MINING ASSOCIATION RULES FOR LARGE TRANSACTIONS USING NEW SUPPORT AND CONFIDENCE MEASURES

G.JAYALAKSHMI DR.K.NAGESWARA RAO

[1] Lecturer, ITDepartment, V.R.SiddharthaEngineering College,Kanuru,Vijayawada,A.P,India

[2]Prof & HOD, CSE Department, PVP Siddhartha Institute of Technology,Vijayawada,AP,India

E-mail: jaya1123@gmail.com , drknrao@ieee.org

ABSTRACT

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. End users of association rule mining tools encounter several in practice when data bases come with binary attributes. In this paper, we introduce a new measure, which does not require initially allotted weights .The quality of transactions is considered by link based models and a fast mining algorithm is adopted.

Index Terms—*Data Mining, Ranking Association Rules, HITS, Link Analysis.*

1. INTRODUCTION

The classical model of association rule mining employs the support measure, which treats every transaction equally. In contrast, different transactions have different weights in real-life data sets. For example, in the market basket data, each transaction is recorded with some profit. Much effort has been dedicated to association rule mining with allotted weights However, most data types do not come with such allotted weights, such as Web site click-stream data. There should be some notion of importance in those data. For instance, transactions with a large amount of items should be considered more important than transactions with only one item. Current methods, though, are not able to estimate this type of importance and adjust the mining results by emphasizing the important transactions.

In this paper, we introduce w-support, a new measure of item sets in databases with only binary attributes. The basic idea behind w-support is that a frequent item set may not be as important as it appears, because the weights of transac-tions are different. These weights are completely derived from the internal structure of the database based on the assumption that good transactions consist of good items. This assumption is exploited by extending Kleinberg's HITS model and algorithm to bipartite graphs. Therefore, w-support is distinct from weighted support in weighted association rule mining (WARM) where item weights are assigned. Furthermore, a new measurement framework of association rules based on w-support is proposed. Experimental results show that w-support can be worked out without much overhead, and interesting patterns may be discovered through this new measurement.

The rest of this paper is organized as follows: First, WARM is discussed. Next, we present the evaluation of transactions with HITS, followed by the definition of w-support and the corresponding mining algorithm. An interesting real-life example and experimental results on different types of data are given. Concluding remarks are made in the last section

2 WEIGHTED ASSOCIATION RULE MINING

The concept of association rule was first introduced . It proposed the support-confidence measurement framework and reduced association rule mining to the discovery of frequent item sets. The following year a fast mining algorithm, Apriori, was proposed]. Much effort has been dedicated to the classical (binary) association rule mining problem since then. Numerous algorithms have been proposed to extract the rules more efficiently. These algorithms strictly follow the classical measurement framework and produce the same results once the minimum support and minimum

confidence are given.

WARM generalizes the traditional model to the case where items have weights. Ramkumar et al. introduced weighted support of association rules based on the costs assigned to both items as well as transactions. An algorithm called WIS was proposed to derive the rules that have a weighted support larger than a given threshold. Cai et al defined weighted support in a similar way except that they only took item weights into account. The definition broke the downward closure property. As a result, the proposed mining algorithm became more complicated and time consuming. Tao et al. [9] provided another definition to retain the "weighted downward closure property." In

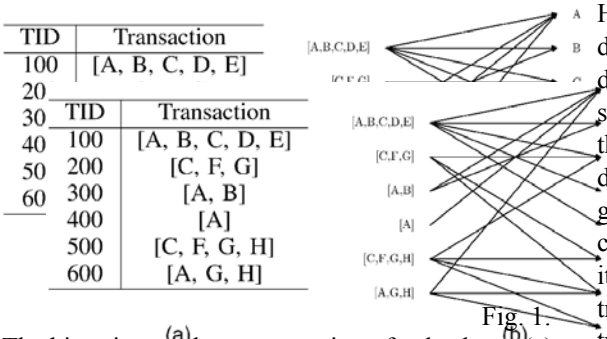


Fig. 1. The bipartite graph representation of a database. (a) Database.
(b) Bipartite graph.

conclusion, the methodology of WARM is to assign weights to items, invent new measures (weighted support) based on these weights, and develop the corresponding mining algorithms.

