Mining generalized fuzzy association rules from web taxonomic structures

Yi-Tsung Tang
NAN HUA University
eric_tang630@hotmail.com

Hung-Pin Chiu
National Taitung University
hpchiu@nttu.edu.tw

Abstract

The discovery of fuzzy association rules is an important data-mining task for which many algorithms have been proposed. However, the efficiency of these algorithms needs to be improved to handle real-world large datasets. In this paper, we present an efficient method named cluster-based fuzzy association rule (CBFAR) to discover generalized fuzzy association rules from web structures. The CBFAR method is to create fuzzy cluster tables by scanning the browse information database (BIDB) once, and then clustering the browse records to the *k*-th cluster table, where the length of a record is *k*. The counts of the fuzzy regions are stored in the Fuzzy_Cluster Tables. This method requires less contrast to generate large itemsets. The CBFAR method is also discussed.

Keyword：Fuzzy data mining; association rules

摘要

模糊關聯法則的挖掘是資料挖掘(Data Mining)中一個重要的部分，也有許多的方法相繼被提出。然而，這些演算法對於處理實際資料上的效率仍然有改進的空間。本研究提出了一個有效率的方法（Cluster-Based Fuzzy Association Rule:CBFAR）來從網頁架構中找出模糊關聯法則，並改進模糊關聯法則挖掘的處理效率，此方法以分群表(cluster table)的關念來儲存瀏覽次數之模糊值，在大項目組的產生過程中，只需掃描瀏覽資料庫一次，也可以去除許多不必要的資料比對時間，有效的減少模糊關聯法則的處理時間，改進效率。

關鍵詞：模糊資料挖掘、關聯法則

1. Introduction

The discovery of fuzzy association rules is an important data-mining task. Association rules are used to discover the relationships, and potential associations, of items or attributes among huge data. These rules can be effective in uncovering unknown relationships, providing results that can be the basis of forecast and decision.

Deriving association rules from transaction database is most commonly seen in data mining. [2][4] It discovers relationships among items. In the past, Agrawal and Srikant proposed the Apriori association rule algorithm.[5] It can discover meaningful itemsets and construct association rules within large databases, but a large number of the candidate itemsets are generated from single itemsets. This method also needs to perform contrasts against all of the transactions, level by level, in the process of creating association rules. The database is repeatedly scanned to contrast each candidate itemset, that performance is dramatically affected.

After Agrawal et al. proposed the Apriori association rule, Tsay et al. have used cluster-based association rule (CBAR) approach.[8] This method used cluster-based table to reduce the number of database scans and requiring less contrast. Recently, the fuzzy set theory[3] has been used more and more frequently in intelligent systems. It's simplicity and similarity to human reasoning.[1] Hong et al. also proposed a fuzzy mining algorithm.[7] The items considered in their approach had hierarchical relationships. However, items in real-world applications are usually organized in some hierarchies. Mining multiple-concept-level fuzzy rules may lead to discovery of more general and important knowledge from data.

In this paper, we present a new method called cluster-based fuzzy association rule (CBFAR), for efficient fuzzy association rules mining. We considered the hierarchical relationships to discover the generalized fuzzy association rules from the browse information database (BIDB) and used the cluster-based concept to reduce the number of database scans. When the customer clicking the web pages, then the click times stored in the browse information database (BIDB).This method not only needs only one database scans, but also requires less contrast.

2. CBFAR Mining Framework

The hierarchical relationships and cluster-based concepts are used to discover generalized fuzzy association rules from browse information database (BIDB). We propose a CBFAR mining framework for discovering generalized fuzzy association rules. The proposed framework is shown in Fig. 1.

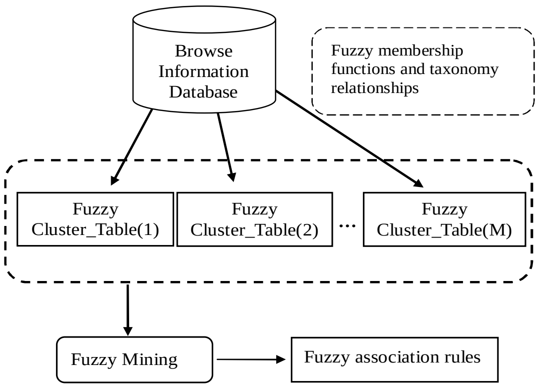

Figure1: CBFAR Mining Framework

We proposed mining framework maintains fuzzy association rules, and uses the hierarchical relationships and cluster-based fuzzy table to derive the fuzzy association rules. Previous studies on data mining focused on finding association rules on the single-concept level. However, relevant web page taxonomies are usually predefined in the web site service and can be represented using hierarchical trees.[6] Terminal nodes on the trees represent actual web pages appearing in networks structure; internal nodes represent classes or concepts of web pages formed by lower-level nodes.

[6]A simple example is given in Fig. 2.

