С.М. Козиенко, С.А. Яблонский

ИНТЕРНЕТ/ИНТРАНЕТ СИСТЕМА СОЗДАНИЯ И ВЕДЕНИЯ КОРПУСОВ ТЕКСТОВ С РАЗВИТЫМИ ЛИНГВИСТИЧЕСКИМИ СРЕДСТВАМИ НА ОСНОВЕ J2EE И ORACLE ТЕХНОЛОГИЙ

За последнее десятилетие в сети Интернет и на CD появились большие коллекции различных текстов на русском языке: художественных, научных, правовых и других, объединенные в электронные библиотеки. Большинство электронных библиотек отличает крайняя бедность средств работы с опубликованными текстами, что не дает в полной мере использовать их в качестве корпусов текстов как для научной работы, так и в целях образования и ознакомления.

Сегодня, обычно выделяют следующие два типа корпусов:

- статический (классический моно- или многоязычный аннотированный корпус);
- динамический (Web как многоязычный частично параллельный корпус).

Для работы с корпусами текстов существует ряд достаточно популярных программных сред: LT-XML (Edinburgh), GATE (Sheffield), IMS Corpus Workbench (Stuttgart) и пр. Однако большая часть из них не ориентирована на работу в Интернете/-Интранете с текстами объемом более 10–50 млн слов, не допускается многоязычие и отсутствуют развитые средства лингвистической поддержки систем поиска и индексирования текстов, особенно для русского языка.

В настоящей работе рассматривается система для создания и ведения корпусов тестов на любом языке в Интернете/Интранете. Разработаны средства создания, ведения и обработки простых и аннотированных моно- и многоязычных корпусов текстов большой размерности (до сотен миллионов слов). В более широком смысле система может рассматриваться как основа любой поиско-

вой системы в Интернете/Интранете и систем документооборота.

Система состоит из подсистем[1] (рис. 1):

- администратора,
- индексирования и поиска документов,
- тезауруса/рубрикатора, конкорданса и толкового словаря,
- морфологического анализатора и лемматизатора,
- обработки новых слов,
- подсистемы интерфейса.

Рассмотрены четыре основных типа представления XML/ SGML размеченных текстов корпуса в базе данных системы[2]:

- классическое реляционное табличное представление;
- объектно-реляционное (object-relational/object-based) представление;
- прямое XML представление (рис. 1);
- гибридное XML представление, сочетающее первый и третий типы;
- используются платформно-независимые технологии J2EE (Servlets и Java Server Pages) и Oracle9i/10g;
- используется встроенное в СУБД Oracle средство работы с текстовыми данными Oracle Text, что позволяет применять:

[1] *Yablonsky S.A.* Corpora as Object-Oriented System: From UML-notation to Implementation // Proceedings of the Third International Conference on Language Resources & Evaluation. Las Palmas, Canary Islands, Spain, 2002; *Yablonsky S.A.* The Corpora Management System Based on Java and Oracle Technologies // Proceedings of the 10th Conference of the European Chapter of the Association for the Computational Linguistics. Budapest, Hungary, 2003.

[2] Ibid.

