Информационно-поисковая система «Манускрипт»: архитектурные и технологические решения[1]

А. Н. Миронов
Удмуртский государственный университет, Ижевск, Россия

The paper deals with the main architectural solutions that are the basis of the structure of the Manuscript system and provide access to the database by various means including the Internet. There are also considered the particularities of the data model allowing description of the arbitrary relationships between text units.

Информационно-поисковая система «Манускрипт» (далее — ИПС) предназначена для хранения и обработки текстов, имеющих сложную графико-орфографическую структуру. Основной ее особенностью является возможность совместного с текстами хранения и обработки дополнительной палеографической, лингвистической, текстологической, археографической и экстралингвистической информации. Хранение в системе структурированных данных о произвольных единицах текстов и связей между ними позволяет выполнять сложные многокритериальные выборки, обеспечивать разнообразную сортировку и представление результатов, получать необходимые перечни и указатели.

Ядром системы является база данных, построенная средствами Oracle SQL Server. База данных обеспечивает надежное хранение больших объемов данных, многопользовательскую работу, выполнение сложных запросов. Применение набора символов UNICODE позволяет хранить в базе необходимое разнообразие знаков, включая современные наборы символов и древнеславянские

[1] Работа по созданию ИПС «Манускрипт» ведется при поддержке Российского фонда фундаментальных исследований (грант № 05-07-90217в), работа по созданию автоматизированного морфологического анализатора — при поддержке Российского гуманитарного научного фонда (грант № 05-04-12408в).

(кириллические), с расширением набора символов глаголическим диапазоном в перспективе. С этой единой базой данных взаимодействуют различные клиентские модули, работающие в непосредственном (on-line) соединении с базой. Информационная модель, лежащая в основе базы данных, обеспечивает хранение и описание любых единиц, составляющих текст (знаков, словоформ, предложений-фраз, фрагментов текста), и связей между ними. Изначально заложенные в модель гибкость и возможность добавления единиц новых типов позволяют непрерывно развивать систему, расширяя возможности описания текстов и составляющих его единиц. Другой стороной этой гибкости является существенное усложнение алгоритмов обработки данных и снижение быстродействия системы. Поддержка моделью связей «много-во-много» между единицами текста дает возможность описания любых лингвистических объектов, однако обработка таких связей в каждом отдельном случае требует достаточно больших трудозатрат. Выделение в сетевой структуре данных отдельных подсетей иерархической структуры позволяет применять для обработки данных менее сложные, универсальные алгоритмы. Окружение данных несколькими оболочками прикладных программных интерфейсов дает возможность вести разработку клиентских модулей, используя объекты данных, соответствующие предметной области, и разделить между исполнителями работы, связанные с манипуляцией данными и с организацией пользовательских интерфейсов.

Проектирование информационной и табличной моделей осуществляется средствами Oracle Designer, обеспечивающим ведение необходимой документации и существенно упрощающим сопровождение и выполнение необходимых доработок табличной модели системы.

Использование нескольких различных специализированных клиентских модулей дает возможность при их разработке сконцентрироваться на создании удобных пользовательских интерфейсов, позволяющих достаточно простыми визуальными средствами и приемами управлять сложными процессами обработки данных, выполняя типовые запросы к базе данных и получая типизированные выборки. Использование протоколов IP и HTTP обеспечивает доступ к данным единой базы с любых компьютеров, подключенных к сети Интернет. Применение сервера приложений (Oracle

application server) дает возможность наряду с достаточно ограниченными интерфейсными возможностями WEB-браузера использовать богатый набор интерфейсных средств Oracle Forms для разработки модуля выборок и запросов и служебных модулей системы.

Разработка специализированного модуля обмена данными и стандартов обмена на базе XML-TEI должна обеспечить возможность обмена данными с приложениями, не использующими непосредственного соединения с базой данных «Манускрипт».

Список литературы

Баранов и др. 2003 — *Баранов, В. А.* Электронные издания древних письменных памятников и технология создания полнотекстовых баз данных / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, О. В. Зуга, А. Н. Миронов [и др.] // Круг идей: электронные ресурсы исторической информатики : тр. VIII конф. Ассоциации «История и компьютер» / под ред. Л. И. Бородкина, В. Н. Владимирова. — М. ; Барнаул : Изд-во Алт. ун-та, 2003. — С. 234–270.

Баранов и др. 2004 — *Баранов, В. А.* Информационно-поисковая система «Манускрипт»: новые возможности электронного издания древнерусских рукописей / В. А. Баранов, А. А. Вотинцев, Р. М. Гнутиков, О. В. Зуга, А. Н. Миронов [и др.] // Информационный бюллетень Ассоциации «История и компьютер». № 32 : Материалы IX конф. АИК (апр. 2004 г.) — Москва ; Томск : Изд-во Том. ун-та, 2004. — С. 7–9.

Baranov et al. 2004 — Victor Baranov, Andrey Votintsev, Roman Gnutikov, Aleksey Mironov, Sergey Oshchepkov, Vitaliy Romanenko. Old Slavic Manuscript Heritage: Electronic Publications and Full-Text Databases // EVA 2004 London (Electronic Imaging, the Visual Arts Conference & Beyond). Conference Proceedings. — University College London. Institute of Archaeology. Principal Editor: James Hemsley. — London, 2004. — 11.1–11.8.

Об одном методе статистической фильтрации текстовой информации

С. В. Моченов, А. М. Бледнов, Ю. А. Луговских
Ижевский государственный технический университет, Россия

The paper presents a review of the statistic methods of analysis of texts in natural languages and shows the possibilities of filtration of the text information on the basis of one of the most informative statistic text characteristics. There is developed a method of filtration of the texts in the Russian language that can be applied to the solution of various tasks of text processing like the decrease of the volume of textual information, writing essays, determination of the semantic component of the text units etc.

Введение

Проблеме анализа и синтеза текстовых документов посвящено значительное количество работ [Математические 1977; Фоменко 1980; Носовский и др. 1989]. Среди них значительное место занимают работы, связанные с задачами автоматического анализа полнотекстовых документов, автоматической классификацией и идентификацией тем документов, автоматическим реферированием, выявлением смысловых связей и др.

Статистические методы достаточно хорошо зарекомендовали себя при построении поисковых систем, выделении ключевых слов и словосочетаний и т. п.

В то же время при решении задач анализа и синтеза текстовой информации, возникающих при построении информационных систем, в частности при формировании профессиональных баз знаний, требуется привлечение алгоритмически более сложных процедур проведения синтаксического и семантического анализа.

В данной статье проводится обзор статистических методов анализа текстов на естественных языках, а также показаны возможности фильтрации текстовой информации на основе одной из наиболее информативных статистических характеристик текста.