Wang and Su proposed a novel approach on item ranking. A directed graph is created where nodes denote items and links represent association rules. A generalized version of HITS is applied to the graph to rank the items, where all nodes and links are allowed to have weights. However, the model has a limitation that it only ranks items but does not provide a measure like weighted support to evaluate an arbitrary item set. Anyway, it may be the first successful attempt to apply link-based models to association rule mining.

3 RANKING TRANSACTIONS WITH HITS

A database of transactions can be depicted as a bipartite graph without loss of information. Let D={T1,T2,…Tm}be a list of transactions and I={i1,i2…in} be the corresponding set of items. Then, clearly D is equivalent to the bipartite graph G = (D,I,E), where

E={(T,i):i €T,T € D, i € I}

Example 1. Consider the database shown in Fig. 1a. It can be equivalently represented as a bipartite graph, as shown in Fig. 1b.

The graph representation of the transaction database is inspiring. It gives us the idea of applying link-based ranking models to the evaluation of transactions. In this bipartite graph, the support of an item i is proportional to its degree, which shows again that the classical support does not consider the difference between transactions. However, it is crucial to have different weights for different transac-tions in order to reflect their different importance. The evaluation of item sets should be derived from these weights. Here comes the question of how to acquire weights in a database with only binary attributes. Intui-tively, a good transaction, which is highly weighted, should contain many good items; at the same time, a good item should be contained by many good transactions. The reinforcing relationship of transactions and items is just like the relationship between hubs and authorities in the HITS model .

Regarding the transactions as "pure" hubs and the items as "pure" authorities, we can apply HITS to this bipartite graph. The following equations are used to perform each iteration:

auth(i)=∑T:i € T hub(T),hub(T)=∑i:i €T auth(i).

When the HITS model eventually converges, the hub weights of all transactions are obtained. These weights represent the potential of transactions to contain high-value items. A transaction with few items may still be a good hub if all component items are top ranked. Conversely, a transaction with many ordinary items may have a low hub weight.

4 W-SUPPORT: A NEW MEASUREMENT

Item set evaluation by support in classical association rule mining is based on counting. In this section, we will introduce a link-based measure called w-support and formulate association rule mining in terms of this new concept.

The previous section has demonstrated the application of the HITS algorithm to the ranking of the transactions. As

the iteration converges, the authority weight authðiÞ ¼
P
T :i2T hubðT Þ represents the "significance" of an item i. Accordingly, we generalize the formula of authðiÞ to depict the significance of an arbitrary item set, as the following definition shows:

$$wsupp(X) = \frac{\sum_{T:X \subset T \wedge T \in D} hub(T)}{\sum_{T:T \in D} hub(T)},$$

Observe that replacing all hubðT Þ with 1 on the right-hand side of (2) gives suppðX Þ. Therefore, w-support can be regarded as a generalization of support, which takes the weights of transactions into account. These weights are not determined by assigning values to items but the global link structure of the database. This is why we call w-support link based. Moreover, we claim that w-support is more reason-able than counting-based measurement. This could be verified through the following example:

Example 2. Consider the database shown in Fig. 1 again. The HITS iteration gives the hub weight of each transaction and w-support of each 1-item set, as shown in Table 1. It is interesting to point out that the best hub (transaction 500 [C F G H]) is not the one with the largest item number, and the most significant 1-item set ({C}) is not the one with the largest support. This shows the intrinsic difference between link-based and counting-based measurement. Transactions 200 and 500 and items C, F, and G form a complete bipartite graph, which implies that a strong cross-selling effect exists between the three items. These items should be highly evaluated because they not only occur frequently by themselves but also reinforce the value of each other by occurring together. On the other hand, although item A has the highest support, it seldom shows up with other valuable items. Thus, A should be ranked somewhat lower. In essence, w-support introduces the cross-selling effect into the evaluation of item sets.

