The Corpora Management System Based on Java and Oracle Technologies

Serge Yablonsky
Petersburg Transport University, Computer Department, Moscow av., 9,
St.-Petersburg, 190031, Russia
Russicon Company, Kazanskaya str., 56, ap.2, St.-Petersburg, 190000, Russia
serge_yablonsky@hotmail.com; root@russicon.spb.su;
http://www.russicon.ru

Abstract

The paper discusses the corpora management system (CMS) design that uses Java and Oracle9i DBMS to support strategic corpora analysis. We present the pilot web-based CMS to support linguists in their daily work. The system offers facilities to assist linguists and internet users as they search for relevant material, and then classify and annotate this material.

1 Introduction

There's a wide class of documental management solutions and products that fall under the rubric "corpora and text mining". They are similar to data mining solutions in that they deal with large volumes of data, but the difference between the two technology solutions is that while data mining extracts, analyzes, and summarizes numerical, structured data, text mining handles large volumes of unstructured, text-based data. Document systems with large-scale linguistic annotation are used by a wide range of research and commercial applications.

This paper presents a web-based text corpora development system (CMS) that focuses on the development of UML-specifications, architecture and actual implementations of DBMS tools to support strategic corpora analysis.

We present the basic features of a prototype corpora management system under development intended to support linguists in their daily work. The system offers facilities to assist linguists and internet users as they search for relevant material, and then classify and annotate this material in a repository.

The CMS is implemented using Java and commercial DBMS Oracle9i.

2 System Overview

The Corpora management system combines Java, XML, XSL, HTML, and Oracle9i components (Yablonsky S.A., 2002). The system was by adapting existing and new DBMS and Java tools to the necessities of the intended task for the Russian language (Yablonsky S.A., 2000).

CMS consists of such main parts (see figure 1):

- Corpora – files in more then 150 different formats (doc, rtf, pdf, htm, xml, etc.);
- Annotated corpora – files in XML format using XML Corpus Encoding Standard (XCES, http://www.xml-ces.org – Ide, N. & Brew, C., 2000, Ide, N., Romary L., 2001) and text formats;
- Oracle 9i DBMS Enterprise Sever (Release 2) with such main counterparts:
    - *Grammatical dictionaries* (inflection paradigms of the given language) for the languages that are not supported by Oracle Text;
    - *Ontologies / WordNet* (Fellbaum C.) / *Domain Thesauruses* for given languages;
    - *Word Index* (index of all entry words or lemmas of Corpora);

