affording to the access frequency. The paths visited most frequently are the chosen navigation paths [5].

To find associations, patterns and correlation between pages that are accessed jointly during a user session, an association rule mining has applied. The possible relationships between pages that are viewed together even if they are not directly connected are discovered by the association rule mining. So that the constructed model envisages users' web page requests, assists the users towards browse web pages, mine user access patterns on web and reduce the access time. Previous ARM algorithms[6][7] assumed all the web pages has same importance in web data, and ignored the significance of the pages in a user session and the period spent for seeing page by user.

The customary association rule problem extended by permitting a weight associated to each page based on weighted association rule mining approach. In *weighted association rule* mining the traditional binary weights are not considered, instead the period spent by each user on each page and visiting frequency of each page is used to assign a weight of a page. The *weighted mining* methodology is two-fold process: first, the association rule analysis method explores a database of URL data concerning access to electronic sources, and determines the different rules that described between set of pages in the website. In conclusion, the recommendation engine resolves the furthermost related rules between pages to the dynamic user session with the uppermost confidence [8] [9].

The rest of this paper is structured as follows: Section 2 exemplify the related work. Section3 describes the algorithm PNTH (Preferred Navigation Tree with HITS). The proposed association rule mining technique and its definitions present in section4. The example illustrated on a web data in section 5 and experimental results are presented and analysed. Lastly, the final remarks are made in the last section.

2. Related Work

Rakesh, Tomasz and Arun [10] proposed an algorithm which combines buffer management, novel estimation and pruning techniques, also produce results by applying this algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm. The measures taken for this algorithm generate association rules based on a count obtained for each item from database. In the worst case, this problem produces an exponential complexity. By using this procedure, A database of 'm' items in which every item appears in every transaction, '2^m'large item sets.

Rakesh Agrawal and Rama Krishnan Srikant [11] proposed the hybrid algorithm, called Apriori Hybrid. Apriori Hybrid also has outstanding scale-up properties with respect to the number of items in the database and the transaction size. As a result this algorithm is an un-weighted mining algorithm that may omit some weight factors in databases, and if the minimum support declines, the execution times of all the algorithms surge because of increases in the total number of candidate and large item sets.

J. M. Kleinberg [12] proposed a new approach**,** link based structure of a hyperlinked environment can be a rich source of information about the content of the environment, provided we have effective means for understanding it. In this paper developed a set of algorithmic tools for mining information from the link structures of such environments, and report on experiments that demonstrate their effectiveness in a variety of contexts on the WWW. The fundamental issue we address within this framework is the distillation of broad search topics, through the discovery of "authoritative" information sources on such topics. They proposed and tested an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of "hub pages" that join them together in the link structure. An Additional direction of work on the combination of links into www search is the creation of search formalisms accomplished of dealing queries that contain predicates over both text and links. The weighted concept which is stemmed from the assumption of HUB and AUTHORITY introduced in this topic.

K.Wang and M.Y.Su[13] proposed an approach based on historical transactions for a necessary problem in business and other applications is ranking items with respect to some notion of profit. The trouble is that the yield of one item not only comes from its own sales, it is called cross-selling effect. A novel approach presented for item ranking based on hub/authority ranking. This ranking approach solves two selection problems size-constrained selection and cost-constrained selection. This method find profitable items called weighted items in the presence of cross selling effect. Here the weight is pre assigned value.

G.D. Ram kumar, S. Ranka, and S. Tsur,[14] implemented an algorithm for Weighted Association rules which addresses a number of problem areas. It does not make sense to assign equal importance to all the items involved in the market basket analysis. The items with more profit may appear less times but it very important; the weighted support of this item is more significant in mining. The proposed work assign a weight value to each item and find the significant items sets in mining process. Experimental results shown that

the weighted mining process found more interesting rules that are not found by traditional support measure.

C.H. Cai, W.W. Kwong and C.H. Cheng, [15] generated association rules with weights to Discovery of association rules has been originate valuable in many applications. All items in a basket database are treated as uniform in previous study. We simplify this to the case wherever items are given weights to replicate their status to the user. The weights may resemble to special promotions on some products, or the productivity of different items. We can mine the weighted association rules with weights. The downward closure property of the support measure in the un-weighted case no longer exists and earlier algorithms cannot be applied. In this paper, two new algorithms MINWAL (W) and MINWAL (O) introduced to handle this difficulty. In these algorithms we make use of a metric called the k-support bound in the mining process. Investigational results show the effectiveness of the algorithms for huge databases. The performance evaluation has been done on these two algorithms. We found that MINWAL(O) outperforms MINWAL(W) in most cases, but MINWAL(W) performs better for the special case with only 0/1 item weights.

