Managing Very Large Databases and Data Warehousing

G.N. Wikramanayake and J.S. Goonetillake
University of Colombo School of Computing
35 Reid Avenue,
Colombo 07,
Sri Lanka
Tel. +94-11-2591065, Fax: +94-11-2587239
email: {gnw, jsg}@ucsc.cmb.ac.lk

Abstract

Major libraries have large collections and circulation. Managing libraries electronically has resulted in the creation and management of large library databases. The interconnection of libraries and sharing resources across libraries has resulted in the management of very large databases. Most large and/or multinational industries worldwide have exploited such opportunities by applying data warehouse technology to their data repositories to discover knowledge that had helped them to gain competitive advantage through decision making. The same can be done for libraries using the available large databases. This paper identifies the changes that had taken placed in libraries due to technology and how the data warehouse technology could assist them to discover knowledge and improve services.

Keywords: Data Warehousing, Data Mining, Digital Libraries, Very Large Databases

Introduction

A library would record data about their books using library catalogues. Each catalogue consists of data about the author, title, subject, publisher, edition, place, year, language and ISBN of the book. In a manual library system one maintains library catalogues only in author sequence as each catalogue entry requires a separate entry card which is placed in the catalogue drawer. Thus a user could search for a catalogue entry only by author name.

All applications use a database to manage data. Database is an organised collection of data managed using a database management system (Elmasiri & Navathe 2003). A library application would manage catalogues using the database technology and such applications are referred to as e-catalogue. The entries in an e-catalogue are reusable since a user could search for a catalogue entry not only by author name but also by any other data field such as title and publisher. Library application packages such as LibSys and CDS/ISIS have enabled effective management of e-catalogues.

Databases are typically used for storing and manipulating of data. Using data processing/manipulation, meaning can be applied to data. For this, data is converted from one form to another. For instance, the books written by Rowling or the most popular book could be derived using the data of a library database. To retrieve the books written by Rowling one needs to list the book titles corresponding to that author. To determine the most popular book one needs to determine the number of books borrowed and/or reserved over a period of time and see which book appeared the most.

Usually most librarians would like to have popular books available in their libraries irrespective of where it is located. Hence it is likely that two popular libraries listing the books written by Rowling would get the same list. However, it is unlikely that two libraries identifying the same book as the most popular book. This is usually because of the differences in the membership of respective libraries, and their social and cultural interests. To identify such differences and other related information such as borrowers with similar interest, e-catalogues have to be exploited effectively.

The paper next introduces the formation of large databases, and the roles the e-catalogues and digital libraries play. This is followed by an introduction to the data warehousing and data mining concepts. Finally the opportunities that have emerged due to the adaptation of technology are discussed.

Large Databases

In computing storage capacity grows rapidly and technology supports increase of storage capacities while decreasing the storage space. Thus “Large” is a relative term that changes with time. What was large five or ten years ago could be small by today’s standards, and what is large today will not be so in a few years from now. However, a Very Large Database (VLDB) is typically a database that contains an extremely high number of tuples (database rows or records), or occupies an extremely large physical file system storage space due to wide tables with large numbers of columns or due to multimedia objects. However, the most common definition of VLDB is a database that occupies more than 1 terabyte or contains several billion rows. Data Warehouses, Decision Support Systems (DSS) and On-Line Transaction Processing (OLTP) systems serving large numbers of users would fall into this category.

Can a library grow up to a terabyte of data? In most cases it does not as even the storage space of computers used by most libraries are not in this range. Let us perform a calculation to determine the storage requirements for a library. If a catalogue consists of 200 characters it would use 200 kilobytes of data. If there are 1 million items then the storage requirement would be 200 million kilobytes or 200 gigabytes. If a library has 50,000 members with each taking 200 kilobytes then the storage requirement would be 10 gigabytes. If a circulation record is using 40 bytes per item and if 1000 item borrowing happens per day then the storage requirement would be 40 kilobytes. Over a year this would accumulate to about 10 gigabytes. Thus the total storage requirement to manage electronic transactions will not fall under the very large database category. However data warehousing and mining can be performed on large e-catalogues as well.

