Discovery of Association Rules at Multiple Levels

Ruchika Yadav
Department of Computer Science and Engineering
GJUS&T, Hisar- 125001, India

Jyoti Vashishtha
Department of Computer Science and Engineering
GJUS&T, Hisar- 125001, India

Abstract- Data mining is extraction of implicit, previously unknown, potentially useful information from the vast amount of data available in the data sets (databases, data warehouses or other information repositories). In previous studies the association rules are generated at the single conceptual level however mining association rules at multiple concept levels may lead to discovery of more specific and concrete knowledge from large transaction databases by extension of some of the existing rule mining techniques. In this paper, we discover multiple-level association rules using MLT2 algorithm. This algorithm discovers association rules for successive levels making use of rules already discovered for upper levels of concept hierarchy. In multilevel association rules we use different minimum support for different conceptual levels.

Keywords – Multiple-level Association rules, Data mining, Support, Confidence

I. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is sometime known as knowledge mining. Many other term carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology and knowledge discovery from data or KDD.[1]

The process of data mining is the use of algorithms to extract the useful information and patterns that the knowledge discovery process strives for. Several Techniques for data mining are Association rules, Clustering, Rule induction, Classification, decision trees, Neural networks etc.

Association rule mining is one of the very important tasks in data mining which is used for discovering relations between items. Association is the degree of relationship or involvement or the connection of objects and such connections sre termed as association rules. It can be applied in different areas like catalog design, store layouts, product placement, target marketing, business planning etc. For example, the association rules can be used to identify the customer buying habits in a market-basket analysis, like "if customers buy personal computer, they are more likely to buy an antivirus or printer as well". In general, every association rule must satisfy two user specified constraints called support and confidence. The support of a rule X=>Y is defined as the percentage of transactions that contain (X ∨ Y), where X and Y are disjoint sets of items from the given dataset. The confidence is defined as the ratio support (X ∨ Y)/support(X). Association rules revel the associative relationship among objects at multiple levels.[4]

II. MULTIPLE LEVEL ASSOCIATION RULES

In multiple-level association rule mining, the items in an itemset are characterized by using a concept hierarchy.[7] Mining occurs at multiple levels in the hierarchy. At lowest levels, it might be that no rules may match the constraints. At highest levels, rules can be extremely general.[2] Generally, a top-down approach is used where the support threshold varies from level to level (support is reduced going from higher to lower levels) .Sometimes, at primitive data level, data does not show any significant pattern. But there are useful information hiding behind. The goal of Multiple-Level Association Analysis is to find the hidden information in or between levels of abstraction.[3] There are two general requirements for multiple-level association rule mining: first provide data at multiple levels of abstraction and secondly find efficient methods for multiple-level rule mining.

Multilevel association rule mining works in two different processes. First of all it finds frequent items at multiple levels and then on the basis these frequent items it generate association rules. The first requirement can be full filled by providing concept taxonomies from the primitive level concepts to higher level. User will provide minimum support and confidence, if minimum support and minimum confidence thresholds at each level are uniform then it may lead to some undesirable result. Because, to find data items at multiple level under the same minimum support and minimum confidence thresholds will not give the desirable result. For example there is a hierarchy in which at level 0 there is food, at level one there are bread, milk and fruit and at level 2 we further put the various brands of these items. Large support is more likely to exist at high concept level such as bread and butter rather than at low concept levels, such as a particular brand of bread and butter. Therefore, if we want to find strong relationship at relatively low level in hierarchy, the minimum support threshold must be reduced substantially.

To remove this problem one should apply different minimum support to different concept levels. This leads to mining interesting association rules at multiple concept levels, which will find nontrivial, informative association rules because of its flexibilities for focusing the attention to different sets of data and applying different thresholds at different levels [9].

MLT2 Algorithm

Algorithm ML_T2L1: Find multiple-level large item sets for mining strong ML association rules in a transaction database.[6]

Input: (1) T [1], a hierarchy-information-encoded and task-relevant set of transaction database, in the format of (TID, Itemset), in which each item in the Itemset contains encoded concept hierarchy information, and (2) the minimum support threshold (minsup[l]) for each concept level 1.

Output: Multiple-level large item sets.

Method: A top-down, progressively deepening process, which collects large itemsets at different concept, levels as follows. Starting at level 1, derive for each level 1, the large K-items sets, L[l, k], for each k, and the large item set, L[l] (for all k's), as follows: -

(1) for (l:=1;$L[l,1]$ <> 0 and l < max_level; l++) do begin
(2) if l = 1 then begin
(3) $L[l,1]$:= *get_large_1_itemsets* ($T[1],l$);
(4) $T[2]$:= *get_filtered_transaction_table* ($T[1]$, $L[1,1]$);
(5) End
(6) else $L[l,1]$:= get_large_1_itemsets ($T[2]$, l);
(7) for (k := 2; $L[l, k-1]$ <> 0; k++) do begin
(8) C_k := *get_candidate_set* ($L[l, k-1]$);
(9) foreach transaction t in $T[2]$ do begin
(10) C_t := *get_subsets* (C_k, t); // Candidates container in t
(11) foreach candidate c in C_t do c.support++;
(12) End
(13) $L[l,$ k] := {c in C_k | c.support >= minsup[l]}
(14) End

(15) $L