W. Wang, J. Yang, and P.S. Yu [16] proposed an algorithm for the convention association rule problem extended by allowing a weight to be coupled with every item in a operation, to reflect interest/intensity of the item in the transaction. This provides us in turn with an opportunity to associate a weight parameter with each item in the resulting association rule is referred as weighted association rule (WAR). WAR not only improves the confidence of the rules, but also provides a method to do more efficient target marketing by identifying or segmenting customers based on their possible degree of loyalty or volume of purchases. This approach mines WARs by first ignoring the weight and discover the frequent item sets (via a traditional frequent item set discovery algorithm), and is followed by introducing the weight during the rule generation. It is shown by investigational results that our approach not only results in shorter average execution times, but also produces higher quality results than the generalization of previous known methods on quantitative association rules.

F. Tao, F. Murtagh, and M. Farid, [17] extracted a model for significant data items. The issues of discovering significant binary relationships in transaction datasets in a weighted setting, where each item is allowed to have a weight, conventional model of association rule mining is adapted to handle weighted association rule mining problems. The goal is to steer the mining focus to those significant relationships involving items with significant weights rather than being flooded in the combinatorial explosion of insignificant relationships. In the iterative process of generating large item sets, we recognize the challenge of using weights. 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 is developed based on the improved model called WARM (Weighted Association Rule Mining). The algorithm is both efficient and scalable during discovering significant relationships in weighted settings as illustrated by experiments performed on replicated datasets.

Srikant, Rakesh. Agrawal [18] addressed the problem of mining generalized association rules introduced for a given a large database of transactions, where each transaction consists of a set of items, and a taxonomy (is-a hierarchy) on the items, so as to uncover associations between items at any level of the categorization. For example, given a categorization that says that jackets are-outerwear is-clothes, we may infer a rule that people who buy outerwear tend to buy shoes". This rule may hold even if rules that people who buy jackets tend to buy shoes", and people who buy clothes tend to buy shoes" do not hold. A clear explanation to the problem is to add all relations of each item in a transaction to the transaction, and then run any of the algorithms for mining association rules on these extended transactions". However, this Basic algorithm is not very fast; so two algorithms present Cumulate and EstMerge, which run 2 to 5 times faster than Basic (and more than 100 times faster on one real-life dataset) and also present a new interest-measure for rules which uses the information in the taxonomy. Given a user-specified minimum interest level", this measure prunes a large number of redundant rules; 40% to 60% of all the rules were pruned on two real-life datasets.

3. Preferred Navigation tree with Hits

The concept of preference to web pages is taken as if there are many different options to leave a page, the options that are selected most frequently and next page viewed reveal *user interest* and *preference*. The projected evaluation for judging the degree of user awareness in web page includes four factors: *authority, hub, selection preference and time preference* of a page. The first two factors are used to rank the web pages via the HITS algorithm. This algorithm ranks web pages by analysing the in-degree and out-degree of web pages. In this algorithm, web pages pointed by

many other pages (in degree) are called *authorities*, while web pages that point to many other pages (out degree) are called hubs. Authorities and hubs have mutual effects on each other. They can be depicted as a bipartite graph shown in Fig 4. Let *auth (p)* and *hub (p)* symbolize the authority and hub values of page p, respectively. This algorithm performs the following equations:

$$auth(p) \longleftarrow \sum_{q:(q,p)\in E} hub(q) \quad \text{Equation…..1}$$

$$hub(p) \longleftarrow \sum_{q:(q,p)\in E} auth(q) \quad \text{Equation…..2}$$

Wherever (p,q) ∈ E denotes that there is a relationship from page p to page q . It is assigning weights to web pages.

