2002 | Том 5 | Выпуск 2

Технологические аспекты организации доступа к разнородным информационным ресурсам в университетской информационной системе Россия

Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н.

**Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова;
АНО Центр информационных исследований**

1. Введение

Университетская информационная система РОССИЯ (далее - УИС РОССИЯ, http://www.cir.ru) [1] создана как ресурсная база Российского университетского информационно-исследовательского консорциума по социальным и гуманитарным наукам (Russian inter-University Social Sciences Information and Analytical consortium - RUSSIA Consortium). Предназначена для проведения исследований по социальным наукам и открыта для коллективного доступа ученым и исследователям из университетов РФ.

В работе над развитием УИС РОССИЯ участвуют специалисты Научно-исследовательского вычислительного центра МГУ им. М.В.Ломоносова (НИВЦ МГУ) и АНО Центр информационных исследований.

В настоящее время УИС РОССИЯ содержит более 400 тысяч полнотекстовых документов, поступающих и более чем 50ти источников, и более 200 тысяч библиографических описаний из коллекции СОЦИОНЕТ/RePEc (Всего более 3.5Гбайт текстов). Все материалы, доступные в УИС РОССИЯ, получены на основе прямых договоров с правообладателями соответствующих ресурсов на условиях гарантии некоммерческого использования полученных документов.

Таблица 1. Основные информационные ресурсы УИС РОССИЯ

	Источник	Ретроспектива	Количество
Правовые акты	НТЦ Система	1990 - ...	50,000
Стенограммы заседаний Государственной Думы	Аппарат ГД ФС РФ	1994 - ...	100,000
Статистические материалы	Госкомстат РФ; Межгос. Стат.	1998 - ...	15,000

Материалы СМИ	Комитет СНГ "Эксперт", "Независимая газета", "Известия", "Комсомольская правда", "Аргументы и факты", "Слово", ...	1997 - ...	180,000
Аналитические материалы	материалы министерств и ведомств РФ, Счетная палата РФ, ЦБ РФ, РЕЦЭП	1996 - ...	15,000
Научные издания	Вестник МГУ, "Соц.исследования"	1998 - ...	600
Библиографические описания материалов по экономике, социологии, ...	СОЦИОНЕТ / RePEc	...	230,000

Решаются следующие задачи интеграции разнородных информационных ресурсов:

1. обеспечение единообразного формата хранения документов разных источников;
2. единообразные способы доступа ко всей коллекции документов;
3. использование специфических поисковых атрибутов для каждой коллекции;
4. тематическая систематизация /классификация документов по тезаурусу, рубрикаторам;
5. аннотирование полнотекстовых документов;
6. создание предметно-ориентированных баз данных, интегрированных в общую систему.

Дополнительные требования:

- максимально возможная автоматизация включения нового источника в интегрированную базу, сопровождение (пополнение, изменение интерфейса доступа) всех источников при минимальных затратах;
- существенная часть материалов предоставлена правообладателями информации на эксклюзивных условиях - бесплатный доступ к материалам, предоставляемыми самими правообладателями на платной основе. Должен быть обеспечен контроль за использованием данных материалов.

Реализованы технологические решения, которые будут рассмотрены далее:

- АРМ Администратора базы данных - CASE средство описания свойств

каждого информационного ресурса;
- Интегрированная библиотека конверторов: автоматическое определение формальных атрибутов документа (тип, дата, номер, и т.п.), преобразование в HTML;
- АЛОТ - Автоматизированная Лингвистическая Обработка Текстов: морфологический анализ, терминологический анализ, тематический анализ, автоматическое рубрицирование, автоматическое аннотирование;
- Поддержка интерфейса доступа к различным коллекциям - динамически формируемый набор специфических поисковых атрибутов для каждой коллекции;
- Средства администрирования доступа пользователей, мониторинг нагрузки системы.

Общая функциональная схема УИС РОССИЯ представлена на Рис.1.