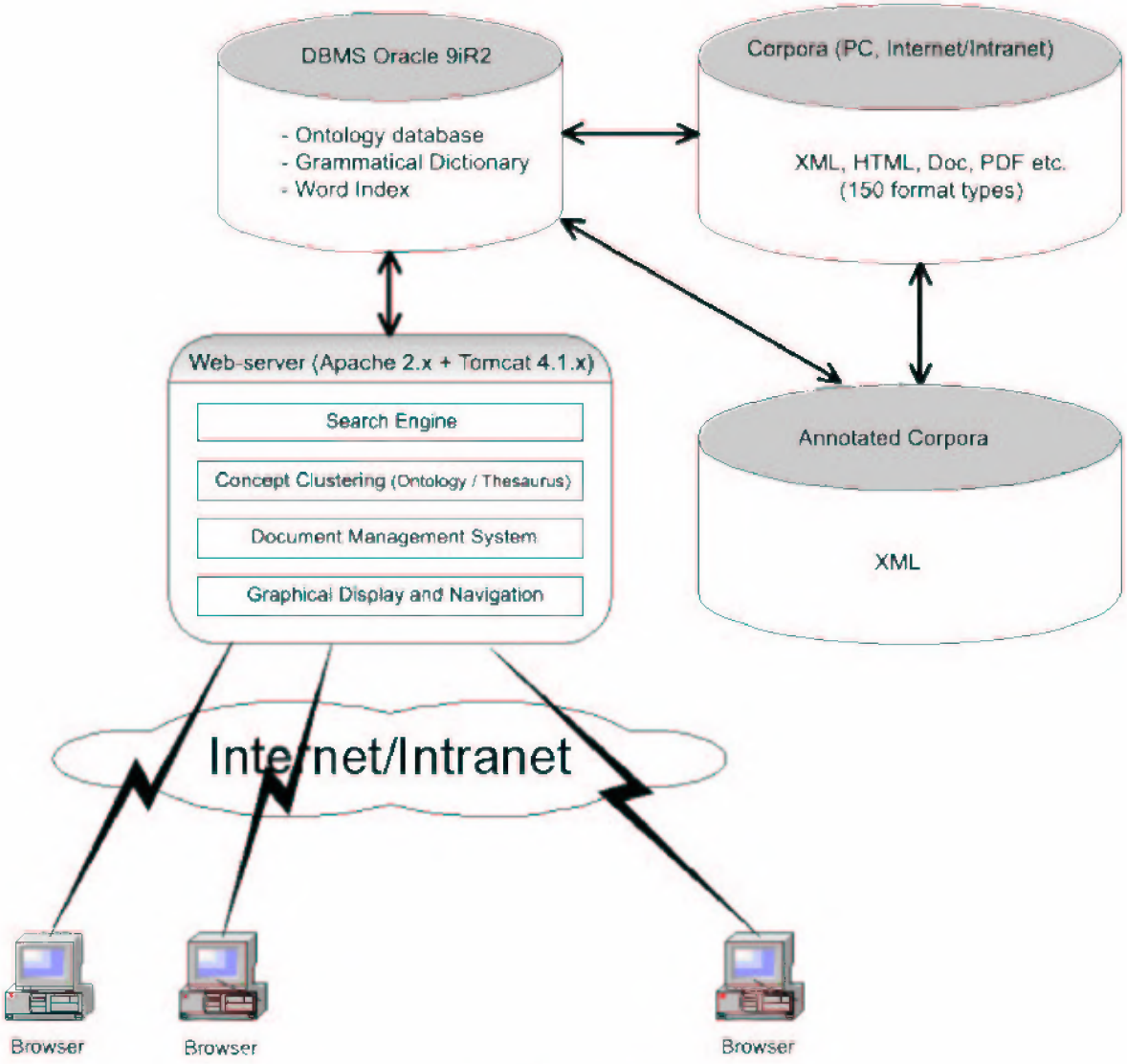

Figure 1. CMS structure

- *Oracle Text.*

- Web Server (for example, Apache 2.x plus Tomcat 4.1.x) with such Java Server Pages (JSP) counterparts:
  - *Search Engine;*
  - *Concept Clustering;*
  - *Document Management System;*
  - *Interface Subsystem.*

We present the fragment of CMS UML-specification built by Rational Rose (Rational Rose Enterprise Edition Documentation, 2001) that could be expanded in future by community of language and speech resources developers (see figure 2).

Here, for example, the relational table DOCS contains such attributes:

- DOCS_NAME – document name;
- DOCS_AUTHOR – document author;
- DOCS_HTTP – document path;
- DOCS_LANG – document language;
- DOCS_LANG_CODE – document coding;
- DOCS_DATE – date of including in the Corpora;
- DOCS_EXTENSION – document file extension.

The set of different types of UML-specifications brings us to full three-level system, including user, business and data services.