The other two factors are *selection preference* and *time preference* is calculated by the following equations:

Let SP denote selection preference. SP is defined as

$$SP_k = C_k / \left(\left(\sum_{i=1}^{n} C_i \right) / n \right) \quad \text{Equation…..3}$$

Where, C is the times that users browse from a parent page to the *k-th* child page. Where, C_k is the times that users browse from a parent page to the k-th child page. $C_i$is the times that users visit the i-th child page from the similar parent page. N is the number of child pages that the parent page offers.

Ex: In a web data after browsing page *A* 3 child pages {*B, C, D*} can be visited. The time for visiting page *A* is 25, and the time for visiting page {B, *C, D*} are {10, 10, 5} respectively. The selection preference of page *B* is calculated as

SP_B=10/ [(10+10+5)/3] = 1.2

The total of the times of all child pages is identical to the times of the parent page (page A).

Let *TP* represent *time preference*. The definition of *TP* is shown as

$$TP_k = T_k / \left(\left(\sum_{i=1}^{n} T_i \right) / n \right) \quad \text{Equation…..4}$$

Where T_k is the browsing time of the k^{th} *child* page, a user visits from a parent page.

T_i is the browsing time of the *i-th* child page a user visits from the same parent page. N is the no. of child pages that the parent page offers.

Ex: The child pages {B, C, D} are visited subsequent to browsing page A. The browsing time on page A is 25, and the browsing time on page {B, C, D} are {34, 78, 20}, respectively.

The *time preference of page B* is

TP_B = 34 / [(34+78+20) / 3] =0.77

If the *k-th* child page is visited from the parent page p based on the four factors stated above, the original preference is modified as

$$PH_k = hub_p \times auth_k \times SP_k \times TP_k \quad \text{Equation..5}$$

Algorithm1: Calculates Authorities and hubs
Input: A web site and the number of iterations *nit.*
Output: Hub (p) and Auth (p) for each page p.

(1) Auth(i) = 1 for each page i ;
(2) For (l = 0; l < nit ; l++) do
(3) $Auth^1$(i) = 0 for each page i ;
(4) For (k=0; k <no pages; k++) do
(5) Hub(k) = $\Sigma_{p\in o(p)}$ Auth(p) , where o(p) denotes the set of all pages pointed to by page p
(6) End For
(7) For (k=0 ; k<no pages ; k++) do
(8) $Auth^1$(k) =$\Sigma_{p\in I(p)}$ Hub(p), where I(p) denotes the set of all pages that point to p
For
(9) Auth(i)= $Auth^1$(i) for each page I
(10) Normalize auth
(11) End For

Fig 1: Algorithm for calculating Authorities and hubs

There are two Algorithms involved to find the preference value. Algorithm1 shown in Fig 1 calculates Hubs and Authorities of web pages of websites and Algorithm2 shown in Fig 2 calculate actual preference value.

Algorithm2: Calculates Preference value.
Input: User access session database D={(i , Si)| Si=(Si1,Si2, ……..Sin), where Sij is a web page ,n is the size of the session Si.
Output: Preference value (PH).

(1) Count is the no. of user sessions in Ds.
(2) For (i = 1; i <num_sessions; i++) do
(3) Let C_k is the Count of k^{th} child page
(4) For (j=1; j<num_childs ; j++) do
(5) C(j)+=C(j)
(6) End For
(7) SP(i)=Ck/(C(j)/nc);
(8) End For
(9) Time Preference (TP) is calculated in the same manner
(10) For (k=1; k<num_pages; k++) do
(11) Calculate Preference value using
(12) PH(k)= Hub(k)*Auth(k)*SP(k)*TP(k)
(13) Store the web page preference for each page
(14) End For

Fig 2: Algorithm for calculating Preference value

4. Association rules in the web

The principle of mining association rules is to find out, which web pages are usually visited together in a session. Association rules are find correlation relationships and interesting associations among large set of items. In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. Association rules discovery is based on the sessions for the web usage mining. Each session is interpreted as a transaction and occurrence of the web pages in the session is relevant to the existence of the items in one basket. in addition, there is a possibility of maintaining a sequence in which web pages were viewed in web usage mining. These web pages sequences are called paths, as the user was following pages in specific order until session ended. When association mining is applied to web usage mining along with web structure mining, one can determine correlations among web pages, exciting access patterns and preferred traversal paths.

