DATA MINING TECHNIQUES FOR DYNAMICALLY CLASSIFYING AND ANALYZING LIBRARY DATABASE

ROOPESH KUMAR DWIVEDI R P BAJPAI

Abstract

Huge amount of data and information is originating in the information era. Library automation can provide some relief, but data mining techniques have to be used for dynamically analyzing the library database and to make strategic decisions for managing the library in an efficient manner. Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Practical data mining can accomplish a limited set of tasks and only under limited circumstances. For library, it can play an important role by dynamically analyzing library database especially data related to the acquisition and circulation. No single data mining tool and technique is equally applicable. In commercial application, data mining is usually employed on very large database. This paper gives the clear picture of some of the most common association rule data mining techniques which can be applied to the library database and it outcomes.

Keywords: Data Warehousing/Data Mining/ Knowledege Discovery Database/ Clustering/ Decision Tree/ Neural Network / Association Rule.

1. Introduction

Data mining which emerged during the late 1980s, has great strides during 1990s. In the last few years, almost in every meeting which has anything to do with the databases, neural networks, genetic algorithms, e-commerce, or artificial intelligence has had a theme or session on data mining and data warehousing. Since the data mining is a multidisciplinary field, the researchers from different disciplines are gradually getting attracted to work in this new frontier of research. Vast amount of operational data are routinely collected and stored away in the archives of many organizations. In the last five years, there has been a tremendous improvement in hardware leading to new computer programs which shift massive amount of operational data, recognize pattern and provide hints to formulate hypothesis for tactical and strategic decision making that can now be executed in a reasonable time.

Moreover libraries are also generating large volume of data. The challenge is to analyze massive amount of data to arrive at meaningful conclusions in reasonable time period. Data mining is one of the most important step in Knowledge Discovery Database (KDD) process and no single data mining technique and algorithm is best for all type of data mining tasks. Data mining tasks can be divided into six broad categories-Classification, Estimation, Prediction, Clustering, Association and Description. Technique like clustering, decision tree and neural network are known in pattern recognition. However, these algorithms are not suitable for the purpose of mining and hence new algorithms and techniques had to be proposed.

For applying data mining process in library database, first of all data from the various sources must be gathered and organized in a useful and consistent way. This is called data warehousing. Though creating a data warehouse is not a pre-requirement for data mining but to get full use of data mining techniques, it is recommended that before applying data mining techniques, first create a data warehouse. The data warehouse provides the enterprise with memory. But memory is of little use without intelligence. This will come through data mining. Although there are so many data mining techniques exists for the last many years or decades, it is only in the last few years that commercial data mining techniques has caught up in a big way.

2. Data Mining Techniques for Library Database.

Some of the popular data mining techniques which are applicable for library databases are divided into the Traditional Techniques (Statistics, Neighborhood and Clustering) and New Generation Techniques (Decision Tree, Neural Network and Association Rule).

2.1 Traditional Techniques

The main techniques that are used 99.9% of the time on existing business problems can be used for mining library databases as well. These cover only those techniques that work consistently are equally useful for library databases and are understandable and explainable.

2.1.1 Statistics : By strict definition a statistics or statistical techniques are not data mining techniques. They were being used even before the term data mining was coined to apply to business application. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. For a data mining problem, one has to solve it with statistical methods or to use other data mining techniques. Hence, it is important to have some idea of how statistical techniques work and how they can be applied.

Difference between statistics and data mining: the techniques used in data mining are successful for precisely the same reasons that statistical techniques are successful. So what is the difference? Why aren't we as excited about statistics as we are about data mining? There are several reasons. The first is that the statistical data mining techniques such as CART, Neural Networks and Nearest Neighbor techniques can not be used by less expert users. The other reason is that the time is right. Because of the use of computers for closed loop business data storage and generation, there now exists large quantities of data that is available to users. Likewise the fact that computer hardware has dramatically improved in order of magnitude in storing and processing the data makes some of the most powerful data mining techniques feasible today.

The list of the most frequently used summary statistics normally available in library automation include:

- Max- the maximum value for given predictor, for example maximum frequent book, maximum frequent use,r etc.
- Min- the minimum value for a given predictor, for example minimum fine collection each day/month/year, minimum number of books issued/ returned each day, etc.
- Mean- the average value for a given predictor, for example average user in a day/ month/year, etc.
- Median- the value for a given predictor that divides the database as nearly as possible into two databases of equal number of records.
- Mode- the most common value for the predicto,r for example predicting the future requirement of the user .
- Variance- the measure of how spread out the values are from the average value.