TID	Transaction	Hub weight
100	[A B C D E]	0.518
200	[C F G]	0.436
300	[A B]	0.233
400	[A]	0.148
500	[C F G H]	0.544
600	[A G H]	0.412

1-itemset	Support	W-support
{A}	0.67	0.57
{B}	0.33	0.33
{C}	0.50	0.65
{D}	0.17	0.23
{E}	0.17	0.23
{F}	0.33	0.43
{G}	0.50	0.61
{H}	0.33	0.42

Furthermore, w-support evaluates item sets in a more distinguishable way. For example, items B, F, and H all have a support of 0.33. However, their w-supports are different. F is ranked first among the three because it is likely to appear together with good items (C and G).

For association rules, we give the following definition.

Definition 2. The w-support of an association rule X) Y is defined as

$$wsupp(X \Rightarrow Y) = wsupp(X \cup Y),$$

and the w-confidence is

$$wconf(X \Rightarrow Y) = \frac{wsupp(X \cup Y)}{wsupp(X)}. \quad -$$

The w-confidence can be understood as the ratio of the hub weights received by X together with Y to the total hub weights received by X. Basically, w-support measures how significantly X and Y appear together; w-confidence mea-sures how strong the rule is. If wconfðX) Y Þ is large, it shows that many good hubs that vote X also vote Y , although the fraction of these hubs may be small. Accordingly, association rule mining is to discover all rules with w-support and w-confidence above some given thresholds.

5 A FAST MINING ALGORITHM

The problem of mining association rules that satisfy some minimum w-support and w-confidence can be decomposed into two subproblems:

1. Find all significant item sets with w-support above the given threshold.
2. Derive rules from the item sets found in Step 1.

The first step is more important and expensive. The key to achieving this step is that if an item set satisfies some minimum w-support, then all its subsets satisfy the mini-mum w-support as well. It is called the downward closure property of w-support.

1) Initialize $auth(i)$ to 1 for each item i
2) **for** $(l = 0; l < num\_it; l{+}{+})$ **do begin**
3) $auth'(i) = 0$ for each item i
4) **for all** transactions $t \in D$ **do begin**
5) $hub(t) = \sum_{i:i \in t} auth(i)$
6) $auth'(i) += hub(t)$ for each item $i \in t$
7) **end**
8) $auth(i) = auth'(i)$ for each item i, normalize $auth$
9) **end**
10) L_1 = {{i} : $wsupp(i) \geq minwsupp$}
11) **for** $(k = 2; L_{k-1} \neq \emptyset; k{+}{+})$ **do begin**
12) C_k = apriori-gen(L_{k-1})
13) **for all** transactions $t \in D$ **do begin**
14) C_t = subset(C_k, t)
15) **for all** candidates $c \in C_t$ **do**
16) c.wsupp += hub(t)
17) H += hub(t)
18) **end**
19) $L_k = \{c \in C_k | c.wsupp/H \geq minwsupp\}$
20) **end**
21) $Answer = \bigcup_k L_k$

Fig. 2. An algorithm for mining significant item sets.

Besides, the hub weights of all transactions are non-negative. Hence,

$$\sum_{T:X \subset T \wedge T \in D} hub(T) \leq \sum_{T:Y \subset T \wedge T \in D} hub(T).$$

Based on this property, we can extract significant item sets in a levelwise manner, as the Apriori-like algorithm demonstrated in Fig. 2.

6 A REAL-LIFE EXAMPLE

In this section, we will demonstrate the usefulness of w-support through an intuitive example. The data set of Netflix [19] is used, which contains over 100 million 1 ₃ 5 star ratings from 480,189 anonymous customers on 17,770 movies.