The algorithm assumes that the traversal path follows the link structure of the graph, where a link structure is representation of web pages along with the embedded hyperlinks. Hence, based on the actual user browsing activity on the web site, needs to be constructed alternatively. During one session, a user browsing sequence or a traversal path is an ordered list of web pages accessed by a user. From the PNTH, User preferred traversal paths are extracted. Association rules of each URL will be extracted from the active user sessions will be calculated upon the preferred path between pages. In the web transactions, based on the navigational patterns of users, association rules capture relationships among pages. Each web page viewed as an "item", and set of web pages accessed by a user within a session as a "transaction". So the idea of mining association rules is to unearth out which web pages are usually visited together in dissimilar sessions.

Let D is user access session database D ={(i, Si) | Si = (S_{i1}, S_{i2}, ---------- , Sin) , where Sij is a Web page, n is the size of the session Si}.

Def. 1: A set *X* of pages $di \in D$ is called the *page set X*. The number of pages in a *page set* is called the *length of the page set*. A page set with the length *k* is denoted as the *K-page set*.

Def. 2: The *i*-th user session *Si* is the page set containing all pages viewed by the user during the *i*-th visit on the web site; $Si \subset D$. *SS* is the set of all user sessions gathered by the system, $Si \in SS$. Each session must consist of at least two pages *card* (*Si*) ≥2. A session *Si* contains the page set *X* if and only if $X \subseteq Si$.

Def. 3: An *association rule* is the relationship X→*Y*, where $X \subseteq D$, $Y \subseteq D$ and $X \cap Y = \phi$. An association rule is described by two measures: *support* and *confidence*.

The rule is defined by X and Y where the page set X is the set of antecedent items and Y is the consequent item of the rule $X \Rightarrow Y$.

The *support* is the no. of transactions that include all items in the antecedent X and consequent Y divided by the total no. of transactions. It is the probability that a transaction contains both X and Y.

The *support* of the rule $X \Rightarrow Y$ is defined as

$$Supp(X \Rightarrow Y) = \frac{\text{(Sessions that contain every item in X and Y)}}{\text{All sessions}}$$

The *support* is reflexive. That is, the support of the rule

$X \Rightarrow Y$ is the same as the support of the rule $Y \Rightarrow X$.

The *confidence* is the ratio of the no. of transactions that include all items in the consequent Y as well as the antecedent X divided by the no. of transactions that include all items in the antecedent X. It is the conditional probability that if a transaction contains X, it also contains Y.

The *confidence* of an association rule $X \Rightarrow Y$ defined as

$$Conf(X \Rightarrow Y) = \frac{\text{(Sessions that contain every item in X and Y)}}{\text{(Sessions that contain the items in X)}}$$

The problem of mining association rules that are strong enough and have the support and confidence value greater than given thresholds: minimum support threshold and minimum confidence (preference) threshold. Consequently, association rule mining is to find out all rules with support and preference above some given thresholds.

The goal of this study is to help the web designers to improve their website by determining related link relations in the website. The results and findings of this experimental study can be used by the web designer in order to plan the upgrading and enhancement to the website. Web mining gives decision support for the changes in the web site navigational structure

The followings are concise conclusion of the proposed study:

•To obtain the interest of the web visitors, the support and the confident values of extracted rules are considered. Accordingly, analysing the visitor approach can increase the no. of hits.

•To prove the web site usability, the patterns which are extracted from the web links can be analysed and the new motivating thoughts can be used for augmented the competence of the web site.

In the results of association mining, we can locate pairs of groups of web pages "visited" together. To develop successful marketing strategies and place their websites for better use, Association rules are more and more supportive for the organizations.

5. Example

This section provides an example for mining process with the proposed algorithm.

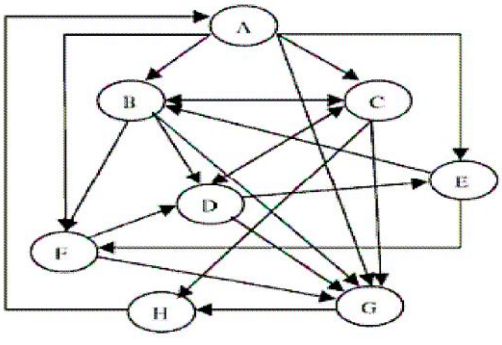

Fig 3: Sample web site

Table1 illustrate the browsing sessions and the elements in each session represented as (web page, Time duration), where Time duration denotes the time spent viewing the selected web page.

