Протокол Z39.50: обзор.

Капустин В.А.[*], Смирнов В.М.[+]

Санкт-Петербургский государственный университет
Санкт-Петербург

[*]*vak@mail.nw.ru* [+]*vms@bmc.usr.pu.ru*

Предыстория

Текстовые информационно-поисковые системы (ИПС) имеют не менее богатую историю, чем системы управления базами данных (СУБД). Данные, с которыми оперируют ИПС, имеют три основные свойства, отличающие ИПС от [реляционных] СУБД:

- Атрибуты могут иметь текстовое значение переменной длины; поиск выполняется не по значению и не по его начальной подстроке, а по сложным комбинациям **слов** из текста атрибута (определение **слова** непросто, однако для целей настоящего обзора можно представлять себе слова естественного языка).
- Атрибуты могут иметь много (или ни одного) экземпляра
- Атрибуты могут иметь структуру (говоря на современном языке, "быть объектами")

Последнее отличие можно снять стандартными приемами нормализации, применяемыми при разработке реляционных СУБД, однако первые два остаются, причем первое является фундаментальным. Именно на поддержку поиска "**по словам**" были и остаются направлены усилия разработчиков ИПС.

Развитие ИПС началось в конце 50-х – начале 60-х годов. Так, одна из ИПС, ведущая свою историю с тех времен, распространена до сих пор – это CDS/ISIS (Computer Document System – 1963 г., компьютер IBM/709)[ISIS] работает сейчас на персональных компьютерах (MS-DOS, Windows), в разнообразных версиях Unix и в среде VAX/VMS. CDS/ISIS распространяется UNESCO и установлен примерно на 14 000 компьютеров по всему миру. Одновременно с развитием локальных ИПС развивались (прежде всего, в США) и так называемые "**базы данных с удаленным доступом**", примером которых может служить переживающий сейчас кризис DIALOG, имеющий в своем составе отдельные тематические ИПС по сотням отраслей (например, Physics Abstracts, New York Times – за пятнадцать лет; ИПС по организациям, предоставляющим гранты – описания нескольких десятков организаций и др.). Доступ к этим базам был и остается дорог, поэтому языки запросов к ним и способы работы с ними важны даже с чисто экономической точки зрения (например, в DIALOG каждый атрибут выводимой записи имеет цену). В конце 70-х годов IBM обобщила подходы к организации поиска в ИПС в своем пакете программ STAIRS, ориентированном на большие ЭВМ. Язык запросов STAIRS стал стандартом де-факто, а затем был стандартизован и ISO[ISO8777]. Все современные ИПС явно или неявно поддерживают STAIRS

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

Модель данных, с которыми оперируют ИПС

Хранимые в ИПС данные представляют собой **записи**. Каждая запись имеет поля, типы которых не имеют для поиска большого значения. Над записями определен набор (обратимых) функций (**"точек входа"**), каждая из которых порождает одно или несколько значений. Множество значений, порождаемых одной из таких функций, называется **атрибутами записи**, а совокупность всех таких множеств для всех точек входа – **поисковым образом** записи – документа (**ПОД**). Именно с **ПОД** имеет дело STAIRS: запрос STAIRS представляет собой булево выражение над частными поисковыми критериями, каждый из которых обращен к единственному атрибуту. В простейшем случае частный критерий определяет требование полного совпадения термина, указанного в критерии, со значением атрибута, однако STAIRS допускает и более сложные частные критерии – прежде всего, левое усечение ("звездочка на конце").

Именно функции точек входа отвечают за разбиение текста, хранимого в полях записей, на слова. Они же отвечают за учет или неучет многоэкземплярности полей при поиске.

Стратегии поиска. Сеанс

Поиск информации выполняется людьми для удовлетворения их информационых потребностей (ИП). Информационная потребность в принципе невербализуема; человек, выполняющий поиск, определяет соответствие найденных документов ИП "на глазок". Документы, соответствующие ИП, называются **пертинентными (pertinent)**. Наличие непертинентных документов в наборе документов, представляющем результат поиска, может быть следствием многих причин, одна из которых – принципиально неточное преобразование

ИП в запрос на формальном языке (STAIRS).

Поэтому стратегия поиска часто ориентирована на **последовательное уточнение запроса** для отражения в нем тонких отличий в представлениях человека-поисковика и человека (коллектива) – проектировщика данных ИПС. Последовательное уточнение запроса может требовать (и требует) возвратов к предыдущим запросам и результатам поиска в них. STAIRS поддерживает ссылку на результаты поиска по предыдущим запросам как на частный критерий.

Результаты поисков всех пользователей крупной ИПС невозможно хранить централизованно из-за их потенциально большого объема. Как мы вскоре увидим, их не всегда возможно сохранить и на клиентской стороне. Поэтому возникает понятие **сеанса**, в течение которого возможна поддержка сохранения результатов поисков средствами ИПС. Если необходимо хранение результатов поиска вне рамок сеанса, то это необходимо делать другими, внешними по отношению к ИПС, средствами. Обратим внимание на то, что самый распространенный в Интернете протокол – HTTP – бессеансовый.