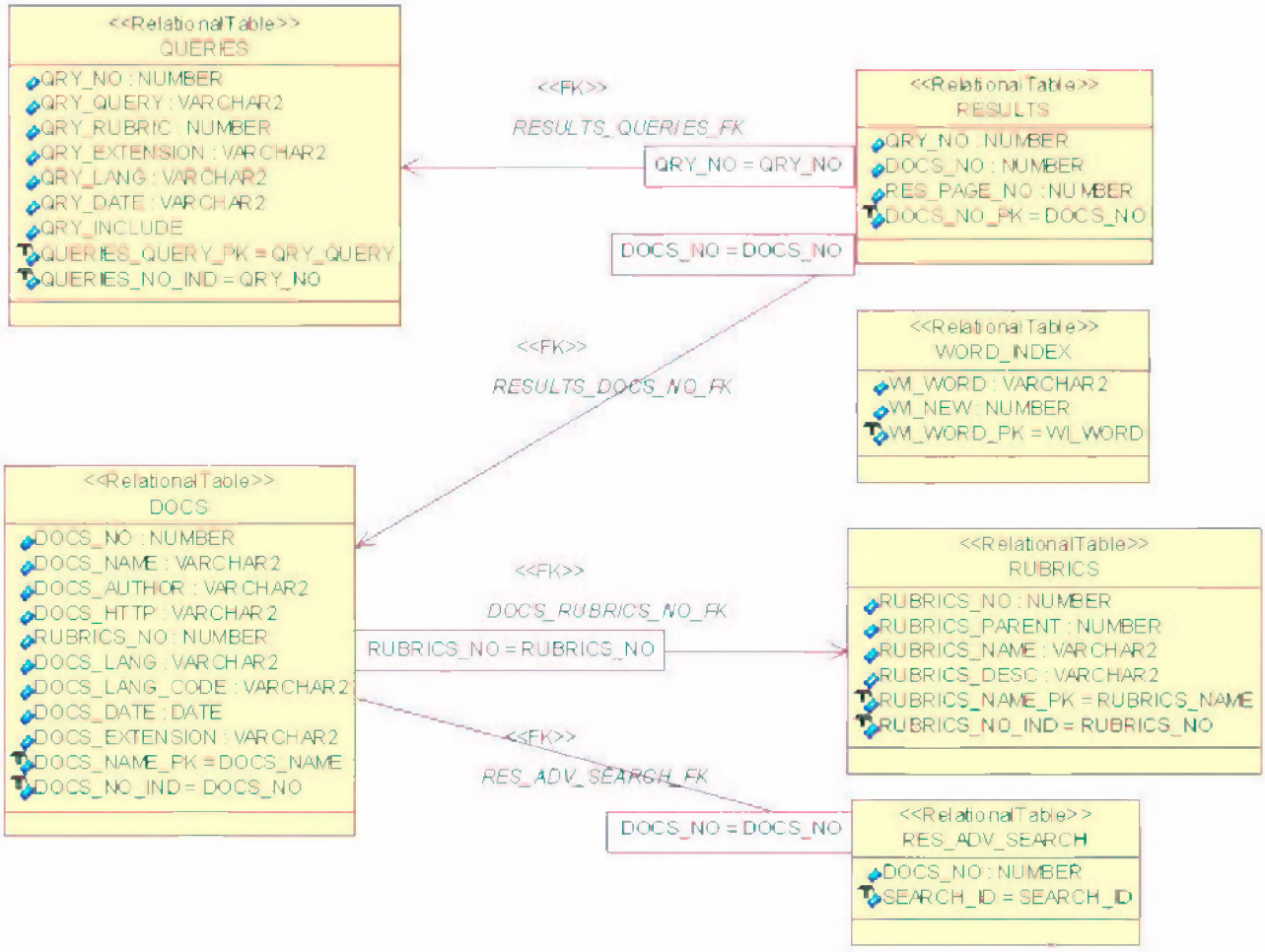

Figure 2. Fragment of UML notation of data model

UML specification of business services uses standard UML notations of standard linguistic annotation and corpora manipulation procedures. Reusability of linguistic and corpora manipulation business services could be achieved by usage of a widely accepted set of UML notation standards for corpus-based work in natural language processing applications.

3 System Features

The powerful search engine of prototype system particular uses advantages of Oracle Text search and services and includes (Oracle9i Database Documentation):

- *Content-based retrieval* on free text with both literal (word) predicates and thematic predicates. It includes: a comprehensive range of operators and index preferences (e.g. Boolean, exact phrase match, proximity, section searching, fuzzy, stemming, wildcard, thesaurus, stopwords, case sensitivity, and search scoring), "about" search, structured search, broad document format support and multi-language support. For example, texts can be searched for stems of words e.g "teach" would return "teaching", "taught" etc. Fuzzy match can be used if you are not sure of the spelling of a word. A search can be done to find words which are close to each other within a word. Documents can be searched on what a document is about as opposed to the existence of specific words. Gists or Theme Summaries can be produced which produce a summary of what a document is about (using themes).