Table 1: Browsing sessions

UID	Browsing sessions
1	(A,23), (B,68), (D,98), (E,130)
2	(A,45), (B,89), (D,102)
3	(A,27), (B,56), (F,86)
4	(A,32), (B,87), (D,45), (G,115), (H,118)
5	(A,30), (C,65), (H,78)
6	(A,12), (C,34), (G,32)
7	(A,78), (C,89), (G,110), (H,123)
8	(A,10), (B,24), (G,45), (H,34)

Now, construct the Bipartite Graph Shown in Fig 4, which can be represented as Hubs and Authorities from the Fig 3. The relationship between Hubs and Authorities are called a mutually reinforcing relationship.

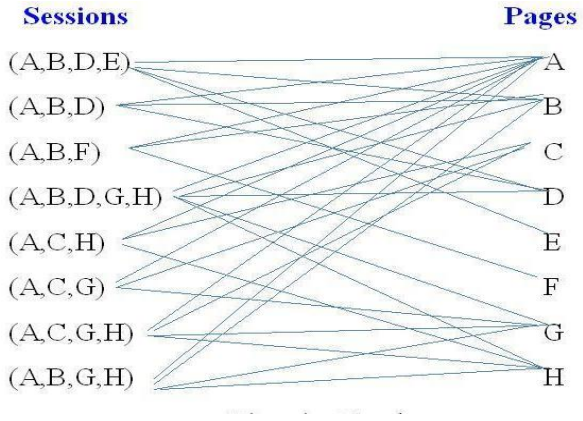

Fig 4 Bipartite Graph

Table 2: Calculated Values

URL	HUB	AUTH	SP	TP	PREFERENCE
A	1.29	0.87	1.0	1.0	0.82
B	1.23	1.13	1.25	1.01	1.84
C	1.22	1.14	0.75	0.98	1.08
D	1.16	1.12	1.25	1.09	2.85
E	1.07	1.08	0.4	1.27	0.73
F	1.08	1.13	0.375	1.02	0.55
G	0.96	1.2	4.1	1.58	8.39
H	0.94	1.03	2.0	2.02	0.98

The values in Table 2 are calculated using the equations specified in section 3.

The classical association rule mining evaluates the *support* value by counting the web pages browsing in user sessions.

The *support* of the X is defined as

$$Supp(X) = \frac{\text{The page X is browse in number of sessions}}{\text{All sessions}}$$

5.1 Experimental Evaluation

The *support* and *preference* thresholds are compared, the support thresholds do not considered the structure of the web pages but in the preference threshold doing that. The differences are listed in the Table 3

Table 3: Calculated Support and preference values

URL	SUPPORT	PREFERENCE
A	1.0	0.82
B	0.625	1.84
C	0.375	1.08
D	0.375	2.85
E	0.125	0.73
F	0.125	0.55
G	0.5	8.39
H	0.5	0.98

All mined paths are {A, B} in classical association but in this approach {B,C} and {C,G} paths are preferred by considering structure of web site.