The first task here is to select the top 10 popular movies. With the assumption that the popularity of a movie is positively correlated to its number of ratings but indepen-dent of the actual rating values, the database can be considered as a transaction database, where each movie corresponds to one item, and all movies rated by a particular user correspond to one transaction.

Results obtained by both support and w-support are shown in Table 2. One can see that the support simply counts the number of ratings of each movie. In contrast, the w-support differentiates the roles of different users. Users who are new to the online rating system or used the system in the past but finally abandoned it are characterized as the inactive users. On the contrary, some old customers have been using the system for a long time. They are character-ized as the active users. When evaluating the popularity, both system users and system maintainers would be more interested in the voice of those active users, because they are considered to be authoritative, responsible, and profitable. Thus, the activity of users and the popularity of movies enjoy a reinforcing relationship, which can be reflected in the HITS model and the w-support measure. For example, The Sixth Sense is absent in the top 10 frequently rated movies. However, its w-support is large. Thus, we can realize that the users who have rated the movie are mostly active, and the ranking of the movie is therefore deserved.

Another interesting problem is to select the top 10 best movies. We proposed a naive algorithm where only five-star ratings are used to construct a transaction database.

The results are given in Table 3, where the column "IMDB" shows the ratings by the Internet Movie Database.[2] W -support ranks the Lord of the Rings series as the top three best movies. This is not an accident. Lord of the Rings: The Two Towers gains most five-star ratings. Consequently, all users rating it as five stars are regarded to be more authoritative. Hence, the rankings of the other two movies in the same series are boosted by those authoritative users. In conclusion, the values of the movies are correlated with each other via w-support.

7 EXPERIMENTS

To evaluate the link-based association rule mining frame-work, we have modified the Apriori implementation by Bodon [10] so that it uses w-support and w-confidence as the rule selection thresholds. Several tests have been carried out on some classical data sets [20], [22], [10], [21]. The experiments were conducted on a 1.8-GHz Sempron 3000+ machine with 1 Gbyte of RAM running FreeBSD 6.1. All code was compiled using g++ Version 3.4.4.

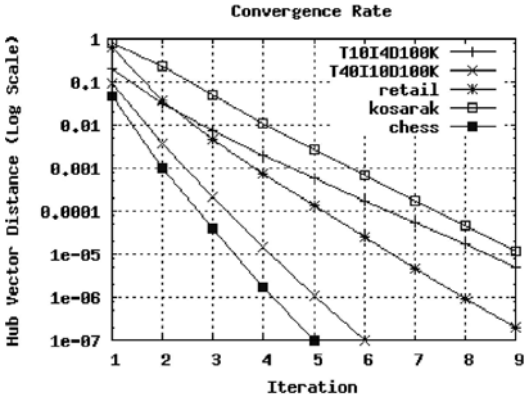
7.1 Performance Study

Compared with Apriori , the proposed mining algorithm (Fig. 2) requires an additional iterative procedure to compute the hub weights of all transactions. The database is scanned exactly once in each iteration. Therefore, the convergence rate of the hub weights is critical to the performance.

Let H_i denote the vector of hub weights after the ith iteration.. It is clear that HITS converges fast on transaction databases. Generally, three or four iterations are enough to achieve a good estimation, which means that our link-based method works at the cost of three or four additional database scans over the traditional techniques.

7.2 Comparison of Support and W-Support

Three representative data sets, the synthetic T10I4T100K, the sparse retail, and the dense chess, are selected. Fig. 4 gives the w-supports and supports of the most significant item sets with more than one item. For each



item set, the left bar gives its w-support, and the number on top shows its rank value by w-support. Similarly, the right bar represents its support and the corresponding rank value.

