Chapter XCII
Full-Text Manipulation in Databases

László Kovács
University of Miskolc, Hungary

Domonkos Tikk
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The textual data format is one of the most important data types in database management. Databases support a wide range of special textual types that can be used to store string data. In the case of textual data, information retrieval mostly concerns the selection and the ranking of documents. In the traditional database management systems (DBMS), text manipulation is related to the usual string manipulation facilities, i.e. the exact matching of substrings. The main disadvantages of the traditional string-level operations are the high cost as they work without task-oriented index structures and the restriction to the syntax level.

The required appropriate full-text management operations belong to text mining, an interdisciplinary field of natural language processing and data mining. As the traditional DBMS engine is inefficient for these operations, database management systems are usually extended with a special full-text search (FTS) engine module. Full-text search engines provide a set of full text manipulation primitives that are based on the semantic aspects of the words and sentences. In

the recent years, the area of semantic query on full-text is treated as a special area of universal ontology. The main goal of ontology is to define a common set of concepts and relationships for knowledge representation.

The market of FTS engines is very promising because the amount of textual information stored in databases rises steadily. According to the study of Meryll Lynch (Blumberg & Arte, 2003), 85% of business information are text documents – e-mails, business and research reports, memos, presentations, advertisements, news, etc. – and their proportion still increases. In 2006, there were more than 20 billion documents available on the Internet (Chang, 2006). The estimated size of the pool reaches 550 billion documents when the documents of the hidden (or deep) web are also considered.

In a broader sense, the area of full-text manipulation is not restricted only to the management of database content. The query interface may be extended with a natural language interface for databases (NLIDB). The NLIDB means that a user can use natural languages to create query expressions, and also the answer can be presented in the same languages. The processing of the incoming queries includes a full-text analysis both at syntactic and semantic levels.

BACKGROUND

The subfield of document management that aims at processing, searching, and analyzing text documents is *text mining*. Text mining is an application oriented interdisciplinary field of machine learning which exploits tools and resources from computational linguistics, natural language processing, information retrieval, and data mining. The general application schema of text mining is depicted in Figure 1 (Fan, Wallace, Rich & Zhang, 2006). For giving a brief summary of text mining, four main areas are presented here: information extraction, text categorization/classification, document clustering, and summarization.

The goal of information extraction (IE) is to collect the text fragments (facts, places, people, etc.) from documents relevant to the given application. The IE module includes among others the following subtasks: named entity recognition (Sibanda & Uzuner, 2006); co-reference resolution (Ponzetto & Strube, 2006); identification of roles and their relations (Ruppenhofer et al, 2006).

Text categorization (TC) techniques aim at sorting documents into a given category system (see Sebastiani, 2002 for a good survey). Typical application examples of TC include among many others: document filtering (Lewis, 1995); patent

Figure 1. The text mining module