



A Data Warehouse for Mining Usage Pattern in Library Transaction Data

Pankaj Kumar Deva Sarma*, Rahul Roy

Department of Computer Science, Assam University, Silchar

Correspondence; *e-mail: pankajgr@rediffmail.com

Abstract

Data Warehousing and Data Mining has evolved and matured into a discipline of research and application over the last decade and a half. Wide ranging applications have emerged across various domains and Data Warehousing and Data Mining tools and applications have yielded much needed decision support to the organizations. An enormous proliferation of database in every aspect of human endeavor has created demand for designing such tools that can turn data into useful task oriented knowledge. A library is no exception in this regard. It contains data of various operations starting from the stage of procurement, stacking to subscriber transactions like issue, return, and requisition and so on. As time progresses such data accumulates to huge size and contains within it patterns which, if known, may guide the future courses of actions better with scope of effective decision making which in turn could result in money saving and better planning and utilization of resources. However, typical library management software lacks the back end support of a built in Data Warehouse which can be used for mining the volumes of data accumulated over the years of operations. Considering this to be bottom liner we have developed and implemented a Data Warehouse which can be the backend of a decision support system which rely on the discovery of patterns of transaction data in a library.

Keywords: Data Mining, Data Warehousing, Snow Flake Schema.

Introduction

Library management concerns with the management of resources which basically includes books, manuscripts, journals etc. and providing effective and efficient services to its users. Since the manual management of books and other resources in a library and keeping track of every books of the library accessed by the user is tedious job and hence often technological support is expected. Further, to keep track of the books issued and returned across the counters, journals, periodicals and manuscripts consulted by the users and so on needs additional book keeping on additional parameters and are generally not done in a typical library which operates manually. The situation becomes acute when the library does procurement of resources. Most of the time, the procurement is done by looking into the availability of the stocks which may result in inefficient procurement. Even various library management tools do not provide with such additional support.

Generally, libraries are concerned with mere issue and return of books and to some extent delayed returns are kept track off manually. Though there are various library management software packages yet, more often than not these do not keep track of the operational or transactional details occurred on account of operations of the library over longer duration of time. It is obvious that the usage patterns of the library books and journals developed over the years are embedded in these transactions. If these patterns can be discovered then these can effectively influence the library operations, investment and procurement plans for the future growth of a library and so on. So there is need for automated computer based system that can help us in management of the resources of the library. Further, it is expected that it should be intelligent enough to discover the usage trends or pattern of the resource utilization so that efficient procurement and management

of resources can be carried out. To implement such a system, a Data Warehouse is required to store the necessary historical data generated through the operation of the library. The purpose of this warehouse is to assist a Data Mining tool which can run on top of it.

In the proposed system for discovery of usage pattern from library transaction data accumulated over the years, the design is done keeping in mind the automation necessary for management of books and journals stored in the stacks and then the corresponding issue and return of these items. Here attempt is made not to design and implement complete library management software since there are plenty of such systems available. Rather, the proposed system is expected to complement traditional library management software by allowing it to be used for discovering the usage pattern of the books and journals as developed over the years based on the actual transaction data.

Thus, in the automated library transaction management system proposed here, there is an online transaction system that stores the current information regarding every transaction of the library. The system stores the data in an operational database. Also these transactions are backed up in a Data Warehouse which stores subject oriented, time variant, non volatile data. These data are then extracted by the data mining tools for knowledge discovery. The features that are incorporated while designing this library transaction management system are:

- * Keeping track of the books that are issued to the members.
- * Allowing users to track the books and journals in the library.
- * Faster mining and retrieval of information.
- * Reduced work load of employee.