Now let us look at some storage requirements for digital material. A scan A4 page would on average would take about 50 kilobytes of storage (White paper 22009). Standard storage box or a drawer is estimated to have about 2500 pages of information. Thus a CD could store content of 4 drawer file cabinet. This type of calculation is good for organisations who want to put their records electronically. However for book, it is not practical to scan in the first place as it could damage the book, especially if the book is quite old. However if we look at how digital copiers are performing their photocopying activities we could see that this is not that difficult. If we assume that a book on average has 250 pages then 40 books may go into one CD. This calculation would go up if one considers high resolution colours material. Similarly the requirements will go up for audio and video content. Thus when digital material is included e-catalogues are considered as digital libraries using VLDB.

e-Catalogue

A library managing an e-catalogue would have three categories of data, namely bibliographic (catalogue), circulation (borrowing) and acquisition (purchasing). For each catalogue entry there would be some information about its acquisition process. Typically this is restricted only to the price and seller information. However, the recording of the details on how the decision process to purchase the item would help the future decision making process. With respect to borrowing there would be many entries for a catalogue if the item is on high circulation while there would be few or on entries for low circulation items. However on average there are many entries for a catalogue and it keeps growing.

Library circulation data is required to be kept in the database until the items borrowed are returned. However, as we would see later such data would serve a librarian in decision making if we retain these data.

The global interconnections of locations through computer networks have allowed data across libraries to be shared across their users. This has allowed effective sharing of library resources and performing activities such as inter-library borrowing. The concept of data warehousing allows inferring knowledge beyond the capabilities of local databases. Most large and/or multinational industries worldwide have already applied data warehouse technology to their data repositories and have discovered knowledge to gain competitive advantage through decision making. Thus researchers in libraries are trying to do the same to discover similar opportunities (Guenther 2000; Needamangala 2000; Baruque & Melo 2004; Dwivedi & Bajpai 2004; Prakash, Chand & Gohel 2004). For instance, a decision like “on which books should a library invest to serve their members better?” could be made after such discovery.

Digital library

Digital library, also called an electronic library is being widely adopted across many libraries and thus have moved from relatively few people’s research interests to a wider application and use. Digital libraries have integrated different information sources and increased the use of information.

In addition to the three categories of e-catalogues a digital library would have the fourth categories, namely the digital content (electronic book/journal/video). Content of the library item would be kept digitally if online access to digital content is to be provided to its users. Publishers of electronic documents (e.g. e-books, e-journals) do have their documents online.

Functions of a library have grown beyond maintaining books, magazines and newspapers. Many libraries also provide CD/video lending, and online searching, reservations and browsing e-journals. Certain universities and libraries have even moved beyond this level and provide full-text books, multimedia manuscripts and periodicals (Chen 2000). We also see newspapers, technical documents being made available on the web along with its print edition (Associated Newspapers of Ceylon Ltd., Online).

Digital libraries allow not only easy of reaching information through search engines and indexing techniques. It has also provided solutions for other issues such as space, preservation of books and rescuing the content of fragile material.

Education is now moving towards an electronic learning environment (Wikramanayake 2005) and digital libraries will play a major role in achieving it. Following are among the benefits it would bring to the libraries in particular and society in general (Fox 1993). There will not be any boundaries in the distribution and dissemination of information. The performance of a library would increase immensely while accessibility to all kinds of items being provided through a single workstation located anywhere. Although dealing with more people and information, the administrative overheads are very minimal as computers do most of the tasks without any human intervention.

