Поисковая навигация в Интернет на основе динамической классификации крупных информационных массивов в PCBIRS технологии

В.Ю. Бугаев, А.С. Паринов

Публикация информации в сети Интернет в виде отдельных страниц, связанных гипертекстовыми ссылками, сталкивается с проблемами классификации документов. Для того чтобы упростить конечному пользователю процедуру поиска необходимой информации, администраторы WEB-узлов обычно размещают документы в отдельных каталогах в соответствии с их тематикой. Такой способ размещения информации имеет ряд недостатков.

Во-первых, каталог создается на этапе разработки и пользователь не в состоянии его менять. Во-вторых, документы сортируются администратором, который не может учесть всех интересов потенциальных пользователей. Чаще всего документ принадлежит нескольким тематическим областям и должен содержаться в нескольких каталогах, что создает дополнительные трудности при его размещении.

Другим существенным недостатком публикаций в виде отдельных страниц является жесткость форм ее отображения. В том случае, когда в результате поиска пользователь находит множество документов, необходим механизм получения обзора содержания последних, а это связано с вариацией форм представления информации (таблицы, списки данных и т.д.).

Очевидное решение этих проблем - публикация информации в виде полнотекстовых баз данных. Кроме преимуществ, связанных с хранением и администрированием, публикация информации в виде баз данных, позволяет реализовать ряд механизмов повышающих эффективность поисковой навигации.

К таким механизмам, прежде всего, относится динамическая классификация документов, которую конечный пользователь может осуществлять самостоятельно в соответствии с областью собственных интересов. Этот механизм является одним из основных средств поиска и анализа информации в системе PCBIRS [1-3], которая обеспечивает работу с крупными массивами текстов в локальной среде. Однако подобные механизмы практически отсутствуют в среде Интернет, что на наш взгляд является серьезным недостатком.

©Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

С технической точки зрения PCBIRS представляет собой систему управления документально-фактографическими базами данных. Все документы, поступающие в базы данных, проходят автоматическую лексическую индексацию текста, на основе которой пользователю предоставляется возможность поддержки множества динамических понятийных классификаторов.

Классификатор - это пакет именованных контекстных запросов. Каждому запросу, входящему в классификатор, присваивается имя - понятие. Предполагается, что если документ удовлетворяет запросу, то он содержит информацию о соответствующем понятии. Пользователь формулирует область своих интересов в виде списков понятий, которые он может накапливать и проецировать на различные информационные массивы для получения обзора содержания последних. Тексты запросов могут содержать гиперссылки на ранее определенные понятия, которые в процессе выполнения запросов интерпретируются, как макро поисковые термины. Таким образом, список понятий может отображать как простые представления (например, синонимию терминов), так и сложные сети отношений. Однородные группы понятий, образующие с точки зрения пользователя некое новое понятие, могут объединяться в подкаталоги, образуя тем самым иерархическое дерево классификации.

Следует подчеркнуть, что сам классификатор никак не связан с тем или иным информационным массивом. Он всего лишь является отражением области представлений и интересов пользователя, сформулированной на языке контекстных запросов. По сути, классификатор представляет собой базу знаний о том, как те или иные понятия могут быть выражены в текстах документов. Один и тот же классификатор может использоваться для навигации в любых базах данных PCBIRS.

Поскольку динамическая классификация даже крупных информационных массивов в сотни мегабайт занимает в PCBIRS, как правило, несколько секунд, пользователь может свободно менять свою точку зрения, фиксировать различные множества документов и проецировать на них другие списки понятий, что в конечном итоге представляет мощное средство поисковой навигации. С одной стороны, еще до формулировки запросов он видит, что

же содержит данный информационный массив (с его точки зрения), с другой стороны, выполнив запрос на поиск документов, у него имеется возможность понять, что же содержит найденное множество.

Включение всех возможностей поисковой навигации локального варианта PCBIRS (контекстный анализ, извлечение из текстов и создание виртуальных списков данных, мониторинги результатов поиска и т.д.) для доступа к информации в сети Интернет на сегодняшний день является проблематичным, поскольку это связано с серьезным увеличением трафика. Однако исследование механизмов динамической классификации информации в сети Интернет представляется настолько актуальным, что для этих целей был разработан PCBIRS HTML сервер, интегрированный с WEB-сервером через интерфейс ISAPI.

