Weighted Association Rule Mining without Pre-determined Weights

R. Madhuri[1]
[1]Asst.professor, CSE Department, GMR Institute of Technology, A.P, India,

P. Pushpa Latha[2]
[2]Asst.professor, CSE Department, GMR Institute of Technology, A.P, India,

K. Prasad Rao[3]
[3]Asst.professor, CSE Dept, Aditya Institute of Technology, A.P, India,

Abstract

Compared to traditional frequent pattern mining, group pattern mining incurs much more computation and storage. To discover group patterns from a set of mobile users, we have proposed an efficient fast mining algorithm that is based on frequent pattern mining. In this paper, we introduced w-support, a new measure of item sets in databases with only binary attributes which does not require any pre determined weights. These weights are completely derived from the internal structure of the database based on the assumption that good number in transactions consists of good items. A fast miming algorithm is given and experimental results show that w-support can be worked out without much overhead, and interesting patterns may be discovered through this new measurement.

Index Terms: *Apriori, Data mining, HITS, Link based association rule, Weighted association rule.*

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.

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 pre-assigned weights. However, most data types do not come with such pre-assigned weights, such as Web page 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 to bipartite graphs. Therefore, w-support is distinct from weighted support in weighted association rule mining, where item weights are assigned. Furthermore, a new measurement framework of association rules based on w-support is proposed. We used HITS and Apriori algorithm on mobile and laptop database. Experimental results show that w-support can be worked out without much overhead, and interesting patterns may be discovered through this new measurement. This work would be better useful for marketing people.

2. Data Mining Techniques:

2.1. Association Rule Mining:

Association Rule mining is one of the fundamental data mining method. Agarwal first introduced the problem of association rule mining in 1993. Since then it is one of the most popular research area on the field of knowledge discovery. The association rule-mining problem is commonly known as the market basket analysis, but there are several applications that use association rules as well i.e. biological research areas, telecommunication and network analysis etc.

Regarding the diversity of the applications that use association rule mining, several algorithms have been developed. All of these algorithms have their own advantages and disadvantages, so it is useful to compare them.

Goal of Association rule mining helps in finding interesting association relationships among large set of data items. The discovery of such associations can help develop strategies to predict.

An Association Rule is a rule of the format

LHS (left hand side) => RHS (ride hand side).

Let X and Y are two item sets where X, Y $\subseteq$ I (both side contains sets of items) and X∩Y=Ø (don't share common items).

Briefly, an Association Rule is an expression;

X =>Y, where X and Y are set of items.

Each rule is assigned with two interesting measures:

- Support
- Confidence

Let A= { l1, l2, l3 ……ln}. be the set of items. Let T be the set of task relevant data containing database D with transactions Ti, where Ti is a subset of the items in A.

• Support:

A transaction T is said to support an item I1. If I1 is present in T it is said to support a subset of items X$\subseteq$A, if it supports l1 in X. An item X$\subseteq$A has support s in T denoted by s(X)T , if s% of transactions in T support X.

Support X =>Y = σ (XUY) = s (XUY) / |D|.

(Where probability denotes as σ).

• Frequent Item set:

Let T be the transaction database and s be the user specified minimum support. An item set X$\subseteq$A is said to be frequent with respect to s, if $s(X)_T \geq \sigma$.

• Confidence:

The confidence of an association rule is defined as the measure of the probability of an item set dependency on the other item sets in the association rule.

Confidence X UY = σ (X | Y) = s (XUY) / s(X).

• Association Rule:

For a given transaction database T, an association rule is an expression of the form X=>Y, where X and Y are subsets of A and X=>Y holds with confidence τ,if τ%of transactions in T that support X also support Y. The rule X=>Y has support σ in the transaction set T if σ% of the transactions in T support XUY.

Generally association rule mining is performed in two steps:

• Find all frequent item sets:

The basic foundation of Association Rule algorithm is fact that any subset of a frequent item set must also be a frequent item set. i.e., if {AB} is a frequent item set, both {A} and {B} should be a frequent item set. Iteratively find frequent item sets with cardinality from 1 to k (k-item set)

• Use frequent item sets to generate strong rules having minimum confidence:

The first step, find all frequent item sets, is expensive in terms of computation, memory usage and I/O resources. Much of the research effort in these algorithms has been related to improving the efficiency of this first step.

The second step, use frequent item sets to generate strong rules having minimum confidence, is relatively certain, but it can still be very expensive when solving real-world problems.

The aim of the algorithm is to discover all association rules with

Support≥ minimum Support

Confidence ≥minimum Confidence

These rules are very simple, understandable and useful for association rule mining. However, the determination of these rules is a major challenge due to the very large data sets and the large number of potential rule.

2.2. Weighted Association Rule Mining:

The methodology of weighted association rule mining 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 algorithm which is shown below 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.