Вывод найденного

Вывод найденных документов также может представлять проблему:

- документов может быть слишком много (невозможно не только просмотреть. но даже и передать по каналу связи за разумное время)
- документы могут быть слишком большими
- Полный набор полей документа может быть нежелательным (например, по экономическим причинам)

Поэтому ИПС должна иметь средства управления выводом.

Z39.50

Разработка стандарта, обеспечивающего взаимодействие поисковой программы-клиента с удаленной ИПС с поддержкой необходимого для поиска сервиса (в виде протокола уровня приложения) началась в начале 80-х годов под руководством Библиотеки Конгресса США. В 1984 г. была подготовлена предварительная версия стандарта, обеспечивавшая поиск только библиографических данных, а в 1988 г. – версия 1 протокола Z39.50 (по названию рабочей группы), формально обеспечивающая поиск и данных других типов (не только библиографии). Эта версия протокола к настоящему времени окончательно устарела. В 1992 г. была утверждена вторая версия, соответствующая расширениям протокола, появившимся в процессе его доведения до уровня стандарта ISO[ISO10163].

Разработка действующей версии стандарта[Z39.50-1995] была начата в декабре 1991 г. на базе Агентства по сопровождению Z39.50, образованного при Библиотеке конгресса США и проводилась по апрель 1995 г., причем каждая черновая версия проекта проходила обсуждение на совещаниях специальной Группы ZIG, через списки рассылки и экспертизу Агентства Z39.50. Результаты этой работы учитывались в последующих версиях стандарта, которые позволили в 1995 г. создать ныне действующий стандарт Z39.50-1995. В настоящее время ведется разработка четвертой версии стандарта, наиболее значительные отличия которой от предыдущих версий включают поддержку SQL в качестве одного из допустимых языков запросов и регулярных выражений в качестве локальных критериев других языков запросов.

Стандарт определяет протокол типа клиент/сервер для информационного поиска. Он включает процедуры и структуры для поиска в разнородных базах данных для клиентов, обеспечиваемых сервером. Поддерживаются контроль доступа, удаленное обслуживание и средства помощи. Протокол определяет форматы и процедуры, управляющие обменом сообщениями типа запрос/ответ между клиентом и сервером, необходимыми при выполнении поиска в базах данных и идентификации записей, которые отвечают заданным критериям, а также получения (извлечения) некоторых или всех идентифицированных записей.

Клиент может инициировать запрос в интересах пользователя, протокол адресует передачи между соответствующими приложениями информационного поиска клиента и сервера, которые могут быть реализованы на разных компьютерах. Взаимодействие между [программой-]клиентом и пользователем лежит вне рамок протокола Z39.50. Протокол включает в себя понятие Z-ассоциации - сессии (сеанса), постоянно поддерживаемой между клиентом и сервером, и разрываемой только при помощи специальной службы.

Таким образом, все взаимодействия между клиентом и сервером происходят только в рамках установленной Z-ассоциации, что обеспечивает протоколу наличие памяти и позволяет избежать явного сохранения промежуточных данных в процессе взаимодействия.

Протокол предусматривает существование различных наборов атрибутов для поиска и синтаксисов описания записей. Агентство по сопровождению Z39.50 ведет реестр наборов атрибутов.

Основные службы протокола Z39.50

В самом протоколе объявлены 11 сервисов, которые и выполняют весь спектр действий предусмотренных в протоколе:

- служба **Init**: инициализирует Z-ассоциацию, и позволяет клиенту и серверу обменяться информацией о поддерживаемых службах.
- служба **Search**: создает результирующее множество в соответствии с заданными клиентом критериями поиска. В качестве средства для отбора записей предусмотрено три типа запросов: 0, 1, 101.
  - **0** – это запрос с произвольным синтаксисом.
  - **1** – это запрос, записываемый в обратной польской записи (RPN) с операторами **AND**, **OR**, **ANDNOT** (язык, равносильный STAIRS).
  - **101** – запрос аналогичный 1, но с дополнительным оператором **PROXY**, описывающим синтагматическую близость связываемых этим оператором терминов в тексте искомой записи.

В качестве терминов используется текст, сопряженный с произвольным набором атрибутов, несущих как служебную, так и поисковую информацию. В качестве результата служба возвращает идентификатор результирующего множества и его мощность.

- служба **Present**: отвечает за передачу записей между клиентом и сервером. Передача предусматривает выбор синтаксиса (если это необходимо), также возможна сегментация результирующего множества

и/или отдельных документов.

