Modeling Web-Based Data in a Data Warehouse

Hadrian Peter
University of the West Indies, Barbados

Charles Greenidge
University of the West Indies, Barbados

INTRODUCTION

Good database design generates effective operational databases through which we can track customers, sales, inventories, and other variables of interest. The main reason for generating, storing, and managing good data is to enhance the decision-making process. The tool used during this process is the decision support system (DSS). The information requirements of the DSS have become so complex, that it is difficult for it to extract all the necessary information from the data structures typically found in operational databases. For this reason, a new storage facility called a data warehouse has been developed. Data in the data warehouse have characteristics that are quite distinct from those in the operational database (Rob & Coronel, 2002).

The data warehouse extracts or obtains its data from operational databases as well as from other sources, thus providing a comprehensive data pool from which useful information can be extracted. The other sources represent the external data component of the data warehouse. External information in corporate data warehouses has increased during the past few years because of the wealth of information available, the desire to work more closely with third parties and business partners, and the Internet**.**

In this paper, we focus on the Internet as the source of the external data and propose a model for representing such data. The model represents a timely intervention by two important technologies (i.e., data warehousing and search engine technology) and is a departure from current thinking of major software vendors (e.g., IBM, Oracle, MicroSoft) that seeks to integrate multiple tools into one environment with one administration. The paper is organized as follows: background on the topic is discussed in the following section, including a literature review on the main areas of the paper; the main thrust section discusses the detailed model, including a brief analysis of the model; finally, we identify and explain the future trends of the topic.

BACKGROUND

Data Warehousing

Data warehousing is a process concerned with the collection, organization, and analysis of data typically from several heterogeneous sources with an aim to augment end-user business functions (Hackathorn, 1998; Strehlo, 1996). It is intended to provide a working model for the easy flow of data from operational systems to decision support systems. The data warehouse structure includes three major levels of information—granular data, archival data, and summary data—and the metadata to support them (Inmon, Zachman & Geiger, 1997).

Data warehousing:

1. Is centered on ad hoc end-user queries posed by business users rather than by information system experts.
2. Is concerned with off-line, historical data rather than online, volatile, operational type data.
3. Must efficiently handle larger volumes of data than those handled by normal operational systems.
4. Must present data in a form that coincides with the expectations and understanding of the business users of the system rather than that of the information system architects.
5. Must consolidate data elements. Diverse operational systems will refer to the same data in different ways.
6. Relies heavily on meta-data. The role of meta-data is particularly vital, as data must remain in its proper context over time.

The main issues in data warehousing design are:

1. performance versus flexibility;
2. cost;
3. maintenance.

Details of data warehousing/warehouses issues are provided in Berson and Smith (1997), Bischoff and Yevich (1996), Devlin (1998), Hackathorn (1995, 1998), Inmon (2002), Mattison (1996), Rob and Coronel (2002), Becker (2002), Kimball and Ross (2002), Lujan-Mora and Trujillo (2003).

External Data and Search Engines

External data represent an essential ingredient in the production of decision-support analysis in the data warehouse environment (Inmon, 2002). The ability to make effective decisions often requires the use of information that is generated outside the domain of the person making the decision (Parsaye, 1996). External data sources provide data that cannot be found within the operational database but are relevant to the business (e.g., stock prices, market indicators, and competitors' data).

Once in the warehouse, the external data go through processes that enable them to be interrogated by the end user and business analysts seeking answers to strategic questions. This external data may augment, revise, or even challenge the data residing in the operational systems (Barquin & Edelstein, 1997).

Our model provides a cooperative nexus between the data warehouse and search engine, and is aimed at enhancing the collection of external data for end-user queries originating in the data warehouse (Bischoff & Yevich, 1996). The model grapples with the complex design differences between data warehouse and search engine technologies by introducing an independent, intermediate, data-staging layer.

Search engines have been used since the 1990s to grant to millions access to global external data via the Internet (Goodman, 2000; Sullivan, 2000). In spite of their shortcomings (Chakrabarti, et al., 2001; Soudah, 2000), search engines still represent the best approach to the retrieval of Internet-based external data (Goodman, 2000; Sullivan, 2000). Importantly, under our model, queries originating with the business expert on the warehouse side can be modified at an intermediate stage before being acted on by the search engine. Results coming from the search engine also can be processed prior to being relayed to the warehouse. This intermediate, independent meta-data bridge is an important concept in this model.

THE DETAILED MODEL

Motivation/Justification

To better understand the importance of the new model, we must examine the search engine and meta-data engine components. The meta-data engine cooperates closely with both the data warehouse and search engine architectures to facilitate queries and efficiently handle results. In this paper, we propose a special-purpose search engine, which is a hybrid between the traditional, general-purpose search engine and the meta-search engine.

The hybrid approach we have taken is an effort to maximize the efficiency of the search process. Although the engine does not need to produce results in a real-time mode, the throughput of the engine may be increased by using multi-threading.

The polling of the major commercial engines by our engine ensures the widest possible access to Internet information (Bauer et al., 2002; Pfaffenberger, 1996; Ray et al., 1998; Sullivan, 2000).

The special-purpose hybrid search engine in Figure 1 goes through the following steps each time a query is processed:

1. Retrieve major search engine's home pages HTML documents;
2. Parse files and submit GET type queries to major engines;
3. Retrieve results and use links found in results to form starting seed links;
4. Perform depth/breath first traversal using seed links; and
5. Capture results from no. 4 to local disk.

Figure 1. Query and retrieval in hybrid search engine