Chapter IV

Text Warehousing: Present and Future

Antonio Badia, University of Louisville, USA

Abstract

Data warehouses, already established as the main repository of data in the enterprise, are now being used to store documents (e-mails, manuals, reports, etc.) so as to capture more domain information. In order to integrate information in natural language (so-called unstructured data) with information in the database (structured and semistructured data), existing techniques from Information Retrieval are being used. In this chapter, which is part overview and part position paper, we review these techniques and discuss their limitations. We argue that true integration cannot be achieved within the framework of Information Retrieval and introduce another paradigm, based in Information Extraction. We discuss the main characteristics of Information Extraction and analyze the challenges that stand on the way of this technology being widely used. Finally, we close with some considerations on future developments in the general area of documents in databases.
Introduction

Data warehouses (DWs) are already established as an important tool for understanding and improving business practices in organizations. Usually, DWs feed from smaller, transaction-oriented databases that an organization already has in place to take care of day-to-day business. Lately, however, it has been realized that this information by itself may not be enough to support decision-taking. Additional sources of information are then sought. There are two main types of sources to achieve data enrichment: external and internal. External data sources bring in information about the environment of the organization (state of the market, competitors, overall economical and technological trends), while internal sources bring additional insight about the workings of the organization. Often, this additional information (both external and internal) comes from documents. Consequently, interest in managing text, especially for applications within business intelligence (decision support, CRM) has grown enormously in recent years, due to the realization that a great deal of business information resides in documents like e-mail messages, memos, internal manuals, sales reports, etc. This is the basic reason to bring text into the DW: The practice of enriching data could be substantially improved. Furthermore, documents usually contain not only information about what happens, but about why things happen.

Database vendors have responded to this trend by adding to their flagship database systems modules for text management (Oracle Text in Oracle 9i — to be called Oracle Multimedia in 10g — and Text Extender in DB2 v8). These modules are based on Information Retrieval (IR) techniques, essentially creating an inverted index for a collection of documents and allowing SQL to query this index. This approach has been sanctioned by the latest SQL standard, which includes basic text search capabilities in the WHERE clause of SQL queries. However, IR approaches have clear limitations that restrict their usefulness for decision-support.

In this chapter, we overview the integration of datasets and text as is presently implemented with the use of IR technology. We then show the shortcomings of this approach, and argue that true integration of information can only happen if the information in texts is analyzed at a deeper level than IR allows. Information Extraction (IE) is a new technology that addresses this issue by doing a focused, basic analysis of document contents. While IE may be the right technology to substitute IR, its widespread use depends on the solution of some challenging technical issues which we present and discuss.

Note that we consider only the case in which the DW contains both structured or semistructured information besides documents; the case of a DW with documents alone (a document or text database) is also interesting, but we argue
Related Content

Notifiable Disease Databases for Client Management and Surveillance
www.irma-international.org/chapter/notifiable-disease-databases-for-client-management-and-surveillance/186949/

Understanding Actors in Complex Security Problems
www.irma-international.org/article/understanding-actors-in-complex-security-problems/203634/

Self-Fulfilling Prophecy in e-Negotiations: Myth or Reality?
www.irma-international.org/article/self-fulfilling-prophecy-negotiations/69514/

A Method for Systematic Artifact Selection Decision Making
www.irma-international.org/chapter/method-systematic-artifact-selection-decision/11302/

Technology Advancement and E-Procurement in the US
www.irma-international.org/chapter/technology-advancement-procurement/72644/