The Design and Implementation of Collaborative Filtering in Data Mining

Wen-Chih Chen* Lu-Ping Chang Hsin-Te Shih Shu-Ling Hsiao
Advanced e-Commerce Technology Lab., Institute for Information Industry, ROC
Phone no : 886-2-23770566,,,,328
Fax no : 886-2-23776194
E-mail : wjchen@iii.org.tw

Abstract

Data mining is the process of discovering explicit knowledge from large amounts of data stored in database, data warehouse or other repositories. There have been many studies about models of data mining such as association rule, sequential pattern and so on. Collaborative filtering is one of data mining models. In this paper, we propose two approaches to solving the mining process of collaborative filtering. Finally, collaborative filtering mining is applied to Knowledge Management system.

Key Word: Data Mining, Collaborative filtering, Knowledge Management

Introduction

Since 1970s, database technology has been developed to manage and maintain large amounts of data. Relational database system has been wildly applied in business applications from 1980s. Data Mining means the mechanism of Mining potential useful knowledge from large amounts of data[10]. Instead that Data warehouse help business executives organize, understand their data to make strategic decision. And On-Line Analytical Processing(OLAP) provides Multidimensional data view[2]. Mining potential useful knowledge can help different business domain do a best marketing and control business risk[9].

The technology of database has evolved from information processing to analytical processing and from analytical processing to Data mining[5]. Data mining supports knowledge discovery in database by finding hidden patterns, performing classification, making prediction, estimating the results and constructing association analytical model[8].

Related Work in Data Mining

There are different mechanisms in Data Mining: association rule, sequential pattern, classification, clustering, outlier mining, and regression analysis.

Mining Association rule means finding interesting relationship among items in a given data set by giving support and confidence value[3]. A typical application of association rule is market basket analysis. For instance, if customers buy this brand of

bread, how likely are they to also buy certain brand of milk. The problems of mining association rule have been firstly solved by Apriori algorithms. But the time complexity of Apriori algorithms is $O(N^m)$. Many studies in solving association rule tried to improve the time complexity[4]. Such algorithms are FP-tree, DHP and etc.

Mining sequential pattern was to discover time series knowledge[1]. The input data is a set of sequences, called data-sequences. Each data-sequence is a list of transactions, where each transaction is a set of literals, called items. Typically there is a transaction-time associated with each transaction. A sequential pattern also consists of a list of sets of items. The problem is to find all sequential patterns with a user-specified minimum support. For example, in the book-club, a sequential pattern might be 5% of customers bought "Foundation", then "Foundation and Empire", and then "Second Foundation". The data-sequence corresponding to a customer who bought some other books in between these books still contains this sequential pattern; the data-sequence may also have other books in the same transaction as one of the books in the pattern. The results apply to many scientific and business domains. For instance, in the medical domain, a data-sequence may correspond to the symptoms or diseases of a patient, with a transaction corresponding to the symptoms exhibited or diseases diagnosed during a visit to the doctor. The patterns discovered using this data could be used in disease research to help identify symptoms or diseases that precede certain diseases.

Classification is one of the data analysis forms that can be used to extract models describing important data classes or to predict future data trends. For example, a classification model may be built to categorize bank loan applications as either safe or risky. There are many approaches for data classification model building, including "decision tree induction", "Bayesian classification", "neural networks", "case-based reasoning", "genetic algorithms" etc.

Unlike classification, the class label of each object is unknown in cluster analysis[6]. Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Similarity are measured based on the attribute values describing the objects. Often distance measures are used. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, and market research. Clustering analysis is organized into the following categories: partitioning methods, hierarchical methods, density-based method, grid-based methods, and model-based methods.

For many KDD applications, such as detecting criminal activities in E-commerce, finding the rare instances or the outliers, can be more interesting than finding the common patterns. "What is an outlier?" Very often, there exist data objects that do not

comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers. Related to outlier detection is an extensive body of work on clustering algorithms. From the viewpoint of a clustering algorithm, outliers are objects not located in clusters of a dataset, usually called noise. Outlier mining has wide applications. It can be used in fraud detection, for example, by detecting unusual usage of credit cards or telecommunication services. In addition, it is useful in customized marketing for identifying the spending behavior of customers with extremely low or extremely high incomes, or medical analysis for finding unusual responses to various medical treatments. The methods for outlier detection can be categorized into three approaches: "statistical approach", "distance-based approach", and "deviation-based approach".

Collaborative Filtering Architecture

Collaborative Filtering recommend people for something by finding the similar behavior in the threshold ratio of users or calculating the similar ratings in all users[7][11][12]. Collaborative Filtering comprises the behavior and the rating of relish model in our study.

Behavior model

The behaviors of certain of consumers have fixed patterns among groups of consumers. Consider below simple matrix, in which rows stand for the consumer id and columns stand for the item. The cell of matrix represents whether the consumer bought the item.

Table I Consumers/Item matrix

Consumer/Item	Beer	Bread	Milk	Chips	Soda
Peter	1	1	0	1	0
Rita	1	1	0	1	1
John	0	1	1	1	0
May	1	1	0	1	1
Eric	1	1	1	0	1

In our study, three parameters are given to solve the behavior model in collaborative filtering. First parameter named Support means that the number of the same behavior consumers meet the minimum Support value. Second parameter named Length means that meet the minimum number of behaviors of the consumer. Third parameter named Difference means that maximum number of difference behavior between different consumers must be satisfied. Therefore Let (Support, Length, Difference) to be (2,4,1), above matrix could be mining knowledge. The length of all consumers is calculated first. Length(Peter) equals three; Length(Rita) equals four; Length(John) equals three; Length(May) equals four; finally Length(Eric) equals four. Because Length is set to be "four", the patterns of Rita, May and Eric to be mining

first. Then Support is set to be two. The behaviors of Rita and May are the same. The value of Support is calculated to meet minimum Support. Finally, the difference is set to be "one", Difference(Peter,Rita) equals "one". The behaviors of Peter meet the maximum difference. But Difference(John,Rita) equals "three". The behaviors of John do not meet the maximum difference. The Collaborative filtering system will recommend Peter for Soda. Below algorithms describe behavior model of Collaborative Filtering.

Mining Behavior model of Collaborative Filtering Algorithms:

Input: Given a set of users, where each user consists of a set of items; a minimum Support S; a minimum Length L; a maximum Difference D

Output: finding all recommendations of all users that meet (S,L,D)

Procedure Mining_CF_Behavior(matrix,S,L,D)

Calculate all users' length

Find the candidate users that meet minimum L

From the candidate users, find the candidate patterns that meet minimum S

From candidate patterns, find all recommendations of all users meet maximum D

End Procedure

Rating of relish model

Rating of relish model based on the subjective rating or evaluations of others. Virtual community has been more and more popular in internet. The core idea is to group people who have similar interests or similar evaluations in virtual community. The questionnaire is a common way to understand how much people like subject. According to different scores by different evaluations of others, the fond degree of specific subject will be predicted. For instance, a record store site has history data about the fond degree of specific subject by different members. As table II described. The value of the cell in this matrix range from one to ten. The higher value means much preferences for subject.

Table II member/music matrix

Consumer/music	Folk music	Classical	Pop music	Jazz	Rock music
Peter	4	5	3	4	
Rita	4	4	3	4	6
John	3	5	2	5	5
May	4	5	3	4	6
Eric	3	4	3	5	5

Assume **Peter** is the active user, we want to make some recommendations for him at Rock music. From the Table II, we found other four people, Rita, John, May and Eric, have different ratings with Peter on the Music (Folk music, Classical, Pop music, Jazz) that Peter rated, and these people all rated music Rock music. So we can

predict the rating that Peter will give to the music Rock music according to these four people's rating on this music. The most popular algorithms used in the collaborative filtering are Correlation or Vector Similarity. Using this kind of algorithms, firstly, correlation coefficients of appropriate users are computed based on their rating similarity to the active user. And then a threshold is set for choosing a subset of appropriate users based on their correlation coefficients with the active user. Finally, weighted aggregate of their ratings is used to generate predictions for the active user.

First, we compute correlation coefficient, weights between -1 and 1 that indicate how much Peter tended to agree with each of the others on those articles that they both rated. For example, Peter's correlation coefficient with Rita is computed as:

$$r_{PR} = \frac{\sum_i (P_i - \bar{P})(R_i - \bar{R})}{\sqrt{\sum_i (P_i - \bar{P})^2}\sqrt{\sum_i (R_i - \bar{R})^2}} = \frac{0 + 0.25 + 0.75 + 0}{\sqrt{2}\sqrt{3 \times (0.25)^2 + (0.75)^2}} = 0.816$$

In the formula above, $\bar{P}$ is the average of Peter's ratings. All the summations and averages in the formula are computed only over those articles that Peter and Rita both rated. Similarly, Peter's correlation coefficient with John, Mary and Eric are 0.816, 1 and 0.426.

Second, we set threshold 0.7 for choosing a subset of appropriate users based on their correlation coefficients with the active user. We get Rita, John and Mary. Finally, to predict Peter's score on the Rock music in the matrix, take a weighted average of all the ratings on Rock music according to the following formula:

$$P_{Pred} = \bar{P} + \frac{\sum_{J \in raters} (J_{Rockmusic} - \bar{J}) r_{PJ}}{\sum_J |r_{PJ}|}$$

$$= 4 + \frac{2.25 r_{PR} + 1.25 r_{PJ} + 2 r_{PM}}{|r_{PR}| + |r_{PJ}| + |r_{PM}|} = 5.76$$

Case Study

Knowledge Management System has been developed in our study. Besides basic function like bulletin board, document management in KM, KM systems have been more intelligent by using the mechanism of collaborative filtering. Most organizations construct Knowledge Management System for colleagues to upload documents and share knowledge.

The applicable scenarios in Knowledge Management combined with collaborative filtering are as follows. In KM system, Peter searched terms "Data

Mining", "Agent", "CRM" and Rita searched terms "Clustering", "Text Mining", "CRM" and so on. As table III described. If (S,L,D) is given (2,3,1), the system would find support pattern "Peter" and "May" and give Eric recommendation for searching term "Agent" by the behavior model of collaborative filtering.

Table III Users Search terms in KM system

User/Search	Data Mining	Agent	Clustering	Text Mining	CRM
Peter	1	1	0	0	1
Rita	0	0	1	1	1
John	1	1	1	1	0
May	1	1	0	0	1
Eric	1	0	0	0	1

Another scenario in KM system is the evaluation of specific document by members. As table iv described. For instance, a member named Peter viewed documents and gave his evaluations on specific document. According to correlation coefficient algorithms, the predictive score of Rita's preference in Agent.ppt can be calculated.

Table IV users/document in KM system

User/document	Mining.doc	Agent.ppt	KM.doc	BI.ppt	CRM.doc
Peter		1		9	2
Rita	8		3		1
John	5		8	3	9
May	3	5	7		74
Eric		7	2		1

Conclusion

Data Mining has played an important role in database and business marketing. In this paper, the model of collaborative filtering has been divided into behavior model and rating model. The behavior model can be used in retailer business, KM recommendation system. The rating of relish model can be used for building virtual community, in which people have high similarity about specific subject. The core spirit of collaborative filtering is to group people who have same behaviors or similar evaluations. Depend on these evaluations, the future behavior of someone will be predicted or recommended by system.

By predicting user's preference for specific subject, many business applications such as market segmentation, direct mail and market basket analysis are used for reducing business cost.

Acknowledge

This research was supported by the Software Technology for Advanced Network Application project of Institute for Information Industry and sponsored by MOEA , ROC

References

[1] R. Agrawal and R. Srikant. Mining Sequential Patterns. In Proc. of the 11th Intel Conference on Data Engineering, Taipei, Taiwan, March 1995.

[2] S.Sarawagi, R.Agrawal, and N.Megiddo. Discovery-driven exploration of OLAP data cubes. In Proc. Int. Conf. Very Large Ba ta Bases(VLDB' 96), Bombay, India, Sept. 1996

[3] R. Srikant, Q.Vu, and R. Agrawal. Mining association rules with item constraints. In Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining(KDD' 97), Newport Beach, CA, Aug.1997

[4] N. Pasquier, Y. Bastide, R.Taouil, and L.Lakhal. Discovering frequent closed itemsets for association rules. In Proc. 7th Int. Conf. Database Theory(ICDT' 99), Jerusalem, Israel, Jan. 1999.

[5] T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM,1996

[6] S. Guha, R. rastogi, and K.Shim. Cure: An efficient clustering algorithm for large databases. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data(SIGMOD' 98) Seattle, WA, June 1998

[7] Herlocker,J.L., Konstan,J.A., Borchers,A.,Riedl,J. An algorithmic framework for performing collaborative filtering. *Proceedings of the 1999 Conference on Research and Development in Information R trieval*

[8] U.M. Fayyad and R. Uthurusamy, editors. Proc. 1st Int. Conf. Knowledge Discovery and Data Mining(KDD' 95). Montreal, Canada, Aug. 1995. AAAI Press.

[9] B.-K. Yi, N. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. In Proc. 2000 Int. Conf. Data Engineering(ICDE' 00), San Diego, CA, Feb. 2000.

[10] S.M. Weiss and N. Indurkhya. Predictive Data Mining. San Francisco: Morgan Kaufmann, 1988

[11] Goldberg, D., Nichols, D.,Oki, B. M., Terry D.(1992)"Using collaborative filtering to weave an information tapestry." Comm. ACM, 35(12),pp.61-70.

[12] Zacharia, G.,Maes, P.(1999) "Collaborative Reputation Mechanisms in Electronic Marketplaces. "Proc. 32nd Hawaii International Conf. On System Sciences.