Available online at www.sciencedirect.com

ScienceDirect

Decision Support Systems 42 (2006) 727–744

Decision Support Systems

www.elsevier.com/locate/dsw

The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence[☆]

Frank S.C. Tseng [a,*], Annie Y.H. Chou [b,1]

[a]Department of Information Management, National Kaohsiung First University of Science and Technology, 1 University Road, YenChao, Kaohsiung, Taiwan 824, ROC

[b]Department of Computer and Information Science, Chinese Military Academy, Taiwan

Available online 17 June 2005

Abstract

During the past decade, data warehousing has been widely adopted in the business community. It provides multi-dimensional analyses on cumulated historical business data for helping contemporary administrative decision-making. Nevertheless, it is believed that only about 20% information can be extracted from data warehouses concerning numeric data only, the other 80% information is hidden in non-numeric data or even in documents. Therefore, many researchers now advocate that it is time to conduct research work on document warehousing to capture complete business intelligence. Document warehouses, unlike traditional document management systems, include extensive semantic information about documents, cross-document feature relations, and document grouping or clustering to provide a more accurate and more efficient access to text-oriented business intelligence. In this paper, we discuss the basic concept of document warehousing and present its formal definitions. Then, we propose a general system framework and elaborate some useful applications to illustrate the importance of document warehousing. The work is essential for establishing an infrastructure to help combine text processing with numeric OLAP processing technologies. The combination of data warehousing and document warehousing will be one of the most important kernels of knowledge management and customer relationship management applications.

Keywords: Data warehousing; Document warehousing; Knowledge management; OLAP

1. Introduction

Data warehousing [18] and data mining techniques [17] are gaining popularity as organizations realize the benefits of being able to perform multi-dimensional analyses of cumulated historical business data to help contemporary administrative decision-making [2,4,15, 17,22]. This inspires enterprises to eagerly delve into useful business intelligence (BI) from both internal

[☆] This research was partially supported by National Science Council, TAIWAN, ROC, under Contract No. NSC-91-2416-H-327-005.

[*] Corresponding author. Tel.: +886 7 6011000x4113; fax: +886 7 7659541.

E-mail addresses: imfrank@ccms.nkfust.edu.tw (F.S.C. Tseng), yhchou@cc.cma.edu.tw (A.Y.H. Chou).

[1] Tel.: +886 7 7438179.

0167-9236/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2005.02.011

and external data. Business intelligence is supposed to provide decision-makers with the tactical and strategic information they need for understanding, managing, and coordinating the operations and processes in organizations.

However, much of the efforts have only touched the tip of the information iceberg. While the techniques regarding data warehouses, multi-dimensional models, on-line analytical processing (OLAP), or even ad hoc reports have served enterprises well; they do not completely address the full scope of business intelligence. It is believed that [42], for the business intelligence of an enterprise, only about 20% information can be extracted from formatted data stored in relational databases. The remaining 80% information is hidden in unstructured or semi-structured documents. This is because the most prevalent medium for expressing information and knowledge is text. For instances, market survey reports, project status reports, meeting records, customer complaints, e-mails, patent application sheets, and advertisements of competitors are all recorded in documents.

Despite that, documents in the Web, enterprise repositories, and public document management systems are all growing as well. Therefore, knowledge workers, managers, and executives still have to spend much of the working moment reading dozens, if not hundreds, of various types of electronic documents spread over the Internet. There is just too much text to digest in daily life. The fast-growing and tremendous amount of documents has far exceeded the human ability for comprehension without powerful tools. As a result, when doing important decision-making, some relevant documents may be ignored, and some irrelevant documents may be considered by intuition. We believe that leaving out information induced from relevant documents or keeping information by intuitively guessing from irrelevant documents may be detrimental, causing disaster from the strategy weaved by incomplete information.

To alleviate this phenomenon, Grigsby [14], McCabe et al. [26] and Sullivan [33] have advocated that documents should be properly *warehoused* according to some well-defined concepts for expanding the scope of business intelligence to include textual information. Ishikawa and colleagues [19–21] even advocated this by implementing a prototype system to support management of compound documents, keyword-based and content-based retrieval. They used ECA rules to classify multimedia documents, and SOM (Self-Organizing Map) to cluster a set of collected texts into the number of groups in the retrieval space of manageable dimensions.

Hence, we think one of the next challenges of the information community will be the study of topics about document warehousing and text mining to help enterprises in obtaining complete business intelligence. Although research work regarding text mining have been conducted widely (for examples, the gentle readers are referred to Refs. [44,23–25,40,34]), however, the issues regarding document warehousing are rarely addressed. We have proposed a multi-dimensional indexing structure, called D-tree in Ref. [36] to study the performance measurement for constructing document warehouses. Some theoretical analyses on the properties of indexing a document warehouse were also elaborated in Ref. [35]. With document warehouses, the documents of enterprises can be well organized for effective analysis, or feature extraction to create distilled and fruitful business intelligence.

Since there are usually many diverse concepts involved in a document, a document is multi-dimensional in nature. Document warehouses, unlike traditional document management systems, include extensive semantic information about documents, cross-document feature relations, and document grouping or clustering to provide more accurate and more efficient access to text-oriented business intelligence. To facilitate flexible and effective multi-dimensional on-line analytical document processing and browsing, a multi-dimensional query language for querying document warehouses is indispensable. In Ref. [37], we have devised a multi-dimensional query expression for querying document warehouses to provide users an easy and efficient way of performing on-line analytical processing on documents.

Although issues about document warehousing have been addressed in Refs. [14,26,33], there are still no formal definitions established up to now. In this work, we will first discuss the concept of document warehousing and formally define the related terms. Then, we propose a framework for document warehousing and elaborate some applications of document warehousing to sketch an attractive roadmap of using document warehouses.