Data Warehousing

Data can now be stored in many different types of databases. One type of database architecture that has recently emerged is data warehouse, which is a repository of multiple heterogeneous data sources, organised under a unified schema at a single site in order to facilitate management decision-making (Chaudhuri & Dayal 1997; Chawatte, Garcia-Molina, Hammer, Ireland, Papakonstantinou, Ullman & Wisdom 1994; Han & Kamber 2001). Data warehouse technology includes data cleaning, data integrating, and on-line analytical processing (OLAP) that is, analysis techniques with functionalities such as summarisation, consolidation and aggregation, as well as the ability to view information from different angles. A data warehouse is defined as a “subject-oriented, integrated, time variant, non-volatile collection of data that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. In data warehouses historical, summarised and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be much larger than operational databases. Most queries on data warehouses are ad hoc and are complex queries that can access millions of records and perform a lot of scans, joins, and aggregates. Due to the complexity query throughput and response times are more important than transaction throughput.

Data Mining

Data Mining is the extraction or “Mining” of knowledge from a large amount of data or data warehouse. To do this extraction data mining combines artificial intelligence, statistical analysis and database management systems to attempt to pull knowledge form stored data.

Data mining is the process of applying intelligent methods to extract data patterns. This is done using the front-end tools. The spreadsheet is still the most compiling front-end application for Online Analytical Processing (OLAP). The challenges in supporting a query environment for OLAP can be crudely summarised as that of supporting spreadsheet operation effectively over large multi-gigabytes databases.

To distinguish information extraction through data mining from that of a traditional database querying, the following main observation can be made. In a database application the queries issued are well defined to the level of what we want and the output is precise and is a subset of operational data. In data mining there is no standard query language and the queries are poorly defined. Thus the output is not precise (fuzzy) and do not represent a subset of the database. Beside the data used not the operational data that represents the to day transactions. For instance during the process of building a data warehouse the operational data are summarised over different characteristics, such as borrowings during 3 months period. Queries can be of the type of “identify all borrowers who have similar interest” or “items a member would frequently borrow along with movies”, which is not a precise as the list of books borrowed by a member. The nature of the database and the query result in extracting non-subset of data.

In supermarkets such relationships have already been identified using data mining. Thus related items such as “bread and milk’ or “beer and potato chips” would be kept together. Mobile companies decide on peak hours, rates and special packages based similar market research.

Users can use data mining techniques on the data warehouse to extract different kinds of information which would eventually assist the decision making process of an organisation (figure 1). For example, if certain books are rarely used by members of a particular library, while the same books are frequently used at other libraries then it is appropriate to transfer these books to respective libraries to ensure its effective use. Such knowledge could only be discovered through sharing experiences of librarians or by capturing the knowledge through database and integrating them as done when building data warehouses. Decision support tools assist users in discovering knowledge.

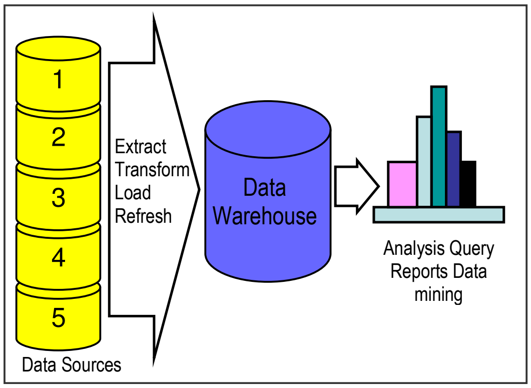

Figure 1: Mining a Data Warehouse

Decision making using a Data Warehouse

A Decision Support System (DSS) is any tool used to improve the process of decision making in complex systems. A DSS can range from a system that answer simple queries and allows a subsequent decision to be made, to a system that employ artificial intelligence and provides detailed querying across a spectrum of related datasets. Amongst the most important application areas of DSS are those complicated systems that directly “answer” questions, in particular high level “what-if” scenario modelling.

