Web Page Extension of Data Warehouses

Anthony Scime
State University of New York College Brockport, USA

INTRODUCTION

Data warehouses are constructed to provide valuable and current information for decision-making. Typically this information is derived from the organization's functional databases. The data warehouse is then providing a consolidated, convenient source of data for the decision-maker. However, the available organizational information may not be sufficient to come to a decision. Information external to the organization is also often necessary for management to arrive at strategic decisions. Such external information may be available on the World Wide Web; and when added to the data warehouse extends decision-making power.

The Web can be considered as a large repository of data. This data is on the whole unstructured and must be gathered and extracted to be made into something valuable for the organizational decision maker. To gather this data and place it into the organization's data warehouse requires an understanding of the data warehouse metadata and the use of Web mining techniques (Laware, 2005).

Typically when conducting a search on the Web, a user initiates the search by using a search engine to find documents that refer to the desired subject. This requires the user to define the domain of interest as a keyword or a collection of keywords that can be processed by the search engine. The searcher may not know how to break the domain down, thus limiting the search to the domain name. However, even given the ability to break down the domain and conduct a search, the search results have two significant problems. One, Web searches return information about a very large number of documents. Two, much of the returned information may be marginally relevant or completely irrelevant to the domain. The decision maker may not have time to sift through results to find the meaningful information.

A data warehouse that has already found domain relevant Web pages can relieve the decision maker from having to decide on search keywords and having to determine the relevant documents from those found in a search. Such a data warehouse requires previously conducted searches to add Web information.

BACKGROUND

To provide an information source within an organization's knowledge management system, database structure has been overlaid on documents (Liongosari, Dempski, & Swaminathan, 1999). This knowledge base provides a source for obtaining organizational knowledge. Data warehouses also can be populated in Web-based interoperational environments created between companies (Triantafillakis, Kanellis & Martakos, 2004). This extends knowledge between cooperating businesses. However, these systems do not explore the public documents available on the Web.

Systems have been designed to extract relevant information from unstructured sources such as the Web. The Topicshop system allows users to gather, evaluate, and organize collections of Web sites (Amento, Terveen, Hill, Hix, & Schulman, 2003). Using topic discovery techniques Usenet news searching can be personalized to categorize contents and optimise delivery contents for review (Manco, Ortale & Tagarelli, 2005). Specialized search engines and indexes have been developed for many domains (Leake & Scherle, 2001). Search engines have been developed to combine the efforts of other engines and select the best search engine for a domain (Meng, Wu, Yu, & Li, 2001). However, these approaches do not organize the search results into accessible, meaningful, searchable data.

Web search queries can be related to each other by the results returned (Wu & Crestani, 2004; Glance, 2000). This knowledge of common results to different queries can assist a new searcher in finding desired information. However, it assumes domain knowledge sufficient to develop a query with keywords, and does not provide corresponding organizational knowledge.

Some Web search engines find information by categorizing the pages in their indexes. One of the first to create a structure as part of their Web index was Yahoo! (http://www.yahoo.com). Yahoo! has developed a hierarchy of documents, which is designed to help users find information faster. This hierarchy acts as a taxonomy of the domain. Yahoo! helps by directing the searcher through

the domain. Again, there is no organizational knowledge to put the Web pages into a local context, so the documents must be accessed and assimilated by the searcher.

DynaCat provides knowledge-based, dynamic categorization of search results in the medical domain (Pratt, 1999). The domain of medical topics is established and matched to predefined query types. Retrieved documents from a medical database are then categorized according to the topics. Such systems use the domain as a starting point, but do not catalog the information and add it to an existing organized body of domain knowledge such as a data warehouse.

Web pages that contain multiple semi-structured records can be parsed and used to populate a relational database. Multiple semi-structured records are data about a subject that is typically composed of separate information instances organized individually, but generally in the same format. For example, a Web page of want ads or obituaries. The first step is to create an ontology of the general structure of the semi-structured data. The ontology is expressed as an Object-Relationship Model. This ontology is then used to define the parsing of the Web page. Parsing into records uses the HTML tags to determine the structure of the Web page, determining when a record starts and ends. The relational database structure is derived from the ontology. The system requires multiple records in the domain, with the Web page having a defined structure to delimit records. However, the Web pages must be given to the system, it cannot find Web pages, or determine if they belong to the domain (Embley et al., 1999).

The Web Ontology Extraction (WebOntEx) project semi-automatically determines ontologies that exist on the Web. These ontologies are domain specific and placed in a relational database schema. Using the belief that HTML tags typically highlight a Web page's concepts, concepts are extracted, by selecting some number of words after the tag as concepts. They are reviewed and may be selected to become entity sets, attributes or relationships in a domain relational database. The determination is based on the idea that nouns are possible entity and attribute types and verbs are possible relationship types. By analyzing a number of pages in a domain an ontology is developed within the relational database structure (Han & Elmasri, 2004). This system creates the database from Web page input, whereas an existing data warehouse needs only to be extended with Web available knowledge.

Web based catalogs are typically taxonomy-directed. A taxonomy-directed Web site has its contents organized in a searchable taxonomy, presenting the instances of a category in an established manner. DataRover is a system that automatically finds and extracts products from taxonomy-directed, online catalogs. It utilizes heuristics to turn the online catalogs into a database of categorized products (Davulcu, Koduri & Nagarajan, 2003). This system is good for structured data, but is not effective on unstructured, text data.

To find domain knowledge in large databases domain experts are queried as to the topics and subtopics of a domain creating an expert level taxonomy (Scime, 2000, 2003). This domain knowledge can be used to assist in restricting the search space. The results found are attached to the taxonomy and evaluated for validity; and create and extend the searchable data repository.

WEB SEARCH FOR WAREHOUSING

Experts within a domain of knowledge are familiar with the facts and the organization of the domain. In the warehouse design process, the analyst extracts from the expert the domain organization. This organization is the foundation for the warehouse structure and specifically the dimensions that represent the characteristics of the domain.

In the Web search process; the data warehouse analyst can use the warehouse dimensions as a starting point for finding more information on the World Wide Web. These dimensions are based on the needs of decision makers and the purpose of the warehouse. They represent the domain organization. The values that populate the dimensions are pieces of the knowledge about the warehouse's domain. These organizational and knowledge faucets can be combined to create a dimension-value pair, which is a special case of a taxonomy tree (Kerschberg, Kim & Scime, 2003; Scime & Kerschberg, 2003). This pair is then used as keywords to search the Web for additional information about the domain and this particular dimension value.

The pages retrieved as a result of dimension-value pair based Web searches are analyzed to determine relevancy. The meta-data of the relevant pages is added to the data warehouse as an extension of the dimension. Keeping the warehouse current with frequent Web searches keeps the knowledge fresh and allows decision makers access to the warehouse and Web knowledge in the domain.

WEB PAGE COLLECTION AND WAREHOUSE EXTENSION

The Data Warehouse Web Extension Architecture (Figure 1) shows the process for adding Web pages to a data warehouse.