). Помимо этой возможности библиотеки, входящие в библиотечные корпорации, предлагают читателям поиск по ресурсам своей корпорации или другим источникам с использованием шлюза HTTP-Z39.50. Для проведения этого типа поиска пользователь, как правило, должен перейти на отдельную WEB-страницу. В результате такой

пестроты интерфейсов и неочевидности выбора информационных источников, пользователь теряет время и не всегда может найти интересующий его электронный ресурс.

С технологической точки зрения наиболее простым способом организации параллельного поиска по нескольким источникам является использование технологии «сервер Z39.50 – шлюз HTTP-Z39.50». Однако в практике создания библиотечных порталов эта технология обычно используется лишь для поиска по ресурсам библиотечных корпораций. Основными сдерживающими факторами являются следующие:

- отсутствие провайдера данных сервера Z39.50 для используемой АБИС или сложности в его настройке;
- отсутствие провайдера данных сервера Z39.50 для СУБД, использующихся в тематических и вспомогательных базах данных, а также сложности конвертирования данных из этих баз в другие системы;
- ограниченная функциональность сервера Z39.50 при поиске по полнотекстовым ресурсам (имеется ввиду, прежде всего, программный пакет ISITE), не позволяющая конкурировать с поисковыми модулями, например, Yandex;
- громоздкость поискового интерфейса шлюза HTTP-Z39.50, требующая от пользователя выбора каталогов, выбора поисковых атрибутов, указаний типа коммуникативного формата и формата представления результатов и т.д.;
- неэффективность поиска при выборе большого числа информационных источников, когда результаты исчисляются сотнями найденных записей, выводимых пользователю беспорядочно;
- отсутствие автоматической обобщенной библиографической записи для одного издания, что приводит к многочисленным повторам почти одинаковых библиографических описаний, выбранных из разных каталогов участников библиотечных корпораций.

При разработке концепции нового WEB-портала ВГБИЛ, были учтены преимущества технологии на основе протокола Z39.50, а также предусмотрены подходы, позволяющие решить часть вышеперечисленных проблем. Учитывая, что во ВГБИЛ внедряется новая АБИС, в основе которой лежит СУБД ORACLE, было предложено организовать общедоступный электронный каталог (OPAC) на основе шлюза HTTP-Z39.50. В качестве сервера Z39.50 будет использоваться сервер Z64, созданный ВГБИЛ совместно с ГПНТБ России, для которого разрабатывается оригинальный провайдер данных для СУБД на основе языка SQL-запросов. Преимуществами сервера Z64 являются простота использования, поддержка UNICODE (UTF-8), наличие провайдера данных для ИРБИС64.

Новый провайдер данных (SQL) будет использоваться, также, для доступа к СУБД электронной библиотеки ВГБИЛ, которая работает на основе MySQL.

Для доступа через портал HTTP-Z39.50 к ресурсам многочисленных тематических баз данных, формируемых отделами ВГБИЛ, осуществляется периодический перенос данных из этих баз в ИРБИС (в дальнейшем – в ИРБИС64), что позволяет использовать встроенный провайдер данных сервера Z32 (Z64). Полный переход к замене программного обеспечения, используемого для формирования этих баз данных, потребует еще немало времени, поэтому в данном случае более оправдано использование коммуникативных форматов для передачи данных во вспомогательные базы данных, подключенные к серверу Z39.50. В настоящее время по этой технологии формируются базы данных Информационного центра ВГБИЛ («Многоаспектная роспись периодики»), сводного каталога страхового регистра микроформ, базы данных редких и трофейных книг отдела редкой книги ВГБИЛ.

Для организации поиска по полнотекстовым материалам и документам WEB-сайта ВГБИЛ в экспериментальном режиме работает сервер Z39.50 на основе программного пакета ISITE. Несмотря на ограниченную функциональность, этот сервер позволяет осуществлять поиск по документам форматах HTML, WORD, RTF, что дает возможность использовать это решение для организации обобщенного поиска по сайту и базам данных ВГБИЛ. Для детального поиска по документам электронной библиотеки ВГБИЛ предполагается использование поискового модуля системы Yandex.

Пользователь портала ВГБИЛ будет взаимодействовать с интерфейсом главной страницы, предлагающей искать по любому слову или фразе, введенной в строке «поиск». В процессе первого шага обработки введенного запроса происходит обращение к шлюзу HTTP-Z39.50 с запросом поиска по любому полю следующих подключенных баз данных локальных источников: электронный каталог, электронная библиотека (поиск по библиографическим описаниям), тематические базы данных ВГБИЛ (многоаспектная роспись), полнотекстовый поиск по документом сайта ВГБИЛ. Результаты первого шага поиска отображаются количеством найденных записей в каждой базе данных. После выбора источника данных (нажатием кнопки мыши) пользователю выводятся найденные библиографические описания (для баз данных) или список документов сайта (для полнотекстового поиска). На этом же этапе можно перейти к работе с формой расширенного поиска, предлагающей заполнить несколько поисковых полей, связанных логическими операторами (И, ИЛИ, НЕ). Если результаты поиска в локальных источниках пользователя не устраивают, то можно перейти к поиску по тому же запросу в расширенном информационном пространстве, включающем: локальные источники, участвовавшие на первом этапе поиска, дополненные региональными источниками из состава корпоративной сети московских библиотек. Результаты поиска выводится в виде количества найденных записей в каждом каталоге. На третьем этапе к вышеуказанным источникам добавляется возможность подключения выбранных российских и зарубежных групп источников (серверов Z39.50). При формировании алгоритма многошагового поиска разработчики исходили из предположения, что для пользователя портала ВГБИЛ наибольший интерес будут иметь локальные источники информации (локальные базы данных). Если с порталом работает читатель библиотеки, то, возможно, его заинтересует возможность найти нужное издание в одной из московских библиотек. Если же с порталом работает пользователь из российских регионов, то его может заинтересовать поиск найденного (или отсутствующего) издания в библиотеках своего региона.

Таким образом, при разработке портала ВГБИЛ разработчики поставили цель реализовать библиотечный поисковый портал в большей степени ориентированный на нужды читателя, сочетающий возможности интегрированного поиска по многим информационным источникам на основе технологии Z39.50 с простотой интерфейса. Помимо описанных в докладе решений, в процессе разработки портала были предложены новые алгоритмы, позволяющие минимизировать задержки, возникающие из-за наличия неработающих или неактивных серверов Z39.50, эффективность которых можно будет оценить после апробации.