2.1.2. Clustering and Nearest Neighbor: Clustering and Nearest Neighbor prediction techniques are among the oldest techniques used in data mining. Clustering means the records are grouped or clustered together. Nearest neighbor is a prediction technique that is quite similar to clustering in order to predict what a prediction value is in one record look for records with similar predictor values in historical database and use the prediction value from the record that it is nearest to the unclassified record. The nearest neighbor prediction algorithm simply stated is: objects that are near to each other will have similar prediction values as well. Thus if you know the prediction value of one of the objects you can predict it for its nearest neighbors. Since Clustering and Nearest Neighbor prediction technique can be used to dynamically classify the records, hence it gives the solid foundation for the new computerize classification scheme to replace the classical/manual classification scheme like DDC, CC, etc. But before using such techniques for classification a set of base classes like those of DDC or CC is required and a set of appropriate subject headings have to be identified.

One of the improvements that is usually made to the basic nearest neighbor algorithm is to take a vote from the nearest neighbors rather than relying on the sole nearest neighbor to the unclassified record. The distance of the nearest neighbor provides a level of confidence. If the neighbor is very close or an exact match then there is much higher confidence in the prediction than if the nearest record is great distance from the unclassified record. The degree of homogeneity amongst the predictions within the nearest neighbors can also be used. If all the nearest neighbors make the same prediction then there is much higher confidence in the prediction than if half the records made one prediction and other half made another prediction.

Clustering is the process of grouping physical or abstract objects into classes of similar objects is called clustering or unsupervised classification. Clustering is the method by which like records are group together. Usually this is done to give the end user a high level view of what is going on in the database. Two of these clustering system are the PRIZM system from Claritas corporation and Micro Vision from Equifax corporation.

Hierarchical and Non-Hierarchical Clustering: There are two main types of clustering techniques, those that create a hierarchy of clusters and those which do not. The hierarchical clustering techniques create a hierarchy of clusters from small to big. The main reason for this is that clustering is an unsupervised learning technique. The hierarchy of clusters is usually viewed as a tree where the smallest clusters merge together to create the next highest level of cluster and those at that level merge together to create the next highest level of clusters.This hierarchy of clustering is created through the algorithm that builds the clusters. There are two main type of clustering algorithms:

- Agglomerative clustering techniques start with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest to each other are merged together to form the next largest cluster. This merging is continued until a hierarchy of clusters is built with just a single cluster containing all the records at the top of the hierarchy.
- Divisive clustering techniques take the opposite approach from agglomerative techniques. These techniques start with all the records in one cluster and then try to split that cluster into smaller pieces and then in turn to try to split those smaller pieces.

Out of these two techniques, the agglomerative techniques are the most commonly used for clustering and have more algorithms developed from them.

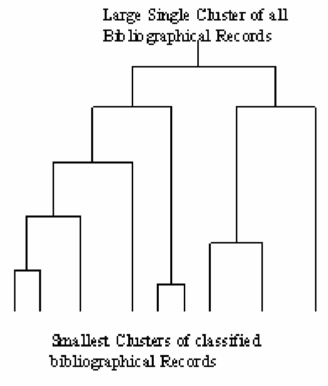


Fig.- Diagram showing a hierarchy of clusters. Clusters at the lowest level are merged together to form largest clusters at the next level of the hierarchy.

Fig.- Diagram showing a hierarchy of clusters. Clusters at the lowest level are merged together to form largest clusters at the next level of the hierarchy.

Non-Hierarchical Clustering: There are two main non-hierarchical clustering techniques. Both of them are very fast to compute on the database but have some drawbacks. The first are the single pass methods. They derive their name from the fact that the database must only be passed through once in order to create the clusters (i.e., each record is read from the database only once). The other classes of techniques are called reallocation methods. They get their name from the movement or reallocation of records from one cluster to another in order to create better clusters. The reallocation techniques do use multiple passes through the database but are relatively fast in comparison to the hierarchical techniques.

2.2 New Generation Techniques

The data mining techniques in this section represent the most often used techniques that have been developed over the last two decades of research in data mining and its use in the field of library database.

2.2.1. Decision Trees : The older decision tree techniques such as CHAID are highly used but the new techniques such as CART are gaining wider acceptance. A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Decision tree can be used for exploration analysis, data preprocessing and prediction work. The process in decision tree algorithms is very similar when they build trees. These algorithms look at all possible distinguishing questions that could possibly break up the original training dataset into segments that are nearly homogeneous with respect to the different classes being predicted. Some decision tree algorithms may use heuristics in order to pick the questions or even pick them at random. CART picks the questions in a much unsophisticated way as it tries them all. After it has tried them all, CART picks the best one, uses it to split the data into two more organized segment and then again ask all possible questions on each of these new segment individually. Most decision tree algorithm stop growing the tree when one of the three criteria are met:

1. The segment contains only one record. Hence there is no further question that you could ask which could further refine a segment of just one.
2. All the records in the segment have identical characteristics. There is no reason to continue asking further question segmentations since all the remaining records are the same.
3. The improvement is not substantial enough to warrant making the split.