- For XML framework XPath searching enables sophisticated queries which can reference and leverage the embedded structure of XML documents – instead of using a text query to find documents, you use a document to find queries. XML path searching is able to perform sophisticated section searches: doctype disambiguation, attribute value searching, automatic section indexing, and more.

In addition to the search capabilities, a number of other features are provided to simplify application development.

- *Corpora Format Support.* In order to index documents stored in a variety of native formats, such as Word, Excel, PowerPoint, WordPerfect, HTML, and Acrobat/PDF, system supplies a broad variety of "filters" that allow documents stored in their native formats to be indexed. Support for more then 150 file formats in order to index files in a large range of formats including Word, Acrobat, HTML, WordPerfect, Powerpoint, Excel Flexible Storage Location - documents can be stored and indexed in the database, in a location pointed to by a URL or in an external file.
- *Corpora Graphical Display and Navigation.* System services can convert any supported document format to either plain text or formatted text (an HTML approximation retaining as much as possible of the original formatting; available for all formats except PDF). Both plain text and HTML versions may be viewed in a standard browser.
- *Document Management System.* System supplies an administration tool through which all major text maintenance and administration functions may be performed.
- *Concept Clustering* identifies the relationships between phrases of the texts and it builds a "lexical network," grouping related phrases and enhancing the most important features of these groupings. The resulting patterns reveal the conceptual backbone of the text collection. Russian WordNet is used for basic conceptual grouping.
- *Automatic Corpora Text* collection, tokenization, part-of-speech tagging. Reusability of linguistic resources is achieved by annotation of texts using a common data model. For that purpose the XML and related standards such as XML Corpus Encoding Standard (XCES, http://www.xml-ces.org) (Ide, *et al.,* 2000) are used in the system.

CNS is localized for Russian language. Oracle 9i Text doesn't have Russian language support. In order to use Oracle Text capabilities we add *Russian Grammatical Dictionary* (morphosyntactic dictionary) that consists of word paradigms with grammatical characteristics of all Corpora word index (Yablonsky S.A., 1998). It helps to perform linguistic (paradigm) search in Corpora and is also used for Corpora text annotation.

System could be easily adapted to different software platforms (Java and Oracle) and the necessities of other languages (Unicode), making the whole system portable to other platforms with minimal changes.

The only language-specific resources are a large-scale morphosyntactic dictionary plus POS tagger.

References

Ide, N., Romary L., 2001. XML Support for Annotated Language Resources. In: *Linguistic Exploration,* Workshop on Web-Based Language Documentation and Description, Dec 12 - Dec 15, 2000, University of Pennsylvania Philadelphia, Pennsylvania, USA.

Ide, N. & Brew, C., 2000. Requirements, Tools, and Architectures for Annotated Corpora. In: *Proceedings of the EAGLES/ISLE Workshop on Meta-Descriptions and Annotation Schemas for Multimodal/ MultimediaLanguage Resources and Data Architectures and Software Support for Large Corpora.* Paris: European Language Resources Association.

Oracle9i Database Documentation (Release 9.0.2), 2002.

Rational Rose Enterprise Edition 2001, Documentation.

Fellbaum C. (ed.). WordNet. An Electronic Lexical Database. Bradford Books.

Yablonsky S.A., 1998. Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo,

Yablonsky S.A., 2000. Russian Monitor Corpora: Composition, Linguistic Encoding and Internet Publication. Proceedings Second International Conference on Language Resources & Evaluation, Athens, Greece, 2000.

Yablonsky S.A., 2002. Corpora as Object-Oriented System. From UML-notation to Implementation. Proceedings Third International Conference on Language Resources & Evaluation, Las Palmas, Canary Islands-Spain, 2002.