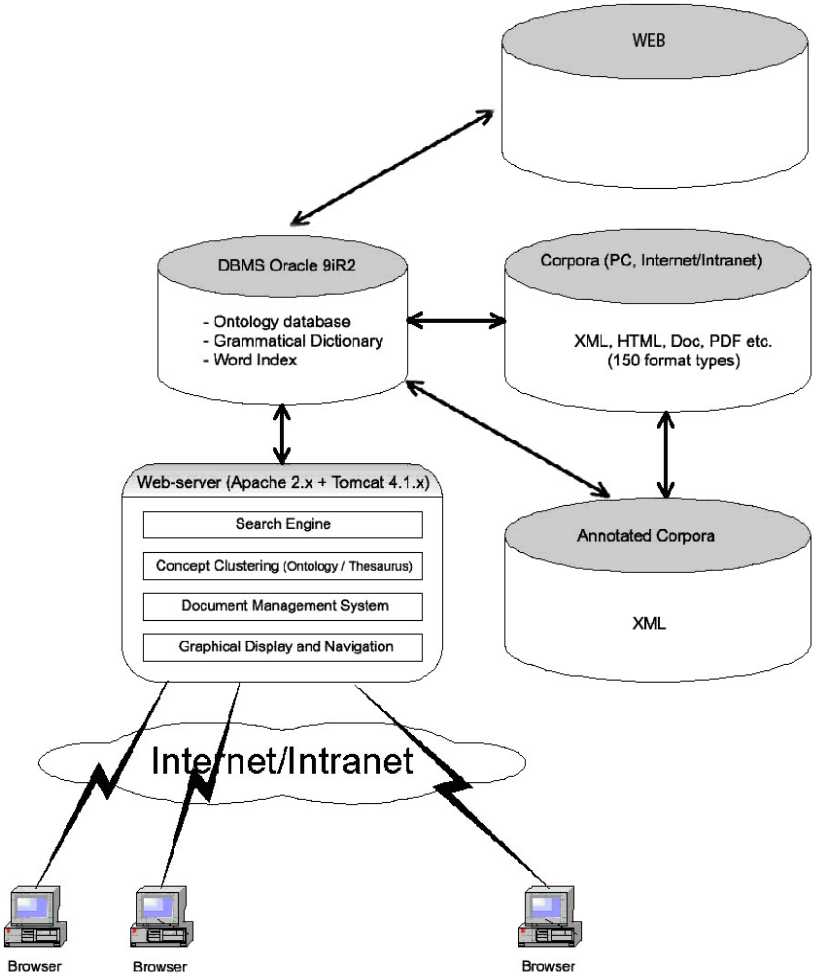

Рис. 1. Структура системы

- встроенные лингвистические средства работы с текстами ряда европейских языков (распознавание числа, времени и других типов слов, а также грамматических форм, неверно написанных слов и слов, сходных по звучанию);
- 150 конвертеров наиболее широко распространенных форматов текстовых файлов (ASCII, doc, rtf, pdf и пр.);
- кроме точного поиска по слову или словосочетанию с выполнением известных булевых операций, также поиск с упорядочиванием по релевантности и заданием списков стоп-слов;
- встроенные и создаваемые пользователями простейшие типы тезаурусов для поиска синонимов и тематически близких слов;
- анализ содержания документа корпуса и автоматическое выделение его ключевых тем с созданием тематического резюме.

- используются апробированные лингвистические ресурсы и программы ЗАО «Руссикон»[1], что позволяет использовать средства OracleText и для русскоязычных текстов (поскольку сам OracleText не поддерживает русский язык);
- реализован полнотекстовый поиск для русскоязычных текстов с учетом морфологии русского языка[2];
- использование нормализатора позволяет создавать компактный индекс для русскоязычных текстов за счет хранения лемм (а не словоформ).

---

[1] *Yablonsky S.A.* Russicon Slavonic Language Resources and Software // Proceedings of the First International Conference on Language Resources and Evaluation / A. Rubio, N. Gallardo, R. Castro, A. Tejada (eds.). Granada, Spain, 1998.

[2] *Yablonsky S.A.* Russian Morphological Analyses // Proceedings of the International Conference VEXTAL, November 22–24, 1999, Venezia, Italia. P. 83–90.

Применяемые решения позволяют работать как с корпусом русских текстов, так и многоязычными корпусами тестов за счет использования Unicode (очевидно, что для каждого языка должен быть включен свой набор лингвистических ресурсов).

Простейший вариант подобной системы реализован в европейском многоязычном портале переводов литературных произведений **www.babelmatrix.org** (см. рис. 3).

Рис. 3. Портал **www.babelmatrix.org**