2.3. Hyperlink-Induced Topic Search (HITS) Algorithm:

Kleinberg's HITS algorithm, “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.

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.

The Hub score and Authority score for a node is calculated with the following algorithm:

- Start with each node having a hub score and authority score of 1.
- Run the Authority Update Rule
- Run the Hub Update Rule
- Normalize the values by dividing each Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores.
- Repeat from the second step as necessary.

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. We have taken the models of mobiles and laptops for finding the hub values which are very important to find the number of hits for each and every model.

Here we have taken the sample data of mobiles and found the hub weights of each model specified.

Following hub weights are obtained for each mobile transactions

Table 1. Hub weights of Mobile Transactions

TID	Transaction	Hub weight
1	X5-01, lumia 900, Xperia mini pro, Txt Pro	0.518
2	Xperia mini pro, 7210, X2-02	0.436
3	X5-01, lumia 900	0.233
4	X5-01	0.148
5	Xperia mini pro, 7210, X2-02, p350	0.544
6	X5-01, X2-02, p350	0.412

In HITS algorithm, as the iteration converges, the authority weight auth(i)=∑T:i€Thub(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:

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

$$\text{Wsupp (x)} = \frac{\sum_{\mathbf{T:\ XcT\Lambda T\epsilon D}} \text{hub(T)}}{\sum_{\mathbf{T:\ T\epsilon D}} \text{hub(T)}}$$

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.

2.4 Mining the Most Interesting Rules:

Several algorithms have been proposed for finding the “best,” “optimal,” or “most interesting” rule(s) in a database according to a variety of metrics including confidence, support, gain, chi-squared value, gini, entropy gain, laplace, lift, and conviction. In this paper, we show that the best rule according to any of these metrics must reside along a support/confidence border. Further, in the case of conjunctive rule mining within categorical data, the number of rules along this border is conveniently small, and can be mined efficiently from a variety of real-world data-sets. We also show how this concept can be generalized to mine all rules that are best according to any of these criteria with respect to an arbitrary subset of the population of interest. We argue that by returning a broader set of rules than previous algorithms, our techniques allow for improved insight into the data and support more user-interaction in the optimized rule-mining process.

3. Problem Description:

In classical models users can rate the products in the website based upon binary attributes such as like or dislike only. Here the top most products are getting by using pre-assigned weights, so apart from top most products it can display the top most products according to the support system. Already existing model is having some limitations such as

- There is no calculation of quality transaction.
- There is no estimation of awareness of user.
- Chance of getting quality product is very less.
- It gives equal priority to all transaction items

In this paper we eliminate equal priority of each transaction item and the priority of transaction items are calculated based on the number of clicks made by the user. Once a registered user rate a product, analysis is made on the database. By considering this rating, we update product rating every time in database. We calculate the W-support values of products to display the top most products. Finally an analysis is made over top products rated by the users and frequent products are mined using an Apriori algorithm to get the perfect results. Hence the results that were observed are

- It provides quality of transactions for the particular user.
- It provides w-support value based on the transaction items.
- By using this system easily we can maintain different user details.

We introduce w-support, a new measure of item sets in databases. 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. The concept of association rule was proposed the support measurement framework and reduced association rule mining to the discovery of frequent item sets. A fast mining algorithm, Apriori, was proposed. Much effort has been dedicated to the classical (binary) association rule mining problem. These algorithms strictly follow the classical measurement framework and produce the same results once the minimum support is given.

4. Algorithms:

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

```
Initialize auth (i) to 1 for each item i
for (l=0; l<num_it; l++) do begin
    auth' (i) =0 for each item i
    for all transactions tϵD do begin
        hub (t) = i: iϵt auth (i)
        auth' (i) +=hub (t) for each item iϵt
    End
        auth (i) =auth' (i) for each item I, normalize auth
End
$L_1$= {{i}: wsupp (i)>=minsupp}
for (k=2; $L_{k-1}$≠ǿ; k++) do begin

  $C_k$=apriori-gen ($L_{k-1}$)

    for all transactions tϵd do begin
            $C_t$=subset ($C_k$, t)
            for all candidates c ϵ $C_t$ do
             C.wsupp+=hub (t)
             H +=hub (t)
    End
    Lk= {cϵ$c_k$|c.wsupp/H>=minwsupp}
End
```

4.2. Apriori algorithm:

For finding frequent patterns, associations, correlations, or causal structures among sets of items in transaction databases, relational databases, and other information repositories we use an influential algorithm known as Apriori Algorithm. Apriori algorithm is used in Basket data analysis, cross-marketing, catalog design, Loss-leader analysis, clustering, classification etc.

In order to express the quality of an association rule one uses measures such as

Support: This metric determines how often a rule is satisfied in the transaction database. It is obtained by dividing the support count for $X \rightarrow Y$ by the total number of transactions.