CART: CART stands for Classification and regression trees and is data exploration and prediction algorithm developed by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone and is nicely detaild in their book entitled-Classification and Regression Tree.

CHAID: Another equally popular decision tree technology to CART is CHAID or Chi-Square Automatic Interaction Detector.

2.2.2. Neural Networks : To be more precise with the term neural network one might better speak of an artificial neural network. True neural network are biological system that detects patterns, make predictions and learn. Artificial neural network derive their name from their historical development which started off with the premise that machines could be made to think if scientists found ways to mimic the structure and functioning of the human brain on the computer. A neural network is loosely based on how some people believe that the human brain is organized and how it learns. There are two main structures of consequence in the neural network.

1. The node- which loosely corresponds to the neuron in the human brain.
2. The link- which loosely corresponds to the connections between neurons in the human brain.

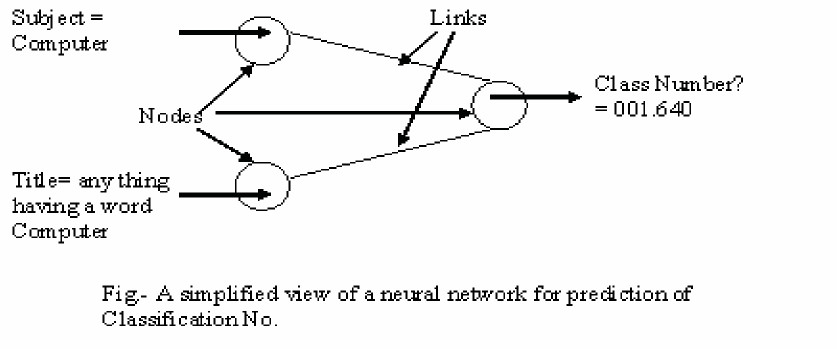


Fig.- A simplified view of a neural network for prediction of Classification No.

In this case the network takes the values for predictors for subject and title of the book and predicts classification number of the book. In order to make prediction, the neural network accepts the values for predictors on what are called the input nodes. These become the values for those nodes, those values are multiplied by values that are stored in the links. These values are then added together at the node at the far right (the output node), a special threshold function is applied and the resulting number is the prediction. In this case the resulting number is 001.640.

Neural network can be used for clustering, outlier analysis, feature extraction and prediction work. There are literally hundreds of variations on the back propagation feed forward neural networks. There are, however, two other neural network architectures that are used often. Kohonen Feature Maps are often used for unsupervised learning and clustering and Radial Basis Function Networks are used for supervised learning and in some ways represent a hybrid between nearest neighbor and neural network classification.

3. Association Rule Mining

The problem was formulated by Agrawal et al. in 1993 and is often referred to as market-basket problem. In this problem, we are given a set of items where items can be referred as books and a large collection of transactions (i.e., issue/return) which are subsets (baskets) of these items/books. The task is to find relationship between the presence of various items within these baskets.

Let A={B1,B2¦..Bm} be the set of books. Let T, the transaction (issue / return) database, where each transaction t is a set of items, then an association rule is an expression of the form X=>Y, where X and Y are the subset of A and X=>Y holds with confidence c if c% of transaction in T that support X also support Y. The rule X=>Y has support s in the transaction set T if s% of transactions in T supports X U Y.

Mathematically support and confidence of rule X=>Y can be defined as Support(X=>Y) = P (X U Y) and Confidence (X=>Y) = P (B/A) where P denotes the probability function.

In fact association rule mining is a two-step process:

1. Find all frequent itemsets / booksets: by definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.
2. Generate strong association rules from the frequent itemsets: by definition, these rules must satisfy minimum support and minimum confidence.

Association rule can be classified in various ways, based on the following criteria:

- Based on the type of values handled in the rule: if a rule concern associations between the presence and absence of items, it is a Boolean association rule, for example

  Computer fundamental => Programming in C [support=5% and confidence=75%]

  If a rule describes associations between quantitative items or attributes, then it is a quantitative association rule. In these rules, quantitative values for items or attributes are partitioned into intervals. For example

  IssueCount (X, 6¦10) => Subject (X, Computer) [supp.=2% and conf.=90%]

- Based on dimensions of data involved in the rule: if the items in association rule reference only one dimension, then it is a single dimensional association rule. Examples given above are single dimensional association rules. If a rule reference two or more dimensions then it is a multidimensional association rule.

