Min-Apriori: An Algorithm for Finding Association Rules in Data with Continuous Attributes *

Eui-Hong (Sam) Han George Karypis, Vipin Kumar

Department of Computer Science and Engineering/Army HPC Research Center
University of Minnesota
4-192 EECS Bldg., 200 Union St. SE
Minneapolis, MN 55455, USA

{han,karypis,kumar}@cs.umn.edu

Last updated on December 31, 1997 at 1:14pm

*This work was supported by NSF ASC-9634719, by Army Research Office contract DA/DAAH04-95-1-0538, by Army High Performance Computing Research Center cooperative agreement number DAAH04-95-2-0003/contract number DAAH04-95-C-0008, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred. Additional support was provided by the IBM Partnership Award, and by the IBM SUR equipment grant. Access to computing facilities was provided by AHPCRC, Minnesota Supercomputer Institute. See **http://www.cs.umn.edu/∼han** for other related papers.

One of the important problems in data mining [SAD+93] is discovering association rules from databases of transactions, where each transaction contains a set of items. Several algorithms for finding association rules have been proposed. Most of the algorithms work on transaction data where each transaction contains a subset of items from the whole item set. The transaction data can be considered as binary where each transaction either contains a specific item or not. The example of this kind of data set includes market basket data. A typical store might have a couple of thousand items and a transaction corresponds to the set of items purchased by a customer.

When the data set contains continuous attributes, the existing algorithms do not work. For example, data might contain age and salaries of the customers in addition to the items the customers purchased. You may want to find out the associations between ages/salaries and items. In this case, the age and salary attributes need to be discretized before the association rules algorithms can be applied. Once these attributes are discretized, association rules like $Age < 30 \Longrightarrow Rollerblade$ can be discovered. The discretization and discovery of the association rules in this type of data is discussed in [SA96].

Another type of data contains entirely continuous attributes. For example, document data consists of set of documents and each document contain words. This original data is transformed into a table where each row corresponds to a word and each column corresponds to a document. Each entry (i, j) of this table corresponds to the frequency of word i in document j. From this table, user might want to find out association rules among different documents or words. Note that in this case the discretization does not apply as the user wants association among documents or words not among the ranges of frequencies each word or document is associated with.

In this paper, we propose a new algorithm to discover association rules in the type of data set discussed in the above paragraph.

Association rules capture the relationship of items that are present in a transaction [AMS+96]. Let T be the set of transactions where each transaction has values for each item in the item-set I. For example, consider a set of transactions from document data as shown in Table 1. The items set I for these transactions is $\{$*doc-1, doc-2, doc-3, doc-4, doc-5*$\}$ and each transaction corresponds to a word in the document data. Each value $T(i, j)$ in the table corresponds to the frequency of the word i in the document j. For example, *word-1* appears twice in *doc-1* and *doc-3*, once in *doc-5* and does not appear in *doc-2* and *doc-4*.

TID	doc-1	doc-2	doc-3	doc-4	doc-5
word-1	2	0	2	0	1
word-2	2	0	3	0	1
word-3	0	1	0	1	1
word-4	0	2	0	0	0
word-5	1	2	0	1	2

Table 1: Transactions from document data.

The first step of *Min-Apriori* is the normalization of values along the column of the table. We normalize the values such that each column adds up to 1.0. The normalized table is shown in Table 2.

Let C be a subset of I, then we define *support* of C with respect to T to be:

$$support(C) = \sum_{i \in T} min\{T(i, j) | j \in C\},$$

where $T(i, j)$ corresponds to the value in the normalized transaction table. For example, the support of $\{$*doc-1, doc-3*$\}$

TID	doc-1	doc-2	doc-3	doc-4	doc-5
word-1	0.4	0.0	0.4	0.0	0.2
word-2	0.4	0.0	0.6	0.0	0.2
word-3	0.0	0.2	0.0	0.5	0.2
word-4	0.0	0.4	0.0	0.0	0.0
word-5	0.2	0.4	0.0	0.5	0.4

Table 2: Normalized transactions table.

is $0.4+0.4+0.0+0.0+0.0 = 0.8$, whereas the support of $\{$*doc-1, doc-3, doc-5*$\}$ is $0.2+0.2+0.0+0.0+0.0 = 0.4$.

The above definition of support satisfies the monotonicity property of *support* defined for the frequent (or large) item set of binary values. In other words, based on the new support definition the support of any subset S of C is greater than or equal to the support of C. This is due to the monotonicity property of *minimum* function, which is if $A \subseteq B$ where A and B are sets of numbers, then the $min(A) \geq min(B)$. Thus, we have

$$\begin{aligned} support(S) &= \sum_{i \in T} min\{T(i, j) | j \in S\} \\ &\geq \sum_{i \in T} min\{T(i, j) | j \in C\} \\ &= support(C) \end{aligned}$$

As this new definition of support satisfies the monotonicity property, the majority of existing algorithms for finding association rules can be used to find association rules based on this new definition. We have implemented *Min-Apriori* based on the *Apriori* algorithm proposed in [AS94].

We have applied *Min-Apriori* algorithm to find association rules for web document data and used association rules to find cluster of words and documents [HBG+97, HKKM97]. The results show that we can find meaningful and useful association rules among documents and words without discretizing the values in the data. The results also suggest that whenever a higher value in the data indicates a stronger relation, *Min-Apriori* is applicable. We plan to apply *Min-Apriori* on different data with this property to measure applicability and usefulness of the algorithm.

References

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, Santiago, Chile, 1994.

[HBG+97] E.H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. Webace: A web agent for document categorization and exploartion. Technical Report TR-97-049, Department of Computer Science, University of Minnesota, M inneapolis, 1997.

[HKKM97] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering in a high-dimensional space using hypergraph models. Technical Report TR-97-063, Department of Computer Science, University of Minnesota, Minneapolis, 1997.

[SA96] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proc. of 1996 ACM-SIGMOD Int. Conf. on Management of Data*, Montreal, Quebec, 1996.

[SAD+93] M. Stonebraker, R. Agrawal, U. Dayal, E. J. Neuhold, and A. Reuter. DBMS research at a crossroads: The vienna update. In *Proc. of the 19th VLDB Conference*, pages 688–692, Dublin, Ireland, 1993.