Confidence: This metric determines how often items in Y appear in transactions that contain X.

Algorithm

Join Step: Ck is generated by joining Lk-1with itself
Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

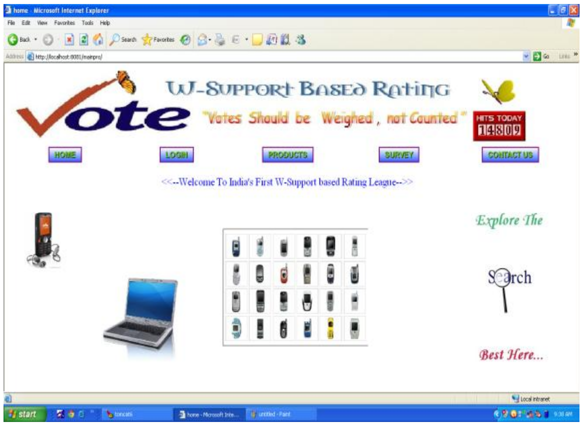
Pseudo-code

```
$C_k$: Candidate itemset of size k.
$L_k$ : frequent itemset of size k.
$L_1$ = {frequent items};
for ($k$ = 1; $Lk$ !=Ø; $k$++) do begin
$C_{k+1}$ = candidates generated from $L_k$;
for each transaction $t$ in database
do
increment the count of all candidates in
$C_{k+1}$ that are contained in $t$
$L_{k+1}$ = candidates in $C_{k+1}$ with min support
end
return $\cup_k L_k$;
```

Generate Candidates

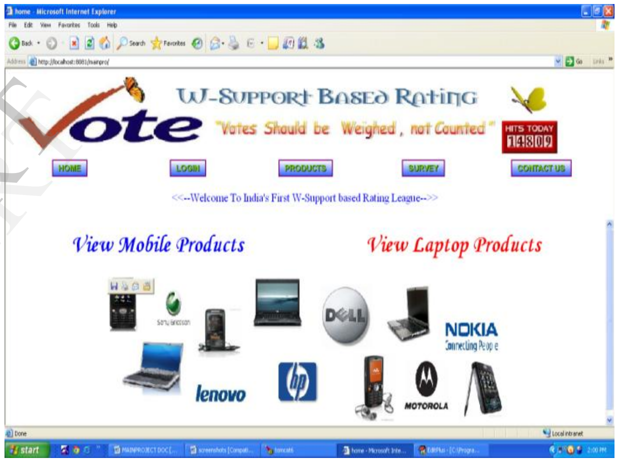
```
Suppose the items in $L_{k-1}$ are listed in an order
Step 1: self-joining $L_{k-1}$
        insert into $C_k$
        select p.item1,p.item2, …,p.itemk-1, q.itemk-1
         from $L_{k-1}$ p, $L_{k-1}$ q
        where p.item1=q.item1, …, p.itemk-2=q.itemk-2, p.itemk-1 < q.itemk-1
Step 2: pruning
         For all itemsets c in $C_k$ do
         For all (k-1)-subsets s of c do
     if (s is not in $L_{k-1}$) then delete c from $C_k$
```

5. Results:



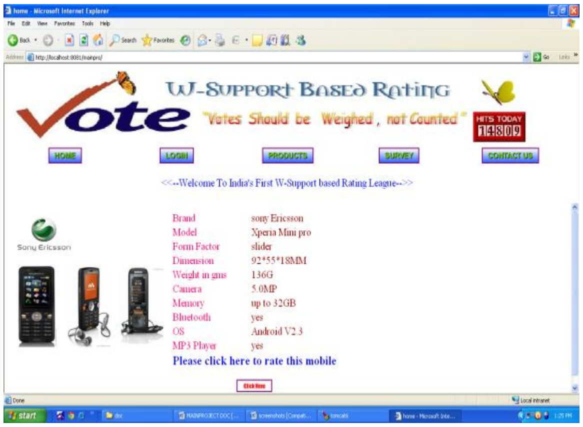
<<--Welcome To India's First W-Support based Rating League-->>

Figure 5.1 Home Page



<<--Welcome To India's First W-Support based Rating League-->>

Figure 5.2 Product details Page

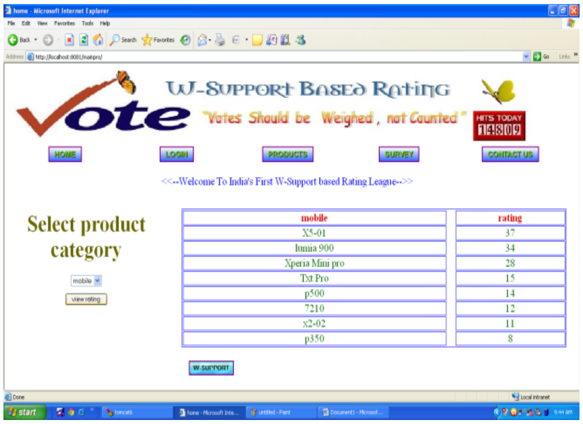


<<--Welcome To India's First W-Support based Rating League-->>

Brand	sony Ericsson
Model	Xperia Mini pro
Form Factor	slider
Dimension	92*55*18MM
Weight in gms	136G
Camera	5.0MP
Memory	up to 32GB
Bluetooth	yes
OS	Android V2.3
MP3 Player	yes

