ISSN:0975-9646

Weighted Association Rule Mining Without Pre-assigned Weights

PURNA PRASAD MUTYALA, KUMAR VASANTHA
Department of CSE, Avanthi Institute of Engg & Tech, Tamaram, Visakhapatnam, A.P., India.

***Abstract*— Association rule mining is a key issue in data mining. However, the classical models ignore the difference between the transactions, and the weighted association rule mining does not work on databases with only binary attributes. In this paper, we introduce a new measure w-support, which does not require pre-assigned weights. It takes the quality of transactions into consideration using link-based models. A fast mining algorithm is given, and a large amount of experimental results are presented. The weights are completely derived from the internal structure of the database based on the Assumption that good transactions consist of good items. Consequently, some item sets, which are not so frequent but accompany good items, may easily be missed by traditional counting-based model but discovered by ours. 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.**

***Keywords*— wsupport, ranking association rules, HITS, link analysis**

1. INTRODUCTION

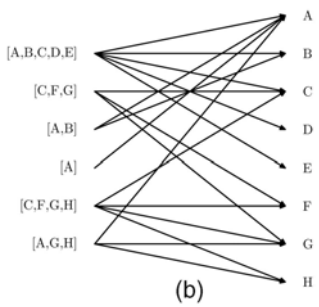
Association rule mining aims to explore large transaction databases for association rules, which may reveal the implicit relationships among the data attributes. It has turned into a thriving research topic in data mining and has numerous practical applications, including cross marketing, classification, text mining, Web log analysis, and recommendation systems [1] 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 preassigned weights [6]. However, most data types do not come with such preassigned 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 transactions 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 [3] to bipartite graphs. Therefore, wsupport is distinct from weighted support in weighted association rule mining (WARM) [6], 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.

2. WEIGHTED ASSOCIATION RULE MINING

The concept of association rule was first introduced in [1]. 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 [2]. 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. [6] 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."

TID	Transaction
100	[A, B, C, D, E]
200	[C, F, G]
300	[A, B]
400	[A]
500	[C, F, G, H]
600	[A, G, H]

(a)



(b)

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

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

$$E = \{(T, i): i \in T, T \in D; i \in I\}$$

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 transactions 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. Intuitively, 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. The following equations are used to perform each iteration:

$$\text{auth}(i) = \sum_{T: i \in T} \text{hub}(T); \qquad \text{hub}(T) = \sum_{i: i \in T} \text{auth}(i);$$

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. A NEW MEASUREMENT: *WSUPPORT*

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 auth (i) = $\sum$ T:i€T hub(T) represents the "significance" of an item i. Accordingly, we generalize the formula of auth(i) to depict the significance of an arbitrary item set, as the following definition shows:

Definition 1. The w-support of an item set X is defined as

$$\text{Wsupp}(X) = \sum_{T: X \subseteq T \in D} \overline{\text{hub}(T)} \Big/ \sum_{T: T \in D} \text{hub}(T)$$

Where hub (T) is the hub weight of transaction T. An item set is said to be significant if its w-support is larger than a user-specified value. Observe that replacing all hub (T) with 1 on the right-hand side of gives supp(X). 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

Table 1: Hubs and W-Supports of the Example Database

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 => Y) = wsupp (XU Y); and the w-confidence is wsupp(X => Y) = wsupp (XU Y) / wsupp(X)

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 measures how strong the rule is. If wconf (X=>Y) 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 sub problems:

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.

Prof. Let X be an item set that satisfies wsupp(X) ≥ minwsupp and Y be a subset of X, we shall prove wsupp(Y) ≥ minwsupp. First, any transaction that contains X must also contain Y, that is,

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.

{T: X C T ; T € D} C {T: Y C T; T € D}:

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

$$\sum_{T: XCT^T€D} Hub(T) \leq \sum_{T: YCT^T€D} hub(T):$$

Divide both sides by T: XCT^T€D hub (T). Then, we have wsupp(X) ≤ wsupp(Y). This gives the desired result

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

6. EXPERIMENTS

To evaluate the link-based association rule mining frame-work, we have modified the Apriori implementation 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.

6.1 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. Fig. 3 shows $\|H_{i+1} - H_i\|$ as a function of i on the data sets in log scale. 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.

NO.	Support	IMDB	Movie Title
1	96535	8.7	Lord of the Rings: The Two Towers
2	95532	8.4	Forrest Gump
3	95150	9.2	The Shawshank Redemption: Special Edition
4	94655	8.7	Lord of the Rings: The Fellowship of the Ring
5	92863	8.2	The Green Mile
6	92805	8.8	Lord of the Rings: The Return of the King
7	82549	8.0	Pirates of the Caribbean: The Curse of the Black Pearl
8	79447	8.2	Finding Nemo (Widescreen)
9	74790	8.2	The Sixth Sense
10	74553	8.2	Indiana Jones and the Last Crusade

(a)

NO.	Support	IMDB	Movie Title
1	165.1	8.7	Lord of the Rings: The Two Towers
2	165.0	8.7	Lord of the Rings: The Fellowship of the Ring
3	160.5	8.8	Lord of the Rings: The Return of the King
4	160.1	9.2	The Shawshank Redemption: Special Edition
5	154.0	8.4	Forrest Gump
6	144.8	8.7	Raiders of the Lost Ark
7	143.2	8.0	Pirates of the Caribbean: The Curse of the Black Pearl
8	142.7	8.2	Finding Nemo (Widescreen)
9	137.8	8.2	The Sixth Sense
10	135.7	8.6	The Matrix

(b)

