Data Warehousing Search Engine

Hadrian Peter
University of the West Indies, Barbados

Charles Greenidge
University of the West Indies, Barbados

INTRODUCTION

Modern database systems have incorporated the use of DSS (Decision Support Systems) to augment their decision-making business function and to allow detailed analysis of off-line data by higher-level business managers (Agosta, 2000; Kimball, 1996).

The data warehouse is an environment that is readily tuned to maximize the efficiency of performing decision support functions. However the advent of commercial uses of the Internet on a large scale has opened new possibilities for data capture and integration into the warehouse.

The data necessary for decision support can be divided roughly into two categories: internal data and external data. In this article, we focus on the need to augment external data capture from Internet sources and provide a tri-partite, high-level model termed the Data Warehouse Search Engine (DWSE) model, to perform the same.

We acknowledge efforts that also are being made to retrieve internal information from Web sources by use of the Web warehouse, which stores the Web user's mouse and keyboard activities online, the so called clickstream data. A number of Web warehouse initiatives have been proposed, including WHOWEDA (Madria et al., 1999).

To be clear, data warehouses have focused on large volumes of long-term historical data (a number of weeks, months, or years old), but the presence of the Internet with its data, which is short-lived and volatile, and improved automation and integration activities make the shorter time scales for the refresh cycle more attractive.

To attain the maximum benefits of the DWSE, significant technical contributions that surpass existing implementations must be encouraged from the wider database, information-retrieval, and search-engine technology communities. Our model recognizes the fact that a cross-disciplinary approach to research is the only way to guarantee further advancement in the area of decision support.

BACKGROUND

Data warehousing methodologies are concerned with the collection, organization, and analysis of data taken from several heterogeneous sources, all aimed at augmenting end-user business function (Berson & Smith, 1997; Inmon, 2003; Wixom & Watson, 2001).

Central to the use of heterogeneous data sources is the need to extract, clean, and load data from a variety of operational sources. Operational data not only need to be cleaned by removing bad records or invalid fields, but also typically must be put through a merge/purge process that removes redundancies and records that are inconsistent and lack integrity (Celko, 1995).

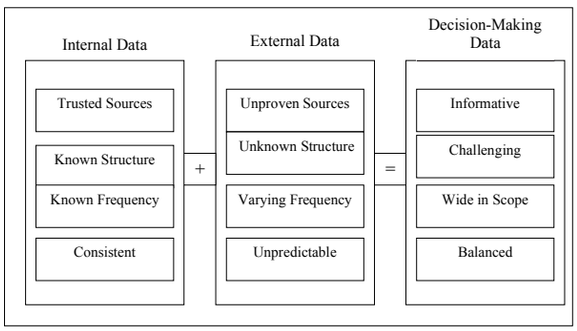
External data is key to business function and decision making, and includes sources of information such as newspapers, magazines, trade publications, personal contacts, and news releases. In the case where external data is being used in addition to data taken from disparate operational sources, this external data may require a similar cleaning/merge/purge process to be applied to guarantee consistency (Higgins, 2003).

The World Wide Web represents a large and growing source of external data but is notorious for the presence of bad, unauthorized, or otherwise irregular data (Brake, 1997; Pfaffenberger, 1996). Thus, the need for cleaning and integrity checking activities increases when the Web is being used to gather external data (Sander-Beuermann & Schomburg, 1998). The prevalence of viruses and worms on the Internet, and the problems with unsolicited e-mail (spam) show that communications coming across the Internet must be filtered. The cleaning process may include pre-programmed rules to adjust values, the use of sophisticated AI techniques, or simply mechanisms to spot anomalies that then can be fixed manually.

MAIN THRUST

Informed decision making in a data warehouse relies on suitable levels of granularity for internal data, where users can drill down from low to higher levels of aggregation. External data in the form of multimedia files, informal contacts, or news sources are often marginalized (Inmon, 2002; Kimball, Barquin & Edelstein, 1997) due to their unstructured nature.

Figure 1. Merits of internal/external data



True decision making embraces as many pertinent sources of information as possible so that a holistic perspective of facts, trends, and individual pieces of data can be obtained. Increasingly, the growth of commerce and business on the Internet has meant that in addition to traditional modes of disseminating information, the Internet has become a forum for ready posting of information of all kinds (Hall, 1999).

The prevalence of so much information on the Internet means that it is potentially a superior source of external data for the data warehouse. Since such data typically originate from a variety of sources (sites), it has to undergo a merging and transformation process before it can be used in the data warehouse. In the case of internal data, which forms the core of the data warehouse, previous manual methods of applying ad-hoc tools and techniques to the data cleaning process are being replaced by more automated forms such as ETL (Extract, Transform, Load). Although current ETL standardized packages are expensive, they offer productivity gains in the long run (Earls, 2003; Songini, 2004).

The existence of the so-called invisible Web and ongoing efforts to gain access to these untapped sources suggest that the future of external data retrieval will enjoy the same interest as that shown in internal data (Inmon, 2002; Sherman & Price, 2003; Smith, 2001). The need for reliable and consistent external data provides the motivation for an intermediate layer between raw data gathered from the Internet and external data storage areas lying within the domain of the data warehouse (Agosta, 2000).

Until there is a maturing of widely available tools with the ability to access the invisible Web, there will be a continued reliance on information retrieval techniques, as contrasted with data retrieval techniques, to gather external data (van Rijsbergen, 1979). The need for three environments to be present to process external data from Web sources into the warehouse suggests a three-tier solution to this problem. Accordingly, we propose a tri-partite model called the Data Warehouse Search Engine Model (DWSE), which has an intermediate data extraction/cleaning layer functionally called the Meta-Data Engine, sandwiched between the data warehouse and search engine environments.

The DWSE Model

The data warehouse and search engine environments serve two distinct and important roles at the current time, but there is scope to utilize the strengths of both in conjunction for maximum usefulness (Barquin & Edelstein, 1997; Sonnenreich & Macinta, 1998). Our proposed DWSE model seeks to allow cooperative links between data warehouse and search engine with an aim of satisfying external data requirements. The model consists of (1) data warehouse (DW), (2) meta-data engine (MDE) and (3) search engine (SE). The MDE is the component that provides a bridge over which information must pass from one environment to the other. The MDE enhances queries coming from the warehouse and also captures, merges, and formats information returned by the search engine (Devlin, 1998).

The new model, through the MDE, seeks to augment the operations of both by allowing external data to be collected for the business analyst, while improving the search engine searches through query modifications of queries emerging from the data warehouse.

The generalized process is as follows. A query originates in the warehouse environment and is modified by the MDE so that it is specific and free of nonsense words. A word that has a high occurrence in a text but conveys little specific information about the subject of the text is deemed to be a nonsense word. Typically, these words include pronouns, conjunctions, and proper names. This term is synonymous with noise word, as found in information retrieval texts (Belew, 2000). The modified query is transmitted to the search engine that performs its operations and retrieves its results documents. The documents returned are analyzed by the MDE, and information is prepared for return to the