Сервер позволяет публиковать в сети как полнотекстовые, так и структурированные базы данных, подготовленные средствами PCBIRS, обеспечивая:

1) выполнение простых и сложных (с учетом структуры документов) запросов на контекстный поиск;

2) выполнение каталогов запросов (пакетная динамическая классификация);

3) хранение на сервере битовых матриц для выполненных запросов с возможностью быстрого перехода на соответствующее множество найденных документов;

4) фиксацию множества документов для последующих запросов;

5) генерацию HTML документов для отображения:

5.1) словарей базы данных (возможность просмотра и выбора поисковых терминов в тексты запросов);

5.2) списков запросов пользователя;

5.3) списков классификационных запросов (иерархия классов и списки понятий);

5.4) текстов документов базы и их списков в различных формах (выбор фрагментов просмотра, таблицы данных и т.д.)

Кроме того, сервер реализует функции сопровождения пользователей, мониторинга их активности, динамического выделения ресурсов для пользователей.

Один из вариантов прикладного интерфейса между пользователями Интернет и PCBIRS HTML сервером реализован на базе обозревателя WEB Microsoft Internet Explorer и механизма DHTML.

При первом подключении к серверу трафик, передаваемый пользователю, составляет около 10 кБ (это связано с необходимостью пересылки клиенту блока программ на языке Java Script). На рис. 1 представлены основные фрагменты рабочего экрана пользователя. В верхней части экрана (фрейм 1) предоставляется возможность ввода запросов на контекстный поиск в базе данных. Результаты поиска отображаются во фреймах 4 и 5 (списки и тексты документов соответственно).

Все запросы пользователя и результаты их выполнения кэшируются на сервере. Список выполненных запросов отображается в 3 фрейме. Помимо этого в 3 фрейме отображается содержимое классификаторов и статистика по выполненным запросам. Предусмотрена также возможность сортировки понятий по различным критериям.

Для удобства пользователя при работе с базой неиспользуемые фреймы можно скрыть, освобождая место для просмотра наиболее актуальных областей. Возможность комбинировать фреймы обеспечивает в известном смысле универсальность данного интерфейса для поиска как в полнотекстовых, так и в структурированных базах данных (рис. 2).

Динамическая классификация баз данных, размещенных в Интернет, проводится в несколько этапов.

Прежде всего, пользователь может работать с множеством готовых каталогов запросов, подготовленных разработчиком базы, и выбирать их из списка по мере необходимости. Таким образом, реализована возможность просмотра содержимого информационного массива с различных точек зрения.

Кроме того, пользователь может создавать собственные каталоги запросов, которые готовятся в виде простых текстовых файлов для передачи серверу. В этих файлах пользователь описывает интересующие его понятия, создает связи между ними и по мере необходимости объединяет в группы. Затем он подключается к PCBIRS HTML серверу, выбирает базу данных или группу баз данных для классификации. После загрузки классификатора на PCBIRS HTML сервер и его выполнения, пользователю предоставляются результаты динамической классификации в виде иерархического списка, в который включаются имена понятий и классов, количество документов, которые удовлетворяют соответствующему запросу. Выбор элемента из классификатора означает переход на просмотр документов, удовлетворяющих соответствующим запросам.

Для анализа содержания множества документов достаточно его фиксировать. При этом сервер автоматически генерирует для него список понятий и классов. Пользователь имеет возможность проецировать на фиксированное множество другие классификаторы. Таким образом поисковая навигация постоянно сопровождается обзором содержания найденных документов.

Следует заметить, что описанная выше технология предъявляет ряд дополнительных требований к WEB серверу. Это связано с необходимостью выполнения больших пакетов запросов в многопользовательском режиме и поддержки рабочих пулов для каждого клиента. Поэтому ее дальнейшая реализация требует продолжения исследований в области измерений временных характеристик и оценки потребляемых ресурсов.