- служба **Scan**: осуществляет сканирование базы данных с целью извлечения терминов для поиска и предоставления их списка клиенту. В качестве параметров принимает термин-шаблон, размер шага, количество возвращаемых терминов.
- служба **Explain**: позволяет клиенту получать с сервера необходимую для поиска и настройки вспомогательную информацию. В протоколе существуют специализированные форматы, призванные описывать синтаксисы поддерживаемых форматов, подключенные к серверу базы данных, поддерживаемые наборы атрибутов и т.д.
- служба **Sort**: осуществляет сортировку результирующего множества.
- служба **delResSet**: позволяет клиенту удалять созданные им результирующие множества.

Остальные службы протокола имеют вспомогательный характер и не относятся непосредственно к процедуре поиска. Особняком стоит слежба **ExtendedService**, которая позволяет присоединять к основным службам протокола произвольные службы, создаваемые владельцами сервера.

Z39.50 – это попытка стандартизировать процедуру поиска. За счет такой стандартизации на основе этого протокола можно строить различные поисковые системы, с самой различной архитектурой. На сегодняшний момент протокол широко используется в области поиска библиографической информации[LOC], для обслуживания программы GILS (Government Information Locator Service[GILS1] в США; Global Information Locator Service[GILS2] и в других странах), в медицинских базах данных и в экологических[GELOS] и геодезических программах типа FGDC[FGDC], а также в распределенных географических информационных системах (ArcView[ESRI] и многочисленных других[NSDI]). Перечень URL со ссылками на географические ИПС можно найти, например, на сервере "GIS Online"[DGI]

Минимальная конфигурация системы на основе Z39.50 – это программа-сервер + программа-клиент, в большинстве случаев связанные между собою по сети TCP/IP – через Интернет (Z39.50 входит в число "известных служб", и имеет порт 210). Также в эту конфигурацию могут добавляться программы-шлюзы, тезаурусы, PROXY-серверы и др. средства.

Реализации Z39.50

Говоря о существующих реализациях протокола, необходимо подразделить эти реализации на четыре категории:

- серверные части
- клиентские части
- программы-шлюзы
- вспомогательное программное обеспечение

Серверные части протокола Z39.50 представлены на сегодняшний момент в двух категориях: коммерческие и бесплатные. Среди коммерческих серверных программ (всего их насчитывается более 20) наиболее выделяется сервер Z39.50 **MetaStar** компании **BlueAngel Technologies**, который обладает большим набором возможностей и во всей полноте реализует протокол. Среди бесплатных серверов существуют лишь два: **Isite**, разрабатываемый организацией **CNIDR** и **Zebra server**, разрабатываемый датской компанией **Indexdata**. Каждая серверная реализация протокола обладает собственной, зачастую нигде более не встречаемой спецификой. К примеру, **Isite** не поддерживает именование результатов и оператор **PROXY**, но корректно поддерживает индексацию кириллицы на платформе Windows NT, а **Zebra server** наоборот.

Все реализации также существенно различаются по типам хранимых данных (тех, для которых могут быть построены индексы, хранящие значения точек входа): например **MetaStar** ориентирован на формат **XML**, **Isite** поддерживает **SGML**, **HTML**, и др., а **Zebra** – **GRS-1**, **Text**, **MARC**. Т.е. то, что в стандарте объявлено универсальным, на практике создает проблему совместимости.

Среди клиентских программ можно выделить две категории: специализированные клиенты, предназначенные для работы со строго определенными серверами или данными (прежде всего, библиотечные клиенты) и универсальные, способные конфигурироваться под каждую конкретную задачу (**Znavigator**, **Willow**).

Работа поддержана РФФИ, грант 98-07-91197.

Литература

[ISIS]
Пакет прикладных программ CDS/ISIS/M версия 2.3. – М:МЦНТИ,1991

[ISO8777]
ISO 8777 – Information and Documentation – Commands for Interactive Text Searching 1987

[ISO10163]
ISO 10163 – Information and Documentation – Search and Retrieve Application Protocol Specification for Open Systems Interconnection 1991

[Z39.50-1995]
ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. – Z39.50 Maintenace Agency Offical Text for Z39.50-1995 July 1995 [http://lcweb.loc.gov/z3950/agency/] and [http://www.niso.org/z39.50/z3950.html]

[LOC]
The Library of Congress Page for gateway access to LC's catalog and those at many other institutions // [http://lcweb.loc.gov/z3950/]

[GILS1]
Government Information Service // [http://www.gils.org]

[GILS2]
Global Information Locator Service // [http://www.gils.net]

[NSDI]
NSDI (National Spatial Data Infrastructure) Clearinghouse Nodes// [http://130.11.52.178/clearinghouse_sites.html]

[ESRI]
Geography Matters // [http://www.esri.com/]

[GELOS]
Global Environmental Information Locator Service //
http://ceo.gelos.org/

[FGDC]
Federal Geographic Data Committee //
[http://www.fgdc.gov/]

[DGI]
DGI Software //
[http://www.geog.byu.edu/gisonline/links/soft.htm]