Volume 3, Issue 10, October 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Efficient Interesting Weighted Association Rule Mining

M. Padmavalli, Research Scholar
Department of Computer Science &Technology
S.K Uiversity, Anantapur-AP, India

Prof.K.Sreenivasa Rao
Department of OR&SQC,
Rayalaseema University, Kurnool, AP, India

Abstract: Mining association rules is an important problem in the field of data mining due to its wide applications. Weighted Association rule mining has recently been proposed, in which transactions are attached with weighted values according to some criteria. The weights in these approaches may be thought of as an extension of traditional support in association-rule mining. Weighted association rules can be discovered in a variety of forms, like weighted association rules, fuzzy weighted association rules, and weighted utility association rules. We have used HITS algorithm to automatically calculate transaction weights. We presented a new algorithm for finding interesting weighted association rules. The interestingness of an item can be computed at the multiplication of weight and support. Interestingness, in some case, can be "the potentially useful for finding association rules. Our experimental result shows that proposed interesting weighted Association Rule Mining out performs the existing algorithm in terms of efficiency, time and valued rules.

Keywords:

1. Introduction

Mining association rules [1] is an important issue in the field of data mining due to its wide applications. Traditional association rules are, however, derived from frequent itemsets, which only consider the occurrence of items but do not reflect any other factors, such as price or profit. Weighted Association rule mining has recently been proposed, in which transactions are attached with weighted values according to some criteria [2][7][8]. It is important because if the same significance is assumed for all the transactions in a database, some interesting association rules may not be found by traditional mining. However, the actual significance of an itemset cannot be easily recognized. The problem of weighted association rule mining is to find the complete set of association rules satisfying a support constraint and a weight constraint in the database. When we compute the weighted support of the rule, we can consider both the support and the weights factors. In the real world, there are more importance have several applications where specific patterns and items within the patterns have more importance or priority than the other patterns. Weighted Association rule mining [9] has been suggested to find important frequent rules by considering the weights of patterns. The concept of Weighted Association rule mining is attractive in that important patterns are discovered. We can use the term, weighted itemset to represent a set of weighted items. A simple way to obtain a weighted itemset is to calculate the average value of the weights of the items in the itemset.

2. Weighted Association-Rule Mining

Weighted association-rule mining [2][8] is concerned with the analysis of significance of items or transactions in a set of data. It is proposed to find out different kinds of interesting patterns from a set of data with item weight or transaction weight. The weights in these approaches may be thought of as an extension of traditional support in association-rule mining. Weighted association rules can be discovered in a variety of forms, like weighted association rules, fuzzy weighted association rules, and weighted utility association rules.

An item weight, w, where $0 \leq w \leq 1$, defines the importance of the item. 0 indicates the least important item, and 1 denotes the most important item. For example, if the weight of the itemset X is 0.95, it tells us the itemset is important in the set of transaction D . The weight of 0.1 indicates a less important set.

A weighted association rule (or association rule with weighted item) has the form $X \Rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \phi$, and the items in X and Y are given by the weights.

Weighted Support: The weighted support of the binary weighted rule $X \Rightarrow Y$ is the adjusting ratio of the support, or mathematically,

$wsupport\ (X\ Y) = (\sum_{j=x \cup y}^{n} (\mathrm{wj}))\ support\ (X, Y)$

where the weights of the items $\{i1,i2,...,im\}$ are $\{w1,w2,...,wm\}$ respectively. In order to find the interesting rules, two thresholds, minimum weighted support (*wminsup*) and minimum confidence (*minconf*) must be specified.

An itemset X is called a large weighted itemset if the weighted support of the itemset X is greater than or equal to the weighted support threshold, or mathematically, $wsupport(X) \geq wminsup$

Weighted Confidence: A weighted association rules $X \Rightarrow Y$ is called an interesting rule if the confidence of itemset ($X \cup Y$) is greater than or equal to a minimum confidence threshold, and ($X \cup Y$) is a large weighted itemset.

2.1. Interesting itemsets

By setting the interestingness of an itemset, we can get a balance between the two measures, weights and supports. If supports are separated from weights, we can only find itemsets having sufficient support. However, this may ignore some interesting knowledge. Special items and special group of items may be specified individually and have higher priority. For example, there are few customers buying bread, but the profit the bread makes is much more than that of other products. As a matter of course, the store clerk will want to put the bread under the promotion rather than others. For this reason, the weight which is a measure of the important of an item is applied.

The interestingness of an item can be computed at the multiplication of weight and support. Interestingness, in some case, can be "the potential usefulness of the knowledge" but it seems to be difficult to understand. It is clear that most end-users are not statisticians, they thus have trouble setting the threshold for min_int. Putting a query "Show me twenty most interesting itemsets" is definitely more comprehensible than "Please list itemsets whose interestingness are greater or equal to 0.5". Furthermore, it is impractical to generate entire set of interesting itemsets. Our purpose is to mine only most interesting ones. Hence, we design a new concept, interestingness. Based on the definitions of weighted itemsets, we extend the definitions of interestingness and interesting itemsets.

Interest: The interestingness of an itemset X , denoted *interest*(X), is the coefficient correlation between the number of transactions in which it occurs as a subset and the total weight of its items, or methametically,

$\text{Interest(X)} = (\sum_{j=x}^{n} (\text{wj}))$ *support (X)*

In order to find the interesting itemsets, the threshold, minimum interestingness (*min_int*) must be specified.

An itemset X is called an interesting itemset if the interestingness of the itemset X is greater than or equal to the interestingness threshold, or mathematically,

$interest(X) \geq min\_int$

Meanwhile, we address the problem that the downwards closure property is invalid in the weighted association rule mining model by setting a factor which relates minimum support with weighted minimum support so as to maintain a property that if an itemset is a weighted frequent itemset under the weighted minimum support then the itemset must be a frequent itemset under the weighted minimum support. Therefore, we can firstly find frequent K-itemset L_K, then generate weighted K-frequent itemset WL_K from L_K.

2.2 Calculating Weights of Transaction using HITS

Kleinberg's HITS algorithm[3], "Hypertext Induced Topic Selection", is a standard algorithm of Link Analysis that rates web pages, developed by Jon Kleinberg. The premise of the HITS algorithm is that a web page serves two purposes: to provide information and to provide links relevant to a topic. This gives two ways to categorize a web page. A web page is an authority on a topic if it provides good information, and it is a hub if it provides links to good authorities. The HITS algorithm is an iterative algorithm developed to quantify each page's value as a hub and an authority.

HITS is a system used for evaluating the usefulness of a web pages on the Internet. Its establishes a website as an Authority or a Hub. Authorities are those sites that are linked to by Hubs. It is a link analysis algorithm that rates Web pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. Item set evaluation by support in classical association rule mining is based on counting. We introduced a link-based measure called w-support and formulate association rule mining in terms of this new concept. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

3. Related Work

C.H.Cai et al[2] have proposed a mining of association rules with Weighted Items.. Downward closure property of the support measure in the unweighted case is not valid in this framework and previous algorithms cannot be applied. The authors proposed, two new algorithms MINWAL (O) and MINWAL (W) to handle this problem. The proposed algorithm for mining weighted association rules is similar to the Apriori Gen Algorithm, but the detailed steps contain some differences. In the beginning large itemsets is generated with increasing sizes. However, since the subset of a large itemset may not be large, k-itemsets is not generated simply from the large (k -1) itemsets as in Apriori Gen. In order to extract such k-itemsets from the database, a new metric called the k-support bound has been used in the mining process.

Lu *et* al[8] have proposed a new mixed weighted mining model for finding weighted association rules from a set of data. The weighted rules might provide more useful and interesting information to decision makers in many applications such as retail marketing, data stream, medical data analysis, and among others. Feng Tao[10] et al address the issues of discovering significant binary relationships in transaction datasets in a weighted setting. Traditional model of association rule mining is adapted to handle weighted association rule mining problems where each item is allowed to have a weight. They identify the challenge of using weights in the iterative process of generating large itemsets. The

problem of invalidation of the "downward closure property" in the weighted setting is solved by using an improved model of weighted support measurements and exploiting a "weighted downward closure property". A new algorithm called WARM (Weighted Association Rule Mining) is developed based on the improved model. The algorithm is both scalable and efficient in discovering significant relationships in weighted settings as illustrated by experiments performed on simulated datasets.

Wei Xie et al [11] have described the challenging problems in the weighted association rules mining is to assign weights to items. For practice, self-assigned weights technique is more useful. In this paper, we proposed a self-assigned weights method to discover positive and negative association rules, instead of assigning the weights by users. To avoid mining misleading and uninteresting rules, a new type parameter, called *sawinterest*, is proposed to eliminate the redundant rules. The rational results are presented.

4. Problem Definition

Let I = {i_1, i_2, … ,im} be a set of literals called items , W={ w_1, w_2, … , w_m } be a set of non-negative real numbers called weights, where wj is the weight of item ij for j=1,2,…, m. Let the database D = { t1, t2, … ,tn } be a set of n transactions, where each transaction is a subset of I. A nonempty subset of I is called itemset. An itemset containing k items is called k-itemset. The support of an itemset X denoted as sup (X) is defines as the fraction of all transactions containing X in D. An itemset is frequent if its support is greater than a user-specified threshold minimum support minisup. The minimum weighted support is denoted as wminisup.

The classical Apriori algorithm for finding binary association rules depends on the downward closure property which governs that subsets of a frequent itemset are also frequent. However, it is not true for the weighted case.

4.1 PROPOSED ALGORITHM

Interesting Weighted ARM:

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 minimum w-support as well. It is called the downward closure property of w-support.

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, {T : X ⊂ T;T ∈ D}⊂{T : Y⊂T, T ∈ D} Besides, the hub weights of all transactions are nonnegative. Hence,

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

Divide both sides by $\sum_{T:T\in D} hub(T)$. Then, we have wsuppX ≤ wsuppY . This gives the desired result. Based on this property, we can extract significant item sets in a levelwise manner, as the Apriori-like algorithm demonstrated in Fig. 1.

Algorithm : Mining weighted Association Rules

Input : database D, set of items I = {i1, i2, … ,im} , set of weights W={ w1, w 2, … , w m };
weighted minimum support threshold: wminisup;
weighted minimum confidence threshold: wminconf ;
weighted mininterest wmint ;
Output: Set of Weighted interesting association rules
Initialize auth(i) to 1 for each item i
For(l=0;l<num_it;l++)
Auth'(i)=0 for each item i
For all transactions t ∈ D do begin
Hub(t)= $\sum i; i \in t$ auth(i)
Auth'(i) += hub(t) for each item i ∈ t
end
Auth(i)=auth'(i) for each item i , normalize auth
end
for each item I_i of K- frequent itemsets in transaction Database D
for each Ii ∈ T do
if (wsupport(i) ≥ minsupport(i))
if(interest ≥ minint)
For(k=2; L_{K-1} ≠ ϕ ; k++) do begin

C_k = apriori-gen(L_{K-1})
For all transactions t ∈ D do begin
C_t=subset(C_k, t)
For all candidates c ∈ C_t do
w.wsupp += hub(t)
H += hub(t)
end
Lk = { c ∈ C_k | c.wsupp/H ≥ minwsupp }
end
Rules = $\bigcup_k L_K$
Figure 1: weighted Association rule Algorithm

Table 1: Symbols Used in Proposed Model

S.No.	Symbols	Explanation
1	Database D = {i1, i2, … ,im}	A original database D consisting of set of items I = {i1, i2, … ,im}
2	I ={i_1,i_2,…i_M}	An item set of length M
3	W	set of weights { w1, w 2, … , w m }
4	Wminisup	weighted minimum support threshold
5	Wminconf	weighted minimum confidence
6	Wmint	weighted mininterest
7	C_k	apriori-gen(L_{K-1})
8	C_t	subset(C_k, t)
9	MinConfidence	User specified Minimum confidence threshold
10	Lk	Large k-itemset

4.2 Example

Consider the database shown in table2. The HITS iteration gives the hub weight of each transaction and w-support of each 1-item set, as shown in Table3. Table 2 gives the transaction database. For each transaction, there will be a transaction identifier (*TID*) and the names of items. Suppose there are only 4 items and totally 5 transactions in the transaction database.

Table 2: Transaction Database

Tid	Items
1	Bread, Milk
2	Bread, Beer, Diaper, eggs
3	Milk, beer, Diaper, coke
4	Bread, Milk, Beer, Diaper
5	Bread, Milk, Diaper, coke

Table 3: Weights of transaction using HITS

Itemset	Support	W-support
B	0.8	0.78
M	0.8	0.81
Be	0.6	0.65
D	0.8	0.88
E	0.2	0.19
C	0.4	0.44

Frequent weighted itemsets selected are

{ B, M, Be, D, {B,M}, {B,D},{ D,M}, {D,Be} }

Table 4. Interest of a transaction

Itemset	Support	W-support	Interest metric
B	**0.8**	**0.78**	**0.624**
M	**0.8**	**0.81**	**0.648**
Be	0.6	0.65	0.39
D	**0.8**	**0.88**	**0.704**
E	0.2	0.19	0.038
C	0.4	0.44	0.176

If the value of *min_int* is 0.5, we obtain three interesting itemsets; these are: {B}, {M}, {D}. It proves that the interestingness of an itemset is made up of its weight and support.

5. Result and Discussions

In order to appraise the performance of the proposed algorithm, we conducted an experiment on various datasets using Proposed weighted algorithm and Traditional Apriori Algorithm. The algorithms were implemented in java and tested on a Windows XP Professional platform. The results show that the performance of our algorithm is much better than that of the existing algorithm in terms of time and number of rules. The experiments estimates performance of the WeightedAR according to three criteria: CPU time requirements, interest and rules as shown in from fig2 to fig 4.

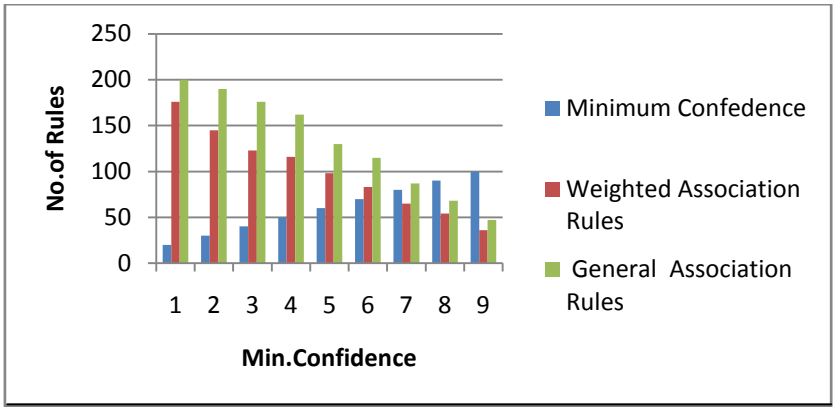

Figure2: Minimum confidence Vs General and Weighted Rules

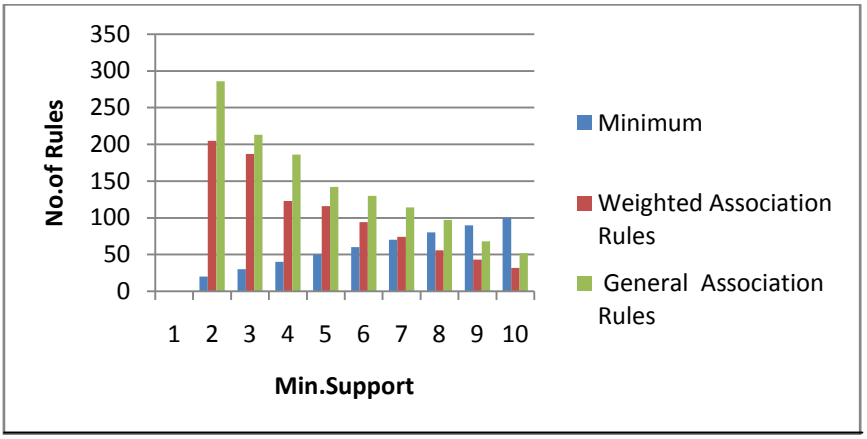

Figure3: Minimum Support Vs General and Weighted Rules

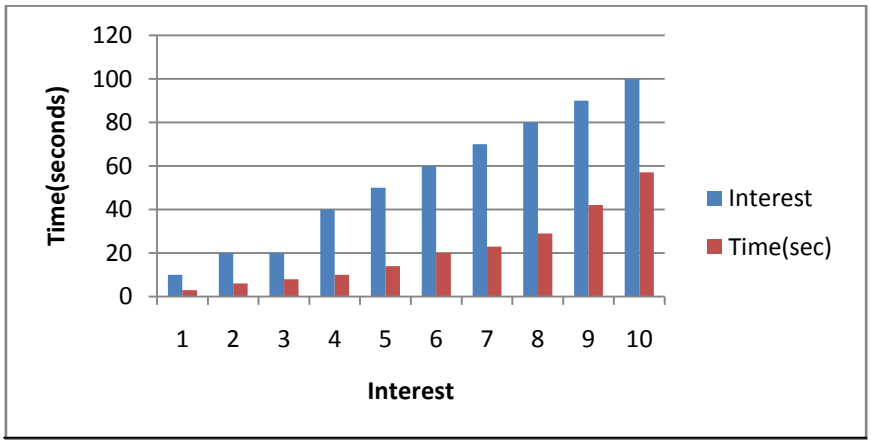

Figure 4: Minimum Interest Vs Time

6. Conclusions

In this paper we have developed a novel framework of weighted association rule mining. First, the HITS algorithms 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 of association rules are defined in analogy to the definition of support. 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. Through comparison, we found that our model and method address emphasis on high-quality transactions. Some interesting patterns discovered when the hub weights of transactions are taken into account.

References

[1] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," The International Conference on Very Large Data Bases, pp. 487-499, 1994.

[2] C. H. Cai, A. W. C. Fu, C. H. Cheng, and W. W. Kwong "Mining association rules with weighted items," The International Database Engineering and Applications Symposium (IDEAS), pp. 68-77, 1998.

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

[7] M. Sulaiman Khan, M. Muyeba, and F. Coenen, "A weighted utility framework for mining association rules," The 2008 Second UKSIM European Symposium on Computer Modeling and Simulation, pp.87-92, 2008.

[8] S. Lu, H. Hu, and F. Li, "Mining weighted association rules," Intelligent Data Analysis, Vol. 5, No. 3, pp. 211-225, 2001.

[9] U.Yun, Efficient Mining of Weighted Interesting patterns with a strong weight and or support affinity, Information Sciences177(17)2007 3477-3499

[10] Feng Tao, Fionn Murtagh and Mohsen Farid, "Weighted Association Rule Mining using Weighted Support and Significance Framework", Department of Electronics and Computer Science, University of Southampton,Southampton, UK

[11] Wei Xie and Jing Wu "Mining Positive and Negative Weighted Association Rules in Medical Records without User-specified Weights Based on HITS Model", 2010 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010), 978-1-4244-6498-2/10 2010 IEEE