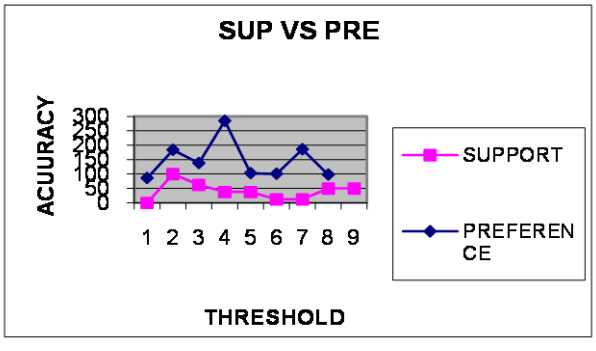

Fig. 5: Comparisons of Support and Preference

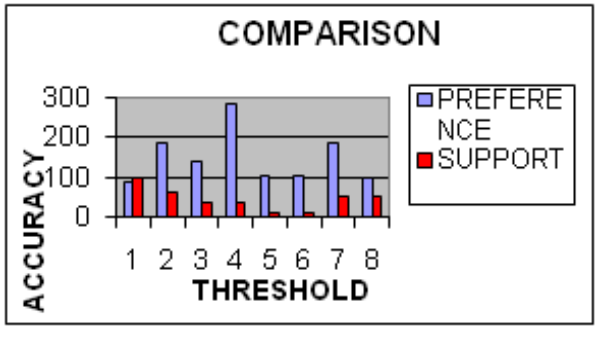

Fig. 6: Accuracy comparisons

We balance the accurateness of the proposed algorithm with support threshold based on the different resources. According to the Fig.5&6 are generated from Table3, PNTH algorithm is more precise than classical mining. It should be perceived that the lower preference thresholds are varying for the support threshold.

6. Conclusion

This manuscript proposes a novel measure called preference, which is to be allowing for web structure to mine the associations between the web pages. Based on, this measure to study preferred web mining process through PNTH algorithm. We have presented a outline in association rule mining. First, the HITS algorithm is used to obtain the Hub and Authority weights. A new measure preference is defined the significance of web pages based on these weights. It differs from the traditional support in taking the time preference and selection preference perspective. Then, the *support* and *confidence* of association rules are defined. To extract association rules whose support and confidence are given thresholds, an Apriori- like algorithm is proposed.

References

[1] Cooley, R. et.al, "*Web mining Information and patterns discovery on the WWW*" in 9th IEEE Int. Conf. on Tools with Artificial Intelligence, pp. 558-567(1997)

[2] Kosala, R., et.al *Web Mining Research*: A Survey. ACM SIGKDD Explorations Newsletter 2, 1-15(2000)

[3] Agrawal, R., et.al, "*Mining sequential patterns"* In 11th Int. Conf. on Data Engg. Pp, 3-14(1995)

[4] J S Yeh, et.al "*Mining preferred trversal paths with HITS"* Springer Berlin, comp. sci., pp. 98-107(2009)

[5] Wu.R.,et.al "*Web mining of preferred Traversal patterns in Fuzzy Environments" i*n: RSFDGrC 2005-LNCS(LNAI),vol.3642,pp 456-465. Springer, Heidelberg(2005)

[6] Xing D., et.al, "E*fficient data mining for web navigation patterns"*, Information and software tech. no.46,pp55-63(2004)

[7] Sun K., et.al, "*Mining weighted association rules without pre assigned weights",* IEEE Trans on knowledge and data engg, Vol20,480-495(2008)

[8] Przemyslaw., et.al "*Mining indirect association rules for web recommendations*.," Int. J. Appl. Math. Comp. Sci., Vol 19.No 1, 165-186(2009)

[9] Resul Das., et.al, "*Extraction of interesting patterns through association rule miming for improvement of web usability*", Journal of EEE .,vol 9, No. 2(2009).

[10] Agrawal, et.al., "Mining association rules between sets of items in large databases". ACM SIGMOD Record, 22, pp. 207–216, 1993.
[11] Agrawal, R., et.al., "*Fast algorithms for mining association rules"* in 20th IntConf Very Large Data Bases, pp. 487-499(1994)
[12] J.M,Kleinberg "*AuthoritativeSources in a Hyperlinked Environment ",* Journal of the ACM, Vol. 46, No. 5, September 1999, pp. 604–632.
[13] K.Wanget.al.,*Item selection by "Hub-Authority" profit ranking*, the eight ACM SIGKDD int. conf. on knowledge discovery and data mining pp 652-657.
[14] G.D.Ramkumar,et.al.,"Weighted Association Rules: Model and Algorithm" KDD1998, 1998.
[15] C.H. Cai, et.al.,"Mining Association Rules with Weighted Items," Proc. IEEE Int"l Database Eng. and Applications Symp. (IDEAS "98), pp. 68-77, 1998.
[16] W. Wang, et.al "Efficient mining of weighted association rules (WAR)", Proc. of the ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, pp.270-274, 2000.
[17] F. Tao,et.al., "Weighted Association Rule Mining Using Weighted Support and Significance Framework," Proc. ACM SIGKDD "03, pp. 661-666, 2003.
[18] Srikant R, et.al., "Mining generalized association rules". In *Proc. Int., Conf. on Very Large Data Bases (VLDB)*, Zurich, Switzerland, 1995, pp.407–419.