Chapter VII

A Textual Warehouse Approach: A Web Data Repository

Kaïs Khrouf, University of Toulouse III, France
Chantal Soulé-Dupuy, University of Toulouse III, France

ABSTRACT

An enterprise memory must be able to be used as a basis for the processes of scientific or technical developments. Indeed, it was proven that information useful to these processes is not found solely in the operational bases of companies; it is also found in textual information and exchanged documents. For that reason, we propose the design and implementation of a documentary memory for business document warehouses. Its main characteristic is to allow the storage, retrieval, interrogation and analysis of information extracted from disseminated sources and, in particular, from the Web.
INTRODUCTION

An enterprise must allow for the sharing of knowledge and information between its employees in order to optimize their tasks. However, the volume of information contained in documents represents a major concern for these companies. Indeed, companies must be fully reactive to any new information and must follow the fast evolution and spread of information. So, a business memory which stores this information and allows end-users to access or analyze it is necessary for every enterprise.

This memory aims to:

- merge information from several sources, such as the World Wide Web, intranets, etc.;
- take the information evolution into account;
- allow end-users to view and analyze information according to their needs;
- facilitate decision-making.

These objectives can be reached by using the concept of textual warehouses, which allows the storage of documents and their exploitation through the techniques of information retrieval, factual data interrogation, and multidimensional analysis of information.

This chapter is organized as follows. First, we outline some work devoted to document querying through information retrieval or database techniques. Then, we propose an architecture and a generic model of textual warehouses. The next section describes the information extraction to feed the warehouse. Finally, we present the techniques we propose to exploit information contained in the warehouse. We describe the information retrieval process and the multidimensional analyses.

BACKGROUND

IRS (Information Retrieval Systems) were initially introduced to exploit non-structured documents, i.e., documents which contain no information about their logical structure. These documents were analyzed to represent their textual content and, therefore, their relevance in response to a non-structured query (free natural language). During the last 20 years, several theoretical models were proposed, and several systems based on those models have been implemented. The most well-known of these systems are: the Boolean model [STAIRS (IBM, 1982)], the vector-space model [SMART (Salton, 1971)],
Ranking Algorithm for Semantic Document Annotations
www.igi-global.com/article/ranking-algorithm-semantic-document-annotations/72703?camid=4v1a