Рис.1.

В целом, УИС РОССИЯ представляет собой (Рис.2) достаточно сложный программный комплекс, реализованный с использованием различных технологий. В качестве СУБД используется Oracle 8.1.7 (800 таблиц и индексов, более 500 миллионов записей).

Рис.2

WEB-интерфейс пользователя УИС РОССИЯ обеспечивается программой автоматической генерации HTML-страниц с использованием технологии Java Servlets. В качестве серверного программного обеспечения используется свободно распространяемое программное обеспечение Apache 1.3.20, Jakarta Tomcat servlet container 3.2.1, Java Developers Kit 1.3. Web-сервер работает под управлением Red Hat Linux 6.1. Взаимодействие Java-программы с базой данных Oracle реализовано при помощи JDBC.

Полнотекстовые документы могут храниться в базе данных или в файловой системе, в том числе в защищенных от внешнего доступа директориях.

Основной платформой является Windows NT, вместе с тем имеется опыт установки программного обеспечения для архитектуры, когда ORACLE функционирует под управлением другой операционной системе - в этом случае часть вычислений переносится на Windows-клиента с использованием механизма Java RMI.

2. Описание информационных ресурсов

Информация о документе в УИС РОССИЯ условно разделяется на:

- атрибуты (заголовок, дата, номер);
- перечислимые поля - атрибутные классификаторы (авторы, разделы);

- индекс для контекстного поиска по леммам (нормализованным словоформам);
- тематический индекс (по терминам тезауруса, рубрикаторам).

Для сопровождения поисковых индексов по атрибутам и атрибутным классификаторам в УИС РОССИЯ поддерживаются, так называемые, "классы документов" - абстрактные объекты, которые для каждой коллекции полнотекстовых документов согласовано описывают:

- как надо обрабатывать и куда загружать документы;
- какие таблицы Oracle и индексы по каким полям таблиц должны быть созданы;
- какие поисковые поля должны присутствовать в карточке запроса;
- каким образом отображать документы;
- какие пользователи имеют права просмотра документов класса и квоты просмотра.

Все данные о классах содержатся в специальных таблицах Oracle. Для управления содержимым этих таблиц на Borland Delphi разработано автоматизированное рабочее место администратора данных УИС РОССИЯ.

Исходные данные поступают в УИС РОССИЯ, в основном, в электронной форме. При этом наблюдается большое разнообразие форматов: WinWord документы из Госкомстата РФ, RTF документы из газеты "Слово", совокупности связанных HTML файлов из "Эксперта", структурированные и слабо структурированные ASCII файлы из "Независимой газеты" и других источников.

Создана библиотека программ-конверторов, которые преобразуют информацию документов разнообразных форматов в единый формат хранения. Дополнительно ставится цель представления документов в максимально удобном пользователю виде.

В рамках УИС РОССИЯ разработаны различные программы-конверторы, преобразующие документы различных провайдеров информации в единообразный формат хранения (HTML), одновременно выделяющие формальные атрибуты (вид документа, даты принятия) и структурные единицы (заголовки, приложения, деление на статьи) документов.

Выделяемые программами-конверторами атрибуты могут быть использованы:

- для управления процессом дальнейшей обработки документа;
- для поиска в виде строки, в том числе с символами усечения;
- для поиска по классификаторам, в том числе с пополняемым списком (например, авторы публикаций). В поисковой системе по описанию АРМа автоматически создается специальный интерфейс поиска элементов классификатора;
- для отображения информации о документе в карточке документа.

В настоящее время существуют конверторы для всех коллекций УИС РОССИЯ -

правовых актов, материалов СМИ, данных Госкомстата и т.д.

Реализован технологический комплекс утилит для автоматизированной подготовки документов для загрузки в полнотекстовую базу данных:

- контроль орфографии в массиве текстов;
- обнаружение и очистка неправильного смешанного употребления кириллицы/латиницы;
- устранение разрядки, переносов;
- разметка таблиц.