It is clear in Fig. 4 that the value of w-support is generally larger than that of support, especially for sparse data. This is due to the mutually reinforcing relationship of hubs and w-supports. Through the HITS iteration, an item set with a large w-support will enlarge the hub weights of all transactions containing it, which in turn will make its w-support even larger. However, in the case of the dense data set, such as chess where about 600 item sets have a support of more than 90 percent, almost all transactions include some significant items. Therefore, it is hard for the hub weights of the transactions to be diverse. As a result, little difference exists between w-supports and supports on dense data sets, as shown in Fig. 4. Hence, the w-support measurement is not recommended for dense data sets.

For the sparse data sets, it is worth noting that the order of item sets given by w-support is very much different from the order by support, as shown in Fig. 4. Using w-support, we are able to discover some significant item sets that are not frequent. Among the most significant item sets of T10I4D100K, [937 368], [682 494], and [529 368] are beyond the top 100 frequent item sets; in the retail data, [38 41 48 39] and [32 41 48 39] are also examples of being significant but not so frequent. Consider the two item sets [38 41 48 39] and [170 39] in retail. The former is less frequent than the latter. However, in the experiments, we found that the average length of item sets containing [38 41 48 39] is 17.6, and it is only 13.1 for [170 39]. Although the hub weight of an item set is not determined by its length, long item sets are more likely to obtain high hub weights. Hence, [38 41 48 39] is ranked somewhat higher by its high-quality hubs. The significance of other infrequent item sets can be explained in a similar way.

7.3 Link-Based Association Rule Mining

Since w-support and w-confidence are normally larger than support and confidence, respectively, a comparison of the two measurement techniques with the same thresholds does not make sense. Instead, we select the thresholds so that the two models produce about the same amount of item sets and association rules.

Consider the data set retail as an example. With minwsupp ¼ 2:4 percent and minwconf ¼ 88 percent in the link-based model, 81 item sets and 19 rules are generated; with minsupp ¼ 1:5 percent and minconf ¼ 75 percent in the traditional model, 84 item sets and 19 rules are discovered. The resulting association rules are shown in Table 4.

Observe that the two models agree well on most of the rules, though they both advocate some rules that are not discovered by the other. Basically, two types of association
rules are likely missing in the traditional model but not in the link-based model:

1. Not so frequent but supported by many good hubs (transactions).
2. With small confidence but many good hubs support-

ing X also support Y (assume that the rules is X) Y).

For example, in Table 4, the first type includes ½170 48 39&) ½38& and ½36 48 39&) ½38&, whereas ½225&) ½39& and ½310&) ½39& are examples of the second type. On the other hand, we do miss some rules that are discovered in the traditional models. The details are omitted here for brevity. In essence, the difference is caused by our basic assumption: the quality of transactions and value of items are in a mutually reinforcing relationship.

7.4 Evaluation

The "cross-selling" effects between items are taken into consideration by the HITS model. Consequently, some item sets, which are not so frequent but accompany good items, may easily be missed by the traditional counting-based model but discovered by ours. We do have found such kind of item sets in the experiments.

Compared with the traditional counting-based measure-ment, the proposed model emphasizes large transactions, because they are generally valuable. In the retail business, maintainers are more interested in customers who buy lots of stuff; in Web click-stream data, longer sessions may correspond to regular visitors; in recommendation systems, a user who has rated many movies (or anything else) is likely to have better taste. The assumption that large transactions are more important may not be a ground truth, but when it is (very likely), the model works.

The experimental results are to show that

1. the resulting rules/item sets put emphasis on good transactions (hubs) and
2. some infrequent rules/item sets may be disclosed with "cross-selling" effects.

It is hard to tell whether a rule/item set is valuable by any objective measure, because traditional association rule mining does lack effective measures. A similar case happens in Web search. It is hard to tell which Web page should be ranked high. All we can argue academically is the model itself. The success of Google proves the efficiency of the link-based models.

8 CONCLUSIONS AND DISCUSSIONS

We have presented a novel framework in association rule mining. First, the HITS model and algorithm are used to derive the weights of transactions from a database with only binary attributes. Based on these weights, a new measure w-support is defined to give the significance of item sets. It differs from the traditional support in taking the quality of transactions into consideration. Then, the w-support and w-confidence of association rules are defined in analogy to the definition of support and confidence. An Apriori-like algorithm is proposed to extract association rules whose w-support and w-confidence are above some given thresholds.

Experimental results show that the computational cost of the link-based model is reasonable. At the expense of three or four additional database scans, we can acquire results different from those obtained by traditional counting-based models. Particularly for sparse data sets, some significant item sets that are not so frequent can be found in the link-based model. Through comparison, we found that our model and method address emphasis on high-quality transactions.

REFERENCE:

[1]. R. Agrawal, T. Imielinski, and A. Swami, "Mining AssociationRules between Sets of Items in Large Datasets," Proc. ACM SIGMOD '93, pp. 207-216, 1993

[2]. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 487-499, 1994

[3]. J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," J. ACM, vol. 46, no. 5, pp. 604-632, 1999.

[4]. O. Kurland and L. Lee, "Respect My Authority! HITS without Hyperlinks, Utilizing Cluster-Based Language Models," Proc. ACM SIGIR, 2006.

[5]. K. Wang and M.-Y. Su, "Item Selection by "Hub-Authority" Profit Ranking," Proc. ACM SIGKDD, 2002.

[6]. W. Wang, J. Yang, and P.S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," Proc. ACM SIGKDD '00, pp. 270-274, 2000

[7]. B. Liu, W. Hsu, and Y. Ma, "Integrating Classification and Association Rule Mining," Proc. ACM SIGKDD '98, pp. 80-86, 1998.

[8]. S. Madria, S. Bhowmick, W. Ng, and E. Lim, "Research Issues in Web Data Mining," Proc. First Int'l Conf. DataWarehousing and Knowledge Discovery (DaWaK '99), pp. 303-312, 1999.

[9]. J. Li, B. Tang, and N. Cercone, "Applying Association Rules for Interesting Recommendations Using Rule Templates," Proc. Eighth Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '04), pp. 166-170, 2004.

[10]. Netflix Prize, http://www.netflixprize.com/, 2007.

[11]. The IBM Synthetic Data Generator, http://www.almaden.ibm.com/software/projects/iis/hdb/Projects/data_mining/data sets/syndata.html, 2007.

[12]. R.J. Bayardo Jr. and R. Agrawal, "Mining the Most Interesting Rules," Proc. ACM SIGKDD '99, pp. 145-154, 1999.

[13]. T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, "The Use of Association Rules for Product Assortment Decisions: A Case Study," Proc. ACM SIGKDD '99, pp. 254-260, 1999.

[14]. J.S. Park, M.S. Chen, and P.S. Yu. An effective hash-based algorithm for mining association rules. SIGMOD, 175-186, 1995.

[15]. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. VLDB, 432-444, 1995

BIOGRAPHY:

1.G.JayaLakshmi received the degree in Computer Science and Engineering from Nagarajuna University in 2000. He is a student MTech of PVPSIT. Currently, she is an Lecturer at VR Siddhartha Engineering College. Her interests are in Data Mining and Network Security.

2.Dr. K.Nageswara Rao did his B.Tech (ECE) from Karnatak University , subsequently M.Tech(CST) and Ph.D (CS&SE) from Andhra University. After serving Andhra University for nearly thirty years, he is now Professor and Head of Computer Science and Engineering, Prasad V. Potluri Siddhartha Institute of Technology.he is a member of PG Board of Studies for M.Sc.(VLSI) & M.Sc. (E. S) A . U, Visakhapatnam andUG Board of Studies B.Tech.(IT), VRSEC (Autonomous), Vijayawada. He has affiliation to various professional societies like IEEE,ACM, IE,IETE,ISTE and CSI. He has been on the NBA committee as a CSE/IT subject expert. He has to his credit 15 publications. His current areas of research include Data Mining and Ware Housing, Software Engineering , Artificial Intelligence .