Search Engines and their Impact on Data Warehouses

Hadrian Peter
University of the West Indies, Barbados

Charles Greenidge
University of the West Indies, Barbados

INTRODUCTION

Over the past ten years or so *data warehousing* has emerged as a new technology in the database environment. "A data warehouse is a global repository that stores pre-processed queries on data which resides in multiple, possibly heterogeneous, operational or legacy sources" (Samtani et al, 2004).

Data warehousing as a specialized field is continuing to grow and mature. Despite the phenomenal upgrades in terms of data storage capability there has been a flood of new streams of data entering the warehouse. During the last decade there has been an increase from 1 terabyte to 100 terabyte and, soon to be 1 petabyte, environments. Therefore, the ability to search, mine and analyze data of such immense proportions remains a significant issue even as analytical capabilities increase.

The data warehouse is an environment which is readily tuned to maximize the efficiency of making useful decisions. 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.

While most of the data necessary for a data warehouse originates from the organization's internal (operational) data sources, additional data is available externally that can add significant value to the data warehouse. One of the major reasons why organizations implement data warehousing is to make it easier, on a regular basis, to query and report data from multiple transaction processing systems and/or from external sources. One important source of this external data is the Internet.

A few researchers (Walters, 1997; Strand & Olsson, 2004; Strand & Wangler, 2004) have investigated the possibility of incorporating external data in data warehouses, however, there is little literature detailing research in which the Internet is the source of the external data. In (Peter & Greenidge, 2005) a high-level model, the Data Warehousing Search Engine (DWSE), was presented. However, in this article we examine in some detail the issues in search engine technology that make the Internet a plausible and reliable source for external data. As John Ladley (Ladley, 2005) states "There is a new generation of Data Warehousing on the horizon that reflects maturing technology and attitudes". Our long-term goal is to design this new generation Data Warehouse.

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; Wixom & Watson, 2001; Inmon, 2003). Central to the use of heterogeneous data sources is the challenge to extract, clean and load data from a variety of operational sources. External data is key to business function and decision-making, and typically 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 will require a cleaning/merge/purge process to be applied to guarantee consistency (Higgins, 2003; Walters, 1997). The Web represents a large and growing source of external data but is notorious for the presence of bad, unauthorized or otherwise irregular data (Kim, 2003). Thus the need for cleaning and integrity checking activities increases when the web is being used to gather external data.

External data has the following advantages (Strand & Olsson, 2004) and disadvantages (Strand & Wangler, 2004):

Advantages

- Ability to compare internal data with external data
- Helps with the acquisition of data related customers
- Helps with the acquisition of additional data about the marketplace

Disadvantages

- Ensuring the quality of the external data
- Making users trust the external data
- Physically integrating the external data with the internal data
- Conceptually mapping the external data with the internal data

Given that the focus of this article is the use of external data (accessed from the web) in the data warehouse, it is in our interest to highlight the advantages and minimize or, if possible, eliminate the disadvantages mentioned above. We believe that one way of doing so is to use the appropriate search engine(s) to access the data from the Internet. Figure 1 illustrates the role played by external data in the decision-making process.

MAIN THRUST

Search Engines: Comparison and Background Information

The literature is replete with examples of search engines for accessing external data from the Internet. In this section we examine a number of search engines and compare their characteristics.

There are substantial limitations to Search Engines. Most of them rely on crawler programs that go through a process of indexing web pages during a search (Butler, 2000). Most search engines only cover a fraction of the web, because vast amounts of data and documents stored in databases cannot be accessed by the current versions of web crawlers. These search engines are therefore unfortunately incapable of extracting information from the deep web or invisible web where much useful external data resides.

A long-standing shortcoming of search engines is that they provide only one way of organizing their search – by *salience (confidence, or certainty*) value (Bean, 2007). This value indicates how likely it is that a returned result matches the query. Such shortcomings

Figure 1. Merits of internal/external data