Одновременно автоматически определяются формальные атрибуты документов, свои для каждого из видов ресурсов: вид, номер, организация для нормативных актов; фамилия выступающего, номер заседания для стенограмм Госдумы, номер, автор, вид приложения для "Независимой газеты" и т.д.

4. Автоматизированная Лингвистическая Обработка Текстов

Программное обеспечение АЛОТ предназначено для автоматической интеллектуальной обработки поступающих потоков документов.

Этапы автоматизированной лингвистической обработки:

- Морфологический анализ;
- Терминологический анализ;
- Рубрицирование;
- Аннотирование.

4.1. Морфологический анализ

В ходе морфологического анализа русскоязычного текста всем словам анализируемого текста сопоставляются леммы (нормализованные словоформы) с соответствующей грамматической информацией (род, число, падеж, категория одушевленности и т.п.).

Размер используемого морфологического словаря - 130 тысяч лемм. В сочетании с простыми словарями, описывающими словообразование, это обеспечивает более чем 99.6% покрытие текстов российских правовых актов и материалов СМИ. Для незнакомых слов порождаются гипотезы, содержащие правильную лемму. Реализован морфологический анализ англоязычных и смешанных русско-английских текстов.

Результаты морфологического анализа позволяют, задавая для поиска одну словоформу, находить документы, содержащие любые возможные словоформы данного слова.

4.2. Терминологический анализ

Реализован на основе Тезауруса по общественно-политической тематике АНО Центр информационных исследований.

Тезаурус - это терминологический ресурс, реализованный в виде словаря понятий и терминов со связями между ними. Основное назначение тезауруса - помощь при информационном поиске: на основе связей тезауруса происходит автоматическое расширение запроса, навигация по связям тезауруса помогает пользователю точнее сформулировать сам запрос.

Тезаурус по общественно-политической тематике, включает более 27,000 понятий, 64,000 терминов, 105,000 прямых и 800,000 наследуемых отношений между понятиями. Существующая версия Тезауруса описывает терминологию, используемую в общественно-политической области, включая экономическую, политическую, военную, законодательную, социальную и другие сферы.

Точное название Тезауруса - *информационно-поисковый тезаурус по общественно-политической тематике для автоматического индексирования*.

Все части определения имеют смысл:

- "информационно-поисковый" - разработан специально для использования в информационном поиске для помощи пользователю при формировании (уточнении) запроса и для автоматического расширения условий запроса при поиске;
- "по общественно-политической тематике" - покрывает 95-99% лексики и терминологии любого русскоязычного текста из коллекции нормативных документов РФ и материалов СМИ с 1991 года;
- "для автоматического индексирования" - является основой для процесса автоматического определения тематики документов - группирования близких по иерархии тезауруса терминов в тематические узлы, автоматического рубрицирования и автоматического аннотирования .

Для многих известных тезаурусов (WordNet, Roget, EuroWordNet) является большой проблемой автоматический вывод по связям тезауруса - когда расширение на ближайшую окрестность верно, но не полно, а попытки расшириться на более широкую окрестность ведут к ошибкам. Имеющиеся попытки, тем не менее, использовать тезаурусы для ручного индексирования в условиях автоматической обработки (Excalibur/Convera, Oracle TextServer) за счет подбора весов, механистичного группирования терминов - подчас приводят к малопонятным результатам.

При построении тезауруса, используемого в УИС РОССИЯ, были предприняты специальные усилия (четкие критерии построения, обратная связь через анализ прозрачных результатов обработки), чтобы добиться возможности расширяться на всю задаваемую иерархическими отношениями окрестность. Кроме того разработана совокупность алгоритмов, существенно использующих фундаментальные свойства связного текста и возможности вывода по иерархии тезауруса.

Тезаурус представляет собой интегрированный комплекс подтезаурусов различных предметных областей, в частности, включает географический подтезаурус и подтезаурус персон. Географический подтезаурус описывает более

6 тысяч географических объектов: города, реки, моря, территории, более подробно - для России и бывшего Советского Союза, а также для наиболее известных географических объектов всего мира. Подтезаурус персон содержит сведения о современных и исторических деятелях, часто упоминаемых в материалах СМИ (более 1000 имен).

На базе Тезауруса осуществляется автоматическое концептуальное индексирование входящего потока текстов - создается список понятий, упомянутых в тексте, и производится процедура разрешения многозначных терминов, что принципиально при поиске по многозначным терминам, обозначающим разные понятия в разных контекстах.

4.3. Построение тематического представления текста

На первой стадии анализа в тексте ищутся термины, описанные в Тезаурусе (как слова, так и словосочетания).

На основе связей Тезауруса термины группируются по смысловой близости в так называемые "тематические узлы".

С учетом свойств связного текста тематические узлы классифицируются на:

- основные тематические узлы - моделирующие в совокупности основную тему документа;
- локальные тематические узлы - моделирующие темы, обсуждаемые в документе как второстепенные;
- все остальные термины - так называемые "упоминавшиеся термины".

В зависимости от того, элементом какой структуры они являются, все встречающиеся в документе понятия получают различные оценки релевантности относительно содержания документа.

4.4. Рубрикация текстов

Тематическое представление содержания документа является основой для ранжированного рубрицирования - вывода соответствующих документу рубрик с оценкой релевантности рубрики содержанию текста.

Реализовано более десяти различных систем рубрикации текстов, в том числе:

- по терминам верхнего уровня тезауруса Исследовательской службы конгресса Библиотеки конгресса США (Legislative Indexing Vocabulary, LIV) (80 рубрик);
- по Общему правовому тематическому классификатору ЦИК РФ (450 рубрик, 4 уровня иерархии);
- по рубрикатору ВЦИОМ (330 рубрик, 4 уровня иерархии);
- по Классификатору правовых актов РФ (Указ Президента РФ от 15.03.2000, около 1200 рубрик, 4 уровня иерархии) .

4.5. Аннотирование текстов

Определение в тексте основных тем позволяет выделить и предложения, в которых тематика документа представлена наиболее ярко (представлены все основные тематические узлы). Данные предложения образуют аннотацию текста.

Существует довольно много программ, которые выделяют в тексте наиболее информативные предложения. Подход, используемый в УИС РОССИЯ, отличается от других тем, что порождаемая аннотация производит впечатление связного текста, значительно облегчая восприятие.

Для текстов некоторых жанров - очень больших текстов, интервью и т.п. - построение хорошей аннотации из фрагментов исходного текста невозможно.

Разработана "структурная тематическая аннотация" [4], представляющая содержание текста в виде совокупностей концептуально связанных терминов. Структурная аннотация позволяет оценить содержание текста любого жанра с одного взгляда, кроме позволяет осуществить перевод на другой язык, подставив соответствующие переводы терминов.

5. Информационно-поисковая система

Информационно-поисковая система, используемая в УИС РОССИЯ, обладает всеми стандартными возможностями систем информационного поиска. Кроме того, УИС РОССИЯ обладает рядом уникальных особенностей, прежде всего динамическим формированием карточки запроса для разных типов архивов, смысловым поиском по тезаурусу и рубрикаторам.

Карточка запроса формируется динамически из поисковых атрибутов и классификаторов, приписанных к конкретной коллекции документов, например, "Номер документа" для нормативных актов, или "Выступающий" для стенограмм ГосДумы (Рис.3).

Рис. 3

Общими поисковыми возможностями для всех коллекций являются:

- поиск по дате создания документа,
- строка запроса по контексту с использованием морфологического разбора и возможностью задания логического выражения любой сложности,
- поиск по общественно-политическому тезаурусу (64 тысячи текстовых входов, сгруппированных в 27 тысяч понятий) и двум рубрикаторам (80 и 180 рубрик).

Поиск можно проводить по любому множеству коллекций, при поиске по одному ресурсу дополнительно доступны специфичные для коллекции атрибуты, описанные как поисковые для соответствующего класса документов.

Обработка запроса начинается (Рис.4) с разбора заполненной карточки запроса на основе метаинформации о классе документов (переход к уникальным идентификаторам элементов классификаторов, чтобы эффективно использовать табличные индексы). При необходимости производится морфологический разбор условий поиска по контексту. Результатом первого этапа является XML-подобная структура дерева условий запроса - внутреннее представление запроса.

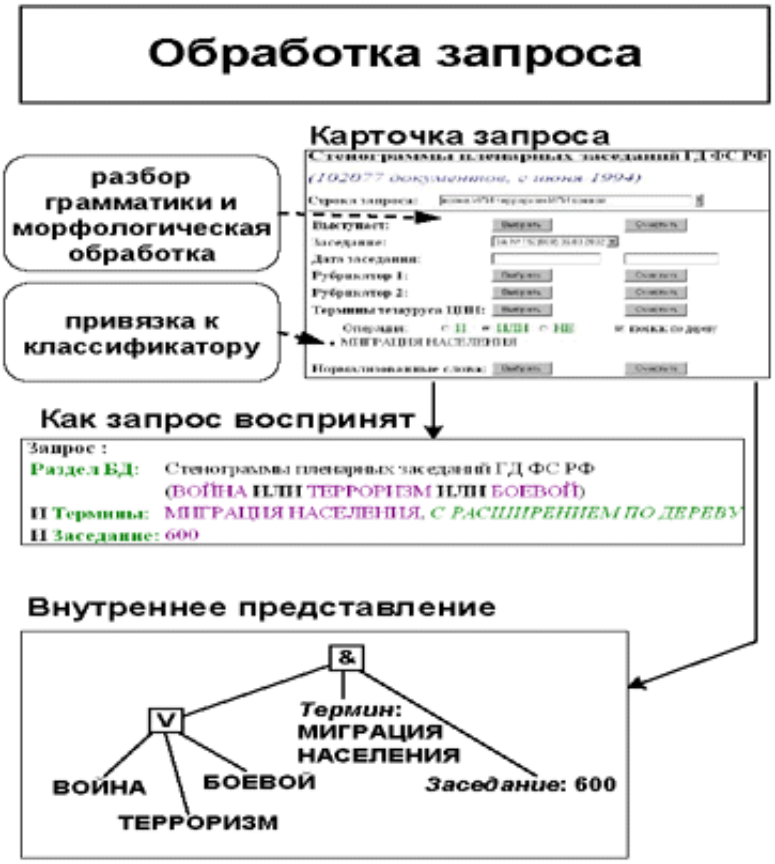

Рис. 4

На втором этапе по дереву внутреннего представления запроса динамически формируется SQL предложение (Рис.5), которое затем исполняется Oracle.

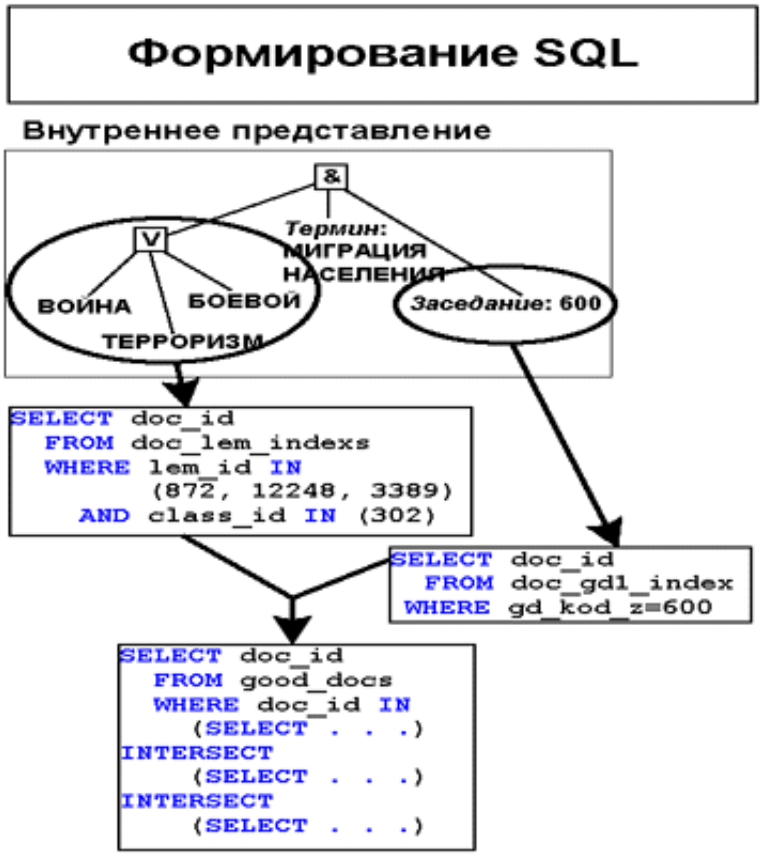

Рис. 5

Результаты поиска ранжируются в соответствии с оценкой релевантности содержимого документа запросу пользователя. Для ранжирования по контекстному поиску используется аналог формулы TF*IDF в формулировке системы INQUERY [5]. Детали ранжирования при поиске по терминам тезауруса и рубрикам описаны в [6, 7].

6. Администрирование доступа

Главные пользователи УИС РОССИЯ - ученые, исследователи из университетов и научных институтов РФ. Существует определенная процедура регистрации как коллективного пользователя (организации, факсом за подписью руководителя организации на имя директора НИВЦ МГУ), так и индивидуального пользователя. Основным условием является обязательство использовать ресурсы УИС РОССИЯ только в исследовательских целях, не распространять их далее в коммерчески значимых масштабах.

Допустима "свободная регистрация", предоставляющая ограниченный ознакомительный доступ к информационным ресурсам. Поисковая система УИС РОССИЯ используется также пользователями сайта "Бюджетная система РФ" (www.budgetrf.ru), которые входят в систему без регистрации и имеют доступ к ограниченному списку открытых источников.

Технологически все пользователи УИС РОССИЯ разделены на группы. С помощью специального программного обеспечения для каждой группы пользователей назначаются те или иные привилегии доступа к информационным ресурсам.

Основные ограничения на группу пользователей:

- возможность работать только с IP адресов, удовлетворяющих условиям на маску;
- возможность доступа к заголовкам ресурса (иначе - ресурс не виден);
- квота на просмотр определенного количества документов отдельного ресурса в течение суток;
- возможность использовать поисковые сервисы в полном объеме (навигация по статьям тезауруса, квота на просмотр статей тезауруса);

Контроль IP адресов позволяет организовать специальный сервис для студентов, использующих УИС РОССИЯ в учебном процессе из учебных классов. С другой стороны - оперативно отключать тех, кто пытается получить неавторизованный доступ к ресурсам и т.п.

Возможность "закрывать" часть ресурсов используется персоналом УИС РОССИЯ для отладки интерфейса нового ресурса, а также для предоставления доступа к специфическим ресурсам клиентам и партнерам УИС РОССИЯ.

Квоты на просмотр документов гарантирует интересы правообладателей информации. Квоты на использование поисковых сервисов предназначены для оптимизации нагрузки на систему со стороны пользователей, пользующихся системой без регистрации (с сайтов, где УИС РОССИЯ используется только как поисковая машина), а также свободно регистрирующихся пользователей, которые не связаны условиями лицензионного соглашения.

Заключение

Таким образом в УИС РОССИЯ в основном решены задачи предоставления единообразного сервиса доступа к большой интегрированной базе данных, состоящей из разнородных коллекций полнотекстовых документов, гибкая настройка на новые виды ресурсов.

В процессе внедрения находятся следующие технологические решения:

- реализация "управляемого разнообразия" интерфейса доступа к документам различных коллекций, когда основное содержимое страницы документа порождается описанной в статье единой для всей базы процедурой, а дополнительное оформление специфических коллекций и подколлекций (ссылки, дополнительные сервисы) описываются с использованием механизма Java Server Pages;
- перенос процедуры аутентификации/идентификации пользователей на уровень документа;
- возможность выхода из защищенной части сайта для просмотра открытых документов, затем возврат в защищенную часть сайта без повторной

идентификации;

Нововведения призваны распределить нагрузку сопровождения специфических коллекций непосредственно на ответственных за конкретный ресурс, облегчить администрирование региональных зеркал УИС РОССИЯ.

Благодарность

Работа частично поддержана грантом РФФИ 01-07-930

Литература

1. Юдина Т.Н., Журавлев С.В., Российский межуниверситетский ресурсный и аналитический центр по гуманитарным исследованиям // Вестник РФФИ. - 1999. - N3. (intra.rfbr.ru/pub/vestnik/V3_99/2_8.htm).

2. Кричел Т., Ляпунов В.М., Паринов С.И., Онлайновые информационные ресурсы для исследователей по экономике: база данных RePEc и веб-портал RuPEc // Электронные библиотеки -1999 - Том 2 - Выпуск 3 (www.elbib.ru/journal/1999/199903/krichel/krichel.ru.html).

3. Лукашевич Н. В.,Салий А. Д., Представление знаний в системе автоматической обработки текстов // НТИ Серия 2. - 1997. - С.27-33. (www.viniti.ru/cgi-bin/nti/nti.pl?action=show&year=2_1997&issue=3&page=27)

4. Loukachevitch N., Dobrov B., Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems // Machine Translation Review - 2000 - 11: 10-20. (www.bcs.org.uk/siggroup/nalatran/mtreview/mtr-11/mtr-11-8.htm)

5. Callan J.P., Croft W.B. and Harding S.M., The INQUERY Retrieval System // A.M. Tjoa and I. Ramos (eds.), Database and Expert System Applications. Proceedings of {DEXA}-92, 3rd International Conference on Database and Expert Systems Applications. - Springer Verlag, New York. - 1992. - pp.78-93. (citeseer.nj.nec.com/26307.html)

6. Добров Б.В., Лукашевич Н.В., Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по Электронным Библиотекам "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - Петрозаводск. - 2001. - С.78-82. (rcdl2001.krc.karelia.ru/papers/papers/dobrov_lukashevich/dobrov_paper.rtf)

7. Добров Б.В., Лукашевич Н.В., Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту - Коломна. - 2002. - (в печати).

Об авторах

Агеев Михаил Сергеевич - аспирант механико-математического факультета МГУ им. М.В. Ломоносова.
E-mail: ageev@mail.cir.ru;

Добров Борис Викторович - канд. физ-мат. наук, снс НИВЦ МГУ им. М.В. Ломоносова.
E -mail:dobroff@mail.cir.ru;

Журавлев Сергей Васильевич - канд. физ-мат. наук, зав. лаб. анализа информационных ресурсов НИВЦ МГУ им. М.В. Ломоносова.
E -mail: zhuravlev@mail.cir.ru;

Лукашевич Наталья Валентиновна - канд. физ-мат. наук, снс НИВЦ МГУ им. М.В. Ломоносова.
E -mail:louk@mail.cir.ru;

Сидоров Алексей Валерьевич - сотрудник АНО Центр информационных исследований.
E -mail:alexeys@mail.cir.ru;

Юдина Татьяна Николаевна - канд. ист. наук, внс НИВЦ МГУ им. М.В. Ломоносова.
E -mail:yudina@mail.cir.ru