Please click here to rate this mobile

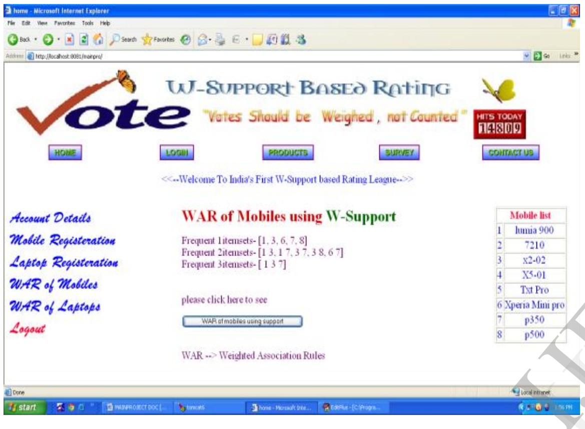
Figure 5.3 Mobile details page



<<--Welcome To India's First W-Support based Rating League-->>

mobile	rating
X5-01	37
lumia 900	34
Xperia Mini pro	28
Txt Pro	15
p500	14
7210	12
x2-02	11
p350	8

Figure 5.4 Mobile ranking page



<<--Welcome To India's First W-Support based Rating League-->>

WAR of Mobiles using W-Support

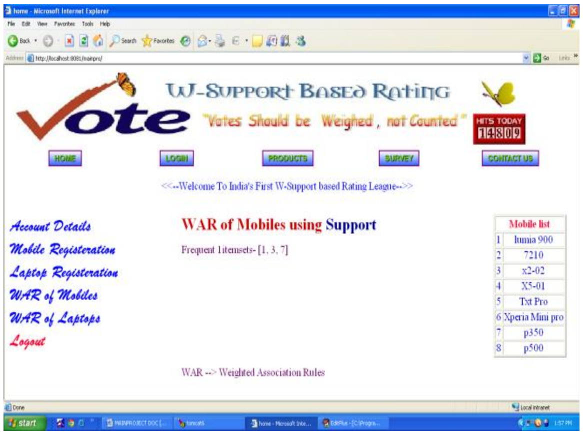
Frequent 1itemsets- [1, 3, 6, 7, 8]
Frequent 2itemsets- [1 3, 1 7, 3 7, 3 8, 6 7]
Frequent 3itemsets- [1 3 7]

please click here to see

WAR --> Weighted Association Rules

	Mobile list
1	lumia 900
2	7210
3	x2-02
4	X5-01
5	Txt Pro
6	Xperia Mini pro
7	p350
8	p500

Figure 5.5 WAR of mobiles using W-Support



<<--Welcome To India's First W-Support based Rating League-->>

WAR of Mobiles using Support

Frequent 1itemsets- [1, 3, 7]

WAR --> Weighted Association Rules

	Mobile list
1	lumia 900
2	7210
3	x2-02
4	X5-01
5	Txt Pro
6	Xperia Mini pro
7	p350
8	p500

Figure 5.6 Frequent item sets using measure

6. Conclusion:

In this paper we have developed 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 of association rules are defined in analogy to the definition of support. 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. 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] G.D.Ramkumar, S.Ranka, and S.Tsur, “Weighted Association Rules: Model and Algorithm,” Proc. ACM SIGKDD, 1998.

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

BIOGRAPHIES

	Madhuri Ravi, is M.Tech in Computer Science from JNTU, A.P., India. She has vast experience in Computer Science and Engineering areas. She is presently working as Assistant Professor in Department of Computer Science and Engineering, GMRIT, Rajam, A.P, India. Her area of research includes Data Mining, Computer Networks and Computer Architecture.
	Pushpa Latha Palli, is M.Tech in Computer Science from Andhra University, A.P., India. She has vast experience in Computer Science and Engineering areas. She is presently working as Assistant Professor in Department of Computer Science and Engineering, GMRIT, Rajam, A.P, India. Her area of research includes Data Mining, Web Technologies and Emerging Technologies.
	Prasad Rao Karu, is M.Tech in Computer Science from JNTU, A.P., India. He has vast experience in Computer Science and Engineering areas. He is presently working as Assistant Professor in Department of Computer Science and Engineering, AITAM, Tekkali, A.P, India. His area of research includes Data Mining, Databases, Web Technologies.