Table 2: (a) Results given by support. (b) Results given by w-support

6.2 Comparison of Support and W-Support

Three representative data sets, the synthetic T10 I4T100K, the sparse retail, and the dense chess, are selected. Fig. 3

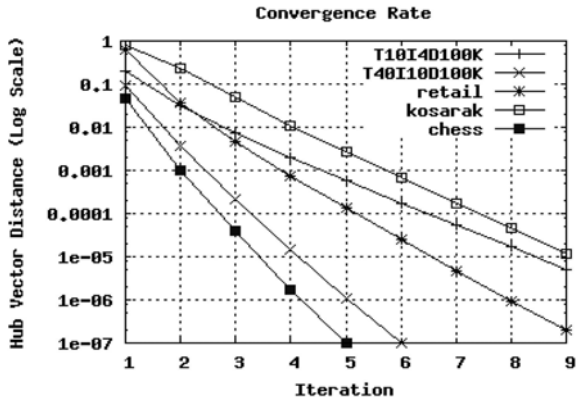


Fig. 3. HITS convergence rate on transaction database

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 data sets. Discovered by the other. Basically, two types of association

6.3 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 supporting X also support Y (assume that the rule is X = > Y).

Rule	W-support	W-confidence	Rule	Support	Confidence
[41] ⇒ [39]	22.5%	88.1%	[41] ⇒ [39]	12.9%	76.4%
[41 48] ⇒ [39]	16.8%	88.6%	[41 48] ⇒ [39]	8.4%	81.7%
[38 41] ⇒ [39]	6.7%	88.1%	*[38 48] ⇒ [39]	6.9%	76.8%
[38 41 48] ⇒ [39]	5.0%	89.4%	[41 38] ⇒ [39]	3.5%	78.3%
[170] ⇒ [38]	4.7%	98.6%	[170] ⇒ [38]	3.4%	97.8%
[36] ⇒ [38]	4.3%	96.4%	[36] ⇒ [38]	3.2%	95.0%
*[225] ⇒ [39]	4.2%	89.1%	[110] ⇒ [38]	3.1%	97.5%
[110] ⇒ [38]	4.1%	98.9%	*[89 39] ⇒ [48]	2.4%	77.3%
[170 39] ⇒ [38]	3.9%	98.6%	*[89 48] ⇒ [39]	2.4%	75.9%
[36 39] ⇒ [38]	3.7%	96.5%	[170 39] ⇒ [38]	2.3%	98.1%
*[310] ⇒ [39]	3.5%	88.1%	[41 38 48] ⇒ [39]	2.3%	83.9%
[110 39] ⇒ [38]	3.4%	99.2%	[36 39] ⇒ [38]	2.2%	95.5%
[170 48] ⇒ [38]	3.2%	99.0%	[110 39] ⇒ [38]	2.0%	98.9%
[225 48] ⇒ [39]	3.0%	88.9%	*[41 32 48] ⇒ [39]	1.9%	79.8%
[310 48] ⇒ [39]	2.9%	88.3%	[170 48] ⇒ [38]	1.7%	98.8%
[36 48] ⇒ [38]	2.9%	96.9%	[225 48] ⇒ [39]	1.6%	80.6%
[110 48] ⇒ [38]	2.8%	99.1%	[110 48] ⇒ [38]	1.5%	98.6%
*[170 48 39] ⇒ [38]	2.8%	99.0%	[36 48] ⇒ [38]	1.5%	96.0%
*[36 48 39] ⇒ [38]	2.5%	97.2%	[310 48] ⇒ [39]	1.5%	79.6%
$minwsupp = 2.4\%$, $minwconf = 88\%$			$minsupp = 1.5\%$, $minconf = 75\%$		

Table3:Association Rules Extracted from the Data Set R e t a i l

For example, in Table 4, the first type includes rules [170 48 39] => [38] and [36 48 39]=>[38], whereas rules [225]=> [39] and [310]=>[39] 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.

6. CONCLUSION

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 Apriority-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. The link-based model is useful in adjusting the mining results given by the traditional techniques. Some interesting patterns may be discovered when the hub weights of transactions are taken into account. Moreover, the transaction ranking approach is precious for estimating customer potential when only binary attributes are available, such as in Web log analysis or recommendation systems.

REFERENCES

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

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 20th Int'l Conf. Very Large Databases (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] C.H. Cai, A.W.C. Fu, C.H. Cheng, and W.W. Kwong, "Mining Association Rules with Weighted Items," Proc. IEEE Int'l Database Eng. and Applications Symp. (IDEAS '98), pp. 68-77, 1998.

[5] F. Tao, F. Murtagh, and M. Farid, "Weighted Association Rule Mining Using Weighted Support and Significance Framework," Proc. ACM SIGKDD '03, pp. 661-666, 2003

[6] G.D. Ramkumar, S. Ranka, and S. Tsur, "Weighted Association Rules: Model and Algorithm," Proc. ACM SIGKDD, 1998.

Mr. Purna Prasad Mutyala received the B.Tech degree from the Department of Information Technology, GEC, JNTUniversity, Kakinada in 2009 and He is currently pursuing M.Tech in the Department Of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Vishakhapatnam, JNTUniversity. His research interests include Data Mining and Association Rules.

Mr. Kumar Vasantha received the M.Tech degree from the Department of Computer Science and Engineering, Avanthi Institute of Engineering and Technology, Vishakhapatnam, JNTUniversity, Kakinada in 2009 and working as a Asst. Prof in Avanthi Institute of Engineering and Technology, Vishakhapatnam His research interests include Information Security and Data Mining.