Library Schema

Once all the requirements were collected and analyzed for the design of the system, the next step was to create a conceptual schema for a database. The conceptual schema provides a concise description of the data requirements by the user and includes description of the entities

type, relationships and constraints (Ramez et al., 2006). In designing the conceptual schema for the database, different types of entities are considered. Some of the prime entities considered for schema are:

1. Book schema
2. Journal schema
3. Transaction details

In the book schema we store the detail information regarding the books. The three entities obtained from normalization of the book schema are book details which provide details of the book, *Isbn_book* provides details regarding the correspondence of the isbn_no with every book and *book_copies* provide details regarding the no of copies in stock for a particular book. The schema is shown below (Figure 1) :

Book_detail

Book_id	Author_name	Rack_no	Book_type	Isbn_no
Isbn_book:				
Isbn_no		Book_name		
Book_copies:				
Book_name		Total_copies		

Figure 1: The Book Schema

The journal schema is same as book schema. Here we have two entities. First, providing the data regarding the journal details and the second providing the correspondences of the ISBN no with the journals. The schema is shown below (Figure 2):

Journal details

journal_id	Rack_no	publication	Isbn_no
Isbn_journal			
Isbn_no		journal_name	

Figure 2: The Journal Schema

The last prime transaction schema is the transaction details which stores information regarding the issue and return of books. Here we divide the entire schema into two types viz. One is issue transaction which provides the details of

the books that are being issued out and the return transaction which provides information for book that are returned back to the library. The schemas are shown below:

Issue_transaction

Book_id	Date_ofissue	Return_date	Member_id	Book_name
---------	--------------	-------------	-----------	-----------

Return transaction:

Return_date	Book_id	Returned_on	Member_id	Delay_in return
-------------	---------	-------------	-----------	-----------------

Figure 3: The Transaction Schema

Data Warehousing

A data warehouse is a system that *stores and consolidates data periodically* from the source systems into a *dimensional or normalized data store*. It usually keeps years of *historical data* and can be *mined* for pattern discovery for *business intelligence* or other *analytical activities*. (Chen et al., 1996) It is typically updated in *batches*, not every time when a transaction happens in the source system. A Data warehouse is maintained separately from an organization’s operational databases. Data warehouse systems allow for the integration of a variety of application systems. There are various models for Data Warehouse. Some among these are (Han and Kamber, 2002):

Star schema

The star schema is a modeling paradigm in which the data warehouse contains

- (i) a central table (fact table)
- (ii) a set of smaller attendant tables (dimension tables) for each dimensions.

The schema graph resembles a starburst, with the radial pattern around the central fact table.

Snowflake schema

The snowflake schema is a variant of star schema where some dimension tables are normalized, there by further splitting the data into additional tables.

Fact constellation

Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars,

and hence is called a galaxy schema or a fact constellation.

Data Warehouse Design

In the data warehouse design, a snowflake schema is used. There are three dimensional tables viz. *book schema, issue transaction schema and return transaction schema*. The *return transaction schema* is normalized and we derived a dimension table *book dimension* to provide the correspondence between book id and book name. The design view is shown below:

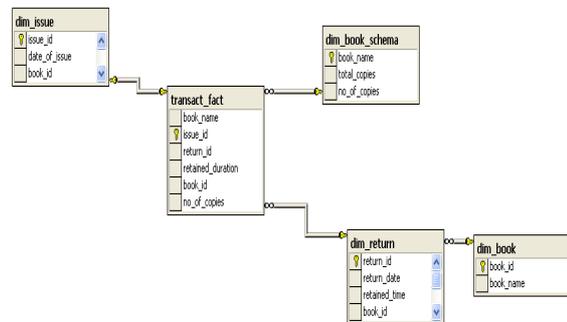


Figure 4: A view of the Data Ware House

Data Mining Approach

Data mining refers to extraction or mining of knowledge from large amount of data (Han and Kamber, 2002) Data mining techniques are employed to find hidden, previously unknown, non trivial but potentially useful information or pattern from the data stored in large databases (Chen, et al., 1996; Pujari, 2002). Such large databases can exist in the form of Data Warehouse. The data ware house designed here can thus be utilized for mining library data. Data mining techniques can now be applied to the library transaction management system to classify the resources (mainly books) as well as to discover other rules, similarities, dissimilarities, correlations, clusters etc. based on their usage patterns obtained from the records of the transaction stored in the Data Warehouse. However, before mining the data there is need to pre process the data.