Over the last decade there was a transition to decision support using data warehouses (Inmon 2002). The data warehouse environment is more controlled and therefore more reliable for decision support than the previous methods. The data warehouse environment supports the entire decision support requirements by providing high-quality information, made available by accurate and effective cleaning routines and using consistent and valid data transformation rules and documented pre-summarisation of data values. It contains one single source of accurate, reliable information that can be used for analysis.

Multi-dimensional Data

To facilitate complex analyses and visualisation, the data in a warehouse is typically modelled multi-dimensionally. For example, in a library data warehouse, time of borrowing, borrower's district, age group and book category might be some of the dimensions of interest (see figure 2). Often, these dimensions are hierarchical, e.g. time of borrowing may be organised as a day-month-quarter-year hierarchy.

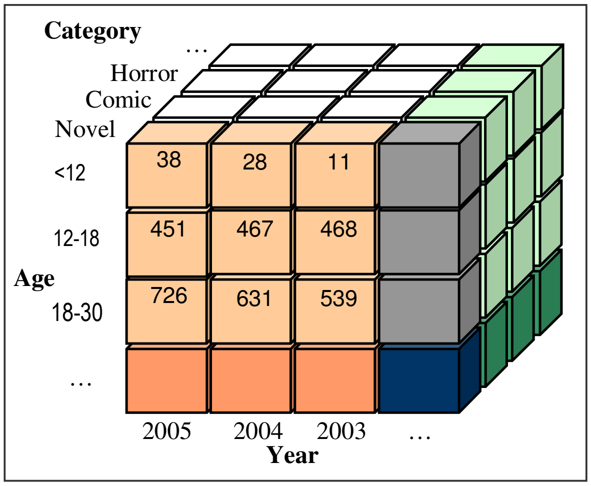

Figure 2: A Multi-dimensional Data Cube

Data Analysis

Typically data analysis is done through OLAP operations such as rollup (increasing level of aggregation) and drill-down (decreasing level of aggregation or increasing details) along one or more dimension hierarchies, slice-and-dice (selection and projection) and pivot (re-orienting the multidimensional view of data).

Analysing and query processing of huge data warehouse is very difficult and time-consuming task. Therefore multi dimensional data cubes consisting of summary tables are created for all possible decision marking activities (Harinarayan, Rajaraman & Ulman 1996; Fernando & Wikramanayake 2004). Front-end tools are available to obtain such information.

Mining Library Data

Libraries arrange their items based on pre-defined subject areas. Although books related to the same subject are kept close to each other, related items that would be identified as in the supermarket example may not be together. For instance the related videos and CD or even magazines or journals would be available elsewhere. Thus the chance of a reader getting to know availability of such items and making use of them is low. Data mining techniques must be applied to data to reveal such information.

Data of a digital library should be organised in a manner in which it could be analysed later. Classification of data according to identified characteristics is one way of achieving it. Classification will allow grouping of discrete values such as by subject of a book. However estimation would be required to deal with continuous values such as age groups of borrowers. Classified and estimated data could be used to predict future behaviours and decision making could be done accordingly. Membership characteristics such as disability, ethnic group income group and social characteristics could help to determine affinity groups. Services offered could target such groups. For example, reduced membership fees or special facilities may be offered to them and ensure that they are not neglected.

Clustering is another technique used to effectively organise information by segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters. Clustering differs from classification as it is based on self-similarity then predefined classes. For example one may group books based on author or publisher due to the high probability of such books falling into the same group.

The applicability of above data mining techniques is highly depended on the size of the database. If the database is small, it is possible to find interesting patterns and relationships by simple inspection of results from familiar tool such as spreadsheets and multidimensional query tools. However in order to generate classification rules, association rules, clusters, or predictions one require a large amount of data.

Since the data of the library continuously growing with an exponential rate, the main problem of referencing the required information form the large amount of redundant information of the library can be reduced using mining techniques. Searching through classification of content of the library and acquisition of books through data mining knowledge are among the main benefits that a library would gain through the management of large databases and data warehousing.