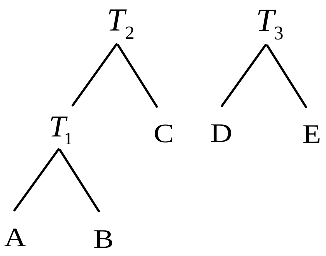

Figure2: An example of taxonomic structures

In this example, the T_2 web class falls into one class and one web page: T_1 and web page C. T_1 can be further classified into page A and page B. Similarly, assume T_3 are divided into page D and page E. Only the terminal web pages (A, B, C, D and E) can appear in browse information records. The CBFAR mining method is divided into four phases.

In the first phase, ancestors of web pages in each given browsed record are added according to the predefined taxonomy.

In the second phase, transform the quantitative value v_{ij} of each browsed data D_i (i=1 to n), for each expanded item name I_j appearing into a fuzzy set f_{ij} .The f_{ij} are represented as (f_{ij1}/R_{j1} + f_{ij2}/R_{j2} + $\cdots$ + f_{ij1}/R_{j1}) using the given membership functions, where h is the number of fuzzy regions for I_j . R_{jl} is the lth fuzzy region of I_j , $1 \le l \le h$, and f_{ijl} is v_{ij} 's fuzzy membership value in region R_{jl} . Calculate the value of each fuzzy region R_{jl} in the browsed data. ($count_{jl} = \sum_{i=1}^{n} f_{ijl}$)

In the third phase, creates M cluster tables. Scan the browse information database once and cluster the browsed data. If the length of browsed data is k, the browsed record and the fuzzy region value of items in this browsed record will be stored in the table, named Fuzzy_Cluster Table (k), $1 \le k \le M$, where M is the length of the longest browsed record in database.

In the fourth phase, the set of candidate itemsets C_n is generated. When the length of candidate itemset is k, the support is calculated with reference to the Fuzzy_Cluster Table(k). If the fuzzy region value of C_n is greater than or equal to the predefined minimum support value α , the candidate itemsets becomes the large itemsets, put C_n in the large itemsets L_n . Otherwise, it is contrasted with the Fuzzy_Cluster Table($k+1$). The large itemsets is $L_n = \{\max - R_j \mid \max - count_j \ge \alpha, 1 \le j \le m\}$. Until the large itemsets L_n is null, this process terminates when the calculated support is greater than or equal to the predefined minimum support or the the end of the Fuzzy_Cluster Table(M) has been reached. Finally, use the predefined minimum confidence value to discover fuzzy association rules. If the candidate fuzzy association rule is larger than or equal to the predefined confidence value, put it in the rule base.

3. An Example

In this section, an example is given to

illustrate the proposed mining method. This is a simple example to show how the proposed method can be used to discover fuzzy association rules from browsed data. There are six browsed records and five items (web pages) in a browse information database: A, B, C, D and E. An example browse information database is shown in Table 1. The taxonomy tree is shown in Fig. 3. The ancestors of appearing web pages are added to browsed records according to the predefined taxonomy tree. The expanded browsed record is shown in Table 2.

Table 1. Six browsed records in this example

BID	Items (Web Pages, Click times)
B1	(A,3) (C,4) (E,2)
B2	(B,3) (C,7) (D,7)
B3	(A,4) (B,2) (C,10) (E,5)
B4	(C,9) (E,10)
B5	(B,3)
B6	(B,8)(D,4)

Table 2: The expanded browsed records

BID	Items (Web Pages, Click times)
B1	(A,3) (C,4) (E,2) (T_1,3) (T_2,4) (T_3,2)
B2	(B,3) (C,7) (D,7) (T_1,3) (T_2,7) (T_3,7)
B3	(A,4) (B,2) (E,5) (T_1,6) (T_3,5)
B4	(C,9) (E,10) (T_2,9) (T_3,10)
B5	(B,3) (T_1,3)
B6	(B,8)(D,4) (T_1,8) (T_3,4)

In this example, assume that the fuzzy membership functions are the same for all the items and are as shown in Fig. 4. The fuzzy membership function is represented by three fuzzy regions: *Low(L)*, *Middle(M)* and *High(H)*, and three fuzzy membership

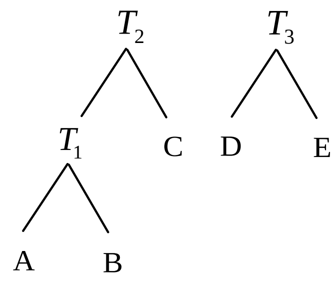

Figure3: Taxonomy tree in this example

values are produced for each item according to the predefined membership function.

The length of the longest browsed record in this database is six, and creates six fuzzy_cluster tables as shown in Table 3. The fuzzy region value of items in this browsed record will be stored in the Fuzzy_Cluster Tables.