The steps involved in preprocessing of the data are (Han and Kamber, 2002):

1. Data Cleaning (to remove the noise or irrelevant data)

2. Data integration (where multiple data source are combined)
3. Data transformations (where data are transformed or consolidated into forms appropriate for mining by performing summary and aggregation operations)

There are various steps for mining data obtained from the warehouse to obtain various useful patterns. These are:

Concept/class description: characterization and discrimination

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions or sometimes concept hierarchies (Chen et al., 1996). These descriptions can be derived via (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms, or (2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

Association Rules

Association analysis concerns with the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis (Chen et al., 1996; Pujari, 2002; Agarwal, 1993). More formally, association rules are of the form $X \Rightarrow Y$, i.e., " $A_1 \wedge A_2 \wedge A_3 \wedge \dots \wedge A_m \Rightarrow B_1 \wedge B_2 \wedge B_3 \wedge \dots \wedge B_n$ " where A_i (for $i \in \{1, \dots, m\}$) are attribute-value pairs (Han and Kamber, 2002). The association rule $X \Rightarrow Y$ is interpreted as "database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y" based on certain input support and confidence threshold (Hipp et al., 2000; Agarwal and Srikant, 1994). Here, for example, Association rule mining technique can be applied to find the trends of reference books that are used along with text books for particular session and many other

association patterns of book usage reflecting subscribers' preferences or choices based on certain threshold.

Classification and prediction

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks [Jiwei-Han and Micheline Kamber (2002, 2006)] [Ming-Syan Chen, Jiawei Han, Phillip S. Yu. (December 1996)] [Pujari, A. K. (2002)].

Clustering analysis

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters (Han and Kamber, 2006; Chen et al., 1996; Pujari, 2002). For mining of Association rules from the library transaction data there are various techniques and algorithms. Among the most popular ones are a priori algorithms and its variations, FP Growth etc. Some implementation issues of these are discussed in (Deva Sarma and Deva Sarma, 2004). An elaborate treatment of the FP Growth algorithm can be found in (Han and Kamber, 2002).

Evolution and deviation analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior

changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis. In the analysis of time-related data, it is often desirable not only to model the general evolutionary trend of the data, but also to identify data deviations which occur over time (Han and Kamber, 2002; Pujari, 2002).

Inference/Results

Based on the above design a prototype system is implemented for the mining of various trends and patterns. The Data Warehouse has been implemented in SQL and various preprocessing routines are implemented.

Future Work

The future works that can be done in this system include incorporation of other mining techniques to find trends related to user preferences in general and subject to conditions such as seasonal, temporal etc. The warehouse schema can further be extended to include more functional areas and operations of library and so on.

Conclusion

This type of Data Ware Housing and Data Mining tools can support library management system with efficient techniques for decision support in connection with procurement of books and journals etc. Scope also exists to develop intelligent library transaction management system that can assist in making proper utilization of resources and provide optimum benefits to the subscribers.

References

- Deva Sarma, P. K.; Deva Sarma, H. K. (2004). "Efficiency Enhancement of Apriori Algorithm", Proceeding of International Conference on Complex Systems Intelligence and Modern Technological Application (CSIMTA), Schrborough, France, Sept-2004.
- Jiwei, H.; Micheline, K. (2002). "Data Mining Concepts and Techniques", Morgan Kaufmann publisher.
- Jochen, H.; Ulrich, G.; Gholamreza, N. (2000). Algorithms for Association Rule Mining – A General Survey and Comparison. *SIGKDD Explorations*. 2: 58 – 64
- Ming-Syan, C.; Jiawei, H.; Phillip, S. (1996). Data Mining: An Overview form a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*. 8: 866-883
- Pujari, A. K. (2002), "Data Mining Techniques", Universities Press, First Edition.
- Agrawal, R.; Srikant, R. (1994). "Fast Algorithm for Mining Association Rules in Large Databases", Proceedings of 20th International Conference on Very Large Databases, Santiago, Chillie. 478 – 499
- Agrawal, R.; Imielinski, T.; Swami, A. (1993). "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of ACM SIGMOD, International Conference on Management of Data, 207 – 216
- Ramez, E.; Shamkant, B.; Navathe, D.; Somayajulu, V.L.N.; Gupta, S.K. (2006). "Fundamentals of Database Systems", Pearson Education.
- Vincent, R. (2005). "Building a data warehouse with examples in SQL server", Apress publication.