Conclusion

Digital libraries have emerged over the last decade and they are used beyond the research community and selected institutions. Libraries should prepare to exploit these digital collections for decision making and provide services to suit the digital society. Knowledge discovered from one library would be different to that of another due to user characteristics. Thus there would be the need to do appropriate research to recognise the needs of urban and rural membership.

References

Abramson I, Abbey M & Corey M 2004, *Oracle Database 10g A Beginner's Guide*, McGraw-Hill/Osborne.

Associated Newspapers of Ceylon Ltd., Available at http://www.dailynews.lk/ [25.10.2006]

Baruque, CB & Melo, RN 2004, *Developing Digital Libraries Using Data Warehousing and Data Mining Techniques*, Available at http://www.pgl.ufl.edu/events/pgl2/CBaruque/digital_libraries.pdf [21.10.2006]

Chaudhuri, S & Dayal, U 1997, *An Overview of Data Warehousing and OLAP Technology*, SIGMOD Record 26:1, Mar, pp 65-74.

Chawatte, SS, Garcia-Molina H, Hammer J, Ireland K, Papakonstantinou Y, Ullman J & Wisdom J 1994, *The TSIMMIS Project: Integration of Heterogeneous Information Sources*, Proc. of IPSJ Conf., Tokyo, Japan, Oct., pp. 7-18.

Chen, H 2000, *The Illinois Digital Library Initiative Project: Federating Repositories and Semantic Research*, Available at http://ai.bpa.arizona.edu/hchen/docs/DLI/ [22.10.2006]

Dwivedi, RK & Bajpai, RP 2004, *Use of Data Mining in the field of Library and Information Science: An Overview*, India.

Elmasiri, R & Navathe, SB 2003, *Fundamentals of Database Systems*, 4th Edition, Addison-Wesley, 2003.

Fernando, MGNAS & Wikramanayake, GN 2004, *Application of Data Warehousing & Data Mining to Exploitation for Supporting the Planning of Higher Education System in Sri Lanka*, Proceedings of 23rd National Information Technology Conference, Published by Computer Society of Sri Lanka, Colombo, Sri Lanka, 8-9 July, pp. 114-120.

Fox, EA (ed.) 1993, *Source Book on Digital Libraries*, TR 93-35, Dept. of Computer Science, Virginia Tech, Available at http://fox.cs.vt.edu/DigitalLibrary/DLSB.pdf [23.10.2006]

Gilheany, S, *Sizing Document Management Systems: Image Size Estimates for All Types of Digitized Documents*, White paper 22009, Computer Storage requirements for various digitized documents. Available at http://www.berghell.com/whitepapers/Storage%20Requirements.pdf [25.10.2006]

Guenther, K 2000, *Applying data mining principles to library data collection*, Computers in Libraries, 20:4, pp. 60-63.

Han, J & Kamber, M 2001, *Data Mining Concepts and Techniques*, Morgan Kaufmann.

Harinarayan, V, Rajaraman, A & Ulman, JD 1996, *Implementing Data Cubes Efficiently*, Proc of SIGMOD Record, 25:2, pp. 205-216.

Inmon, WH 2002, *Building the Data Warehouse*, 3rd Edition, Wiley.

Needamangala, A 2000, *A Library Decision Support System Built on Data Warehousing and Data Mining Concepts and Techniques*, Master of Science Thesis, University of Florida.

Prakash, K, Chand, P & Gohel, U 2004, *Application of Data Mining in Library and Information* Services. 2nd Convention PLANNER - 2004, Manipur Uni., Imphal, India, 4-5 November.

Wikramanayake, GN 2005, *Impact of Digital Technology on Education*, Proceedings of 24th National Information Technology Conference, Published by Computer Society of Sri Lanka, Held in Colombo, Sri Lanka, 15-16 Aug., pp. 82-91.