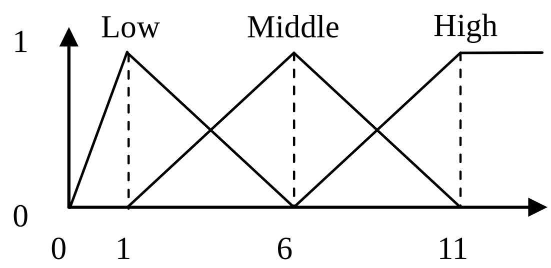

Figure4: The membership function

Assume the minimum support value is 2.0. We can discover the Large-1 itemsets (L_1) which is large than or equal to the predefined minimum support value according to the Fuzzy_Cluster Tables. The itemsets of L_1 are {B.Low = 2.0}, {T_1.Middle = 2.8}, {T_3.Middle = 2.6}.

Table 3: Fuzzy_Cluster Tables

BID	A	B	C	D	E	T_1	T_2	T_3

Fuzzy_Cluster Table(1)

NULL

Fuzzy_Cluster Table(2)

B5	0	L,0.6 M,0.4	0	0	0	L,0.6 M,0.4	0	0

Fuzzy_Cluster Table(3)

NULL

Fuzzy_Cluster Table(4)

B4	0	0	M,0.4 H,0.6	0	M,0.2 H,0.8	0	M,0.4 H,0.6	M,0.2 H,0.8
B6	0	M,0.6 H,0.4	0	L,0.4 M,0.6	0	M,0.6 H,0.4	0	L,0.4 M,0.6

Fuzzy_Cluster Table(5)

B3	L,0.4 M,0.6	L,0.8 M,0.2	0	0	L,0.2 M,0.8	M,1	0	L,0.2 M,0.8

Fuzzy_Cluster Table(6)

B1	L,0.6 M,0.4	0	L,0.4 M,0.6	0	L,0.8 M,0.2	L,0.6 M,0.4	L,0.4 M,0.6	L,0.8 M,0.2
B2	0	L,0.6 M,0.4	M,0.8 H,0.2	M,0.8 H,0.2	0	L,0.6 M,0.4	M,0.8 H,0.2	M,0.8 H,0.2

Generate the large 2-itemsets L_2. Combining the items of L_1 in order to generate candidate 2-itemsets C_2. The procedure is similar to the candidate generation of Apriori algorithm[5]. The itemsets of C_2 are {B.Low, T_1.Middle}, {B.Low, T_3.Middle}, {T_1.Middle, T_3.Middle}. In order to generate L_2, it is necessary to compute the fuzzy region values of each candidate itemset in the Fuzzy_Cluster Table(2). If the value is larger than or equal to the predefined minimum support value, put C_2 in the L_2. Otherwise, compute the fuzzy region values in the next cluster table (Fuzzy_Cluster Table(3)). Until the calculated support is greater than or equal to the predefined minimum support or the end of the Fuzzy_Cluster Table(M). The other large itemsets L_n are in the similar way.

Therefore, the large itemsets in this example are {B.Low}, {T_1.Middle}, {T_3.Middle},{T_1.Middle, T_3.Middle}. Then, we can transform each large itemsets into a fuzzy association rule.

4. Conclusions

In this paper, we have proposed a generalized fuzzy association rules mining framework for extracting fuzzy association rules from browse information database. The cluster-based fuzzy association rule (CBFAR) method creates Fuzzy_Cluster Tables to discover the large itemsets.

Contrasts are performed only against the partial Fuzzy_Cluster Tables that were created in advance. It only requires a single scan of the browse information database, and contrasts with the partial Fuzzy_Cluster Tables. This method not only needs only one database scans, but also requires less contrast.

In the future, we will continuously for the huge database, and discussing with the performance of CBFAR method.

5. References

[1] A. Kandel, Fuzzy Expert Systems, CRC Press, Boca Raton, FL, 1992, pp. 8-19.

[2] J. Han, Y. Fu, Discovery of multiple-level association rules from large database, The Internet. Conf. on Very Large Databases, 1995.

[3] L.A. Zadeh, Fuzzy sets, Inform. and Control 8(3), 1965 pp. 338-353.

[4] R.Agrawal, T. Imielinksi, A. Swami, Mining association rules between sets of items in large database, the 1993 ACM SIGMOD Conf., Washington, DC, USA, 1993.

[5] R. Agrawal, R. Srikant, Fast algorithm for mining association rules in large databasaes, Proceedings of 1994 International Conference on VLDB, 1994 pp. 487-499.
[6] R. Srikant, R. Agrawal, Mining generalized association rules, The Internat. Conf. on Very Large Databases, 1995.
[7] Tzung-Pei Hong, Kuei-Ying Lin, Shyue-Liang Wang, Fuzzy data mining for interesting generalized association rules, Fuzzy Sets and Systems, 2003 pp. 255-269.
[8] Yuh-Jiuan Tsay, Jiunn-Yann Chiang, CBAR: an efficient method for mining association rules, Knowledge-Based Systems, 2005 pp. 99-105.