- Based on levels of abstractions involved in the rule set: some methods for association rule mining can find rules at differing levels of abstraction. For example

  IssueCount (X, 6¦10) => Subject (X, Computer Algorithm)

  IssueCount (X, 6¦10) => Subject (X, Computer)

Here in the above rules, the items bought are referred at different levels of abstraction. (e.g. Computer is a higher level of abstraction of Computer Algorithm). We refer to the rule set mined as consisting of multilevel association rule. If, the rules within a given set do not reference items at different levels of abstraction, then the set contain single-level association rules.

In rule induction systems, the rule itself is of the form of this and this then this. In order for the rules to be useful there are two pieces of information that must be supplied as well as the actual rule:

1. Support- - How often does the rule apply?
2. Confidence- How often is the rule is correct?

Rule	**Support**	**Confidence**
In a semester if a book “Computer Fundamental” was issued to a member then a book “Programming in ‘C’” also issued to the same member.	5%	75%
In a semester if Issue-Count of a book is between 6 to 10 then it is of subject computer.	2%	90%

4. Conclusion

For efficiently and effectively doing the library administration and extending library services, the need of library automation and digital library occur. But simply automating the library or developing digital library is not the only solution unless and until we are not able to explore the hidden information from the large amount of database. This can be done by applying the data mining in the library database. Now we can take a glance on the possibilities opening in the new age of data mining in the field of library and information science.

1. Dynamic Classification of Bibliographical Data of Library: For this purpose, clustering, nearest neighbor, neural network, decision tree data mining techniques are very useful.
2. Sequence Analysis: By using statistical analysis to identify unlinked document that library members are likely to read together. It examines the path that users follow while searching for information and can help to identify which documents users are likely to use together.
3. Association Analysis: By applying association rule mining techniques and algorithms like Aprori, Partition, Pincer-Search, Dynamic Itemset Counting, FP-tree Growth and many more new algorithms for finding association rules one can take advantage of these rules in taking strategic decisions for library management.

References

1. Agrawal, R. and Srikant, R. "Fast algorithm for Mining Association Rules", Proceeding 1994: International Conference on Very Large Database, , Santiago, Chile, Sept. 1994, pp. 487-499.

2. Agrawal, R. and Srikant, R. "Mining Sequential Patterns", Proceeding 1995: International Conference on Data Engineering, Taipei, Taiwan, March 1995, pp. 3-14.

3. Berson, A., Smith, S. and Thearling, K. "An Overview of Data Mining Techniques", White Paper from Internet, 2005.

4. Chen, M.S, Han, J. and. Yu, P.S. "Data Mining: An Overview from a Database Perspective", IEEE Transaction on Knowledge and Data Engineering, Vol.8, 1996, pp. 866-883.

5. Dwivedi, R.K. and Bajpai, R.P. "Use of Data Mining in the field of Library and Information Science: An Overview", Proceedings of second International CALIBER 2004. 2004, pp. 512-519.

6. Dwivedi, R.K. "Critical Study of Data Mining Technique for Un-Directed Knowledge Discovery". International Journal of Natural Sciences & Technology, Vol.1 (1), 2006. pp. 125-130.

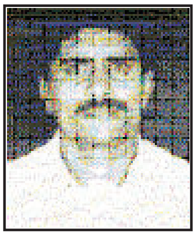
7. Hand, J.D., Mannila , H. and Smith, P. "Principles of Data Mining" , MIT Press, 2000.

8. Han, J and Fu, Y. "Mining Multiple-Level Association Rules in Large Databases", IEEE Transaction on Knowledge and Data Engineering, Vol.11, 1999, pp. 798-805.

BIOGRAPHY OF AUTHORS

Mr. Roopesh Kumar Dwivedi holds MCA and doing PhD in Computer Science. Presently he works as Information Scientist in the University Library of Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, Chitrakoot - 485331, Satna (M.P.). Contributed 7 papers.

Email: rkdwivedi_mgcgv@rediffmail.com



Dr. Raghubansh Prasad Bajpai holds MSc, MLIS, Ph.D.(LIS). Presently works as University Librarian (I/c) and Lecturer and Head, Library and Information Science, Mahatma Gandhi Chitrakoot Gramodaya Vishwavidyalaya, Chitrakoot - 48531, Satna (M.P.). About 12 years worked as Lecturer and About 10 Years as I/C University Librarian. Contributed 10 research papers and two Books.

Email: rpbajpai_mgcgv@redifmail.com