Список литературы

1. Бугаев В.Ю., Белоцерковский А.В. "Информационно поисковая аналитическая система PCBIRS 3.0", "Мир ПК"#12 1997г., с. 54-57

2. Бугаев В.Ю. "Индексация информационных массивов в PCBIRS 3.2", "Мир ПК"#8 1999г., с. 76-77

3. Бугаев В.Ю. "Динамическая классификация информации в системе PCBIRS, как метод поисковой навигации в крупных информационных массивах". Тезисы докладов Всероссийской научно-практической конференции "Проблемы организации использования результатов научно-технической деятельности в интересах экономического и социального развития регионов Российской Федерации"с. 29-33

4. Бугаев В.Ю., Паринов А.С. "Публикация баз данных и доступ к крупным информационным массивам в сети Internet с использованием PCBIRS технологии". Тезисы докладов Всероссийской научно-практической конферен-

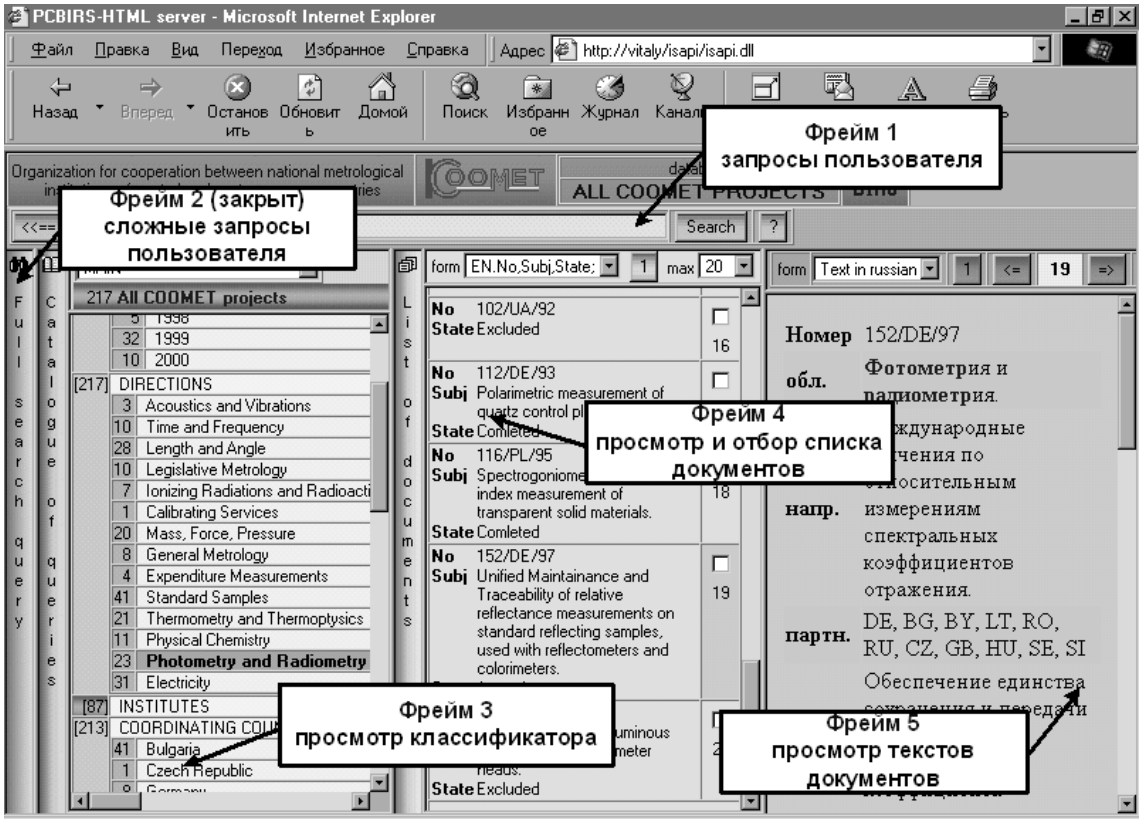
PCBIRS-HTML server - Microsoft Internet Explorer
Фрейм 1
запросы пользователя
Фрейм 2 (закрыт)
сложные запросы
пользователя
Фрейм 4
просмотр и отбор списка
документов
Фрейм 3
просмотр классификатора
Фрейм 5
просмотр текстов
документов

Рис. 1:

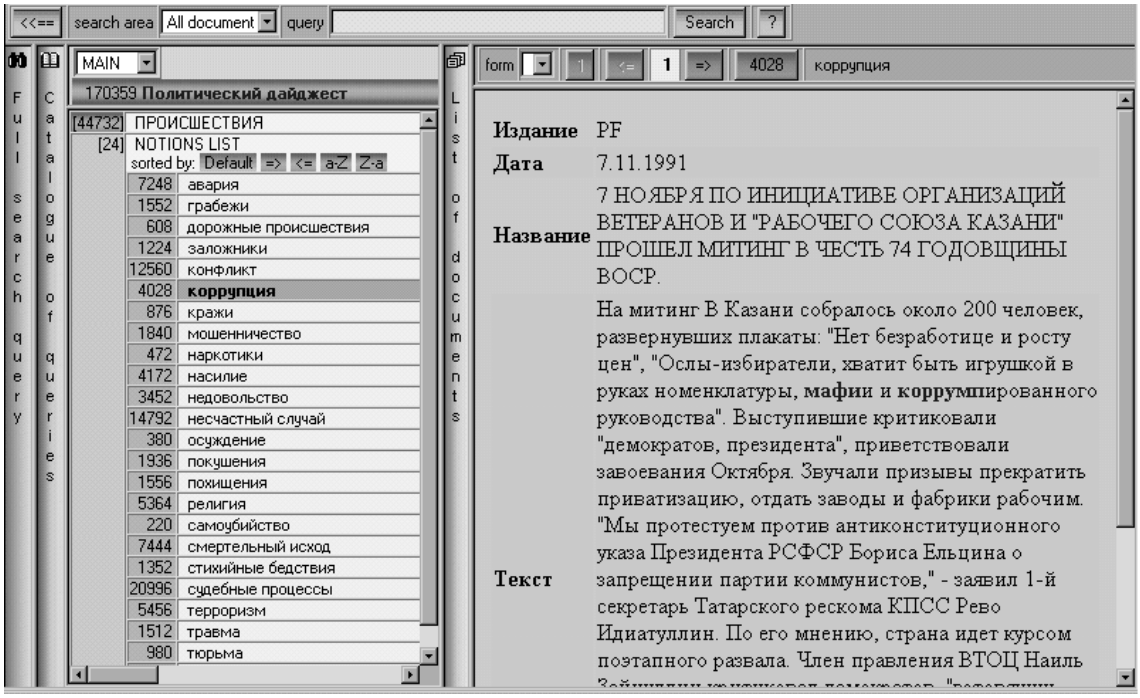
search area
All document
query
Search
MAIN
170359 Политический дайджест
[44732] ПРОИСШЕСТВИЯ
[24] NOTIONS LIST
коррупция
Издание PF
Дата 7.11.1991
Название 7 НОЯБРЯ ПО ИНИЦИАТИВЕ ОРГАНИЗАЦИЙ ВЕТЕРАНОВ И "РАБОЧЕГО СОЮЗА КАЗАНИ" ПРОШЕЛ МИТИНГ В ЧЕСТЬ 74 ГОДОВЩИНЫ ВОСР.
Текст На митинг В Казани собралось около 200 человек, развернувших плакаты: "Нет безработице и росту цен", "Ослы-избиратели, хватит быть игрушкой в руках номенклатуры, мафии и коррумпированного руководства". Выступившие критиковали "демократов, президента", приветствовали завоевания Октября. Звучали призывы прекратить приватизацию, отдать заводы и фабрики рабочим. "Мы протестуем против антиконституционного указа Президента РСФСР Бориса Ельцина о запрещении партии коммунистов," - заявил 1-й секретарь Татарского рескома КПСС Рево Идиатуллин. По его мнению, страна идет курсом поэтапного развала. Член правления ВТОЦ Наиль

Рис. 2:

ции "Проблемы организации использования результатов научно-технической деятельности в интересах экономического и социального развития регионов Российской Федерации"с. 27-29

ОБ АВТОРАХ

Бугаев Виталий Юрьевич - к.ф.-м.н, руководитель лаборатории ВНИИ Физико-технических и радиотехнических измерений (ВНИИФТРИ), автор и разработчик информационно-поисковой аналитической системы PCBIRS, www.chat.ru/~birs, телефон: (095)535-08-52, e-mail: bgv@ftri.extech.msk.su

Паринов Андрей Сергеевич, аспирант ВНИИ Физико-технических и радиотехнических измерений (ВНИИФТРИ), телефон 209-38-54, e-mail: aspmail@mail.ru