Enhanced Knowledge Warehouse

Krzysztof Węcel
The Poznan University of Economics, Poland

Witold Abramowicz
The Poznan University of Economics, Poland

Pawel Jan Kalczynski
University of Toledo, USA

INTRODUCTION

Enhanced knowledge warehouse (eKW) is an extension of the enhanced data warehouse (eDW) system (Abramowicz, 2002). eKW is a Web services-based system that allows the automatic filtering of information from the Web to the data warehouse and automatic retrieval through the data warehouse. Web services technology extends eKW beyond the organization. It makes the system open and allows utilization of external software components, thus enabling the creation of distributed applications.

The enhanced knowledge warehouse is an add-in to the existing data warehouse solutions that offers a possibility of extending legacy data-based information resources with processed information acquired from Web sources of business information.

eKW is characterized by transparent filtering and transparent retrieval. The system automatically acquires interesting documents from Web and internal sources and enables business users, who view data warehouse reports, to access these documents without the necessity of formulating any document retrieval queries. Instead, the system takes advantages of meta-information stored in business metadata and data warehouse reports, and builds and runs appropriate queries itself. The system also has an event-alerting capability.

eKW, by design, is a part of the semantic Web. The main role of eKW in the semantic Web is mediation between the world (external information) and internal information systems utilized within the organization. One of the most popular ways of conforming to semantics is to use ontologies, which provide a means for defining the concepts of the exchanged data. eKW is a kind of data mediator that employs ontologies as a conceptualization layer (W'cel, 2003).

BACKGROUND

Document management systems may be divided into two major categories: information retrieval (IR) and information filtering (IF) systems.

The classical models of information retrieval systems were defined by Salton (1983). As distinct from structured data management, information retrieval is imprecise and incomplete. This is due to the inaccurate representations of document contents and user information needs. The distinctive feature of information retrieval systems is a relatively constant document collection. The collection stores documents and their representations (indices). When a user submits a query to the information retrieval system, the query is compared against all indices in the collection. Documents whose indices match the query are returned as the resulting subset of the collection. Most search engines on the Web and digital libraries available on the market can serve as examples of information retrieval systems (Baeza-Yates, 1999).

The main objective of information filtering systems is to block unwanted information from the incoming stream of documents. As distinct from IR systems, filters do not have any fixed collection of documents. Instead, they hold a collection of standing queries (profiles). The profiles represent long-term information interests of their owners. When a new document arrives at the filter its representation is created and compared against all profiles in the collection. Some libraries maintain systems that inform users about new volumes included in the library. Such systems are usually based on the server-side filtering architectures and their profiles usually consist of semi-structured elements (e.g., author, subject, title). This idea is referred to as selective dissemination of information (SDI).

JUSTIFICATION

Possessing and maintaining large information resources does not itself provide any guarantees that users will manage to find the piece of information (document) they need. First of all, users must be aware that the particular piece of information already exists in their resources. Yet, even if they are aware, searching may be very labor consuming. Such a situation is highly undesirable for organizations, as they do not exploit their potentials to increase the effects. That is why knowledge management is nowadays considered the capability of re-use of information and is becoming an increasingly vital issue.

Usually, users are not eager to learn how to operate in several different information systems. Integration of information resources enables the exploitation of the capabilities of many systems through a single user interface (UI). Data warehouse users should be capable of finding interesting documents through a single UI. Not only should the users be aware that the desired information does exist, but also relevant documents should be disseminated to them mechanically. Such a mechanical distribution of the relevant information gives more time for other tasks.

Business users have certain problems with finding information sources, building correct retrieval queries and formulating proper filtering profiles. Thus, introducing the system capable of relieving the users from the necessity of formulating queries and solving the most commonly reported problems with accessing information on the Web would increase productivity of those seeking relevant external information.

The constantly growing number of content providers and the exploding volume of business information are not accompanied by the corresponding growth of capabilities of exploitation of the resources by the contemporary organizations. Therefore, automatic acquisition, organization and presentation of information became not only possible but also essential for the performance of today's businesses.

THE IDEA OF eKW

The term *data warehouse* was defined by Bill Inmon in 1992 as *a collection of integrated, subject-oriented databases to support the DSS function, where each unit of data is relevant to some moment in time*. Since then a rapid growth of this relatively new idea has been observed.

In the eKW model, Web documents are assumed to be business news published in English by major content providers on the Web (e.g., Reuters, Cnn.com), because this type of external information is the one in favor of business people (Abramowicz, 2000; Webber, 1999). The solution proposed is based on the existing standards in the area of data warehousing, document management, communication among software agents, internetworking and storage.

The basic idea of data warehousing is to create a data model common for the whole organization (metamodel). The model provides a framework for uploading legacy data, stored in heterogeneous databases across the organization, to the warehouse. Before data are uploaded they need transformation that includes filtering, consistency check, field mapping and aggregation.

Data warehouses proved to be useful for the purpose of legacy data analysis and they are commonly implemented by organizations. They provided novel data processing techniques like data mining, basket analysis, dimensional data analysis, drill-down, drill-through or slice-and-dice. These techniques are based on the characteristic features of data and information stored in the data warehouse: non-volatility, relevance to some moment in time (timeliness), correctness, consistency, completeness, business subject orientation, and business descriptions in metadata (Adelman, 1997; Kimball, 1996).

In terms of the data warehouse, metadata1 (warehouse metadata) are data about data. Metadata consist of facts and events concerning database objects. Metadata are usually stored in a database referred to as the metadata repository or the metadata collection. Nowadays, metadata are considered to be an integral part of the data warehouse (Gleason, 1997; Inmon, 1999).

Due to its destination and the sophisticated methods of data processing, the enterprise data warehouse is often claimed to be the corporate knowledge repository. However, we argue that information derived only from structured information (data) produced by the organization is just a fraction of corporate knowledge that may be stored in the repository.

As distinct from database systems, document management systems deal with unstructured and semi-structured information. Because machines are not capable of understanding text, they must rely on mathematical representations of the document content and user information needs.

The most distinctive feature of eKW is that information acquisition, organization and dissemination are performed without any additional actions taken by data warehouse users. In particular, no formulation of keyword-based queries is necessary to access external business news relevant to user information needs. Instead, the queries are formulated mechanically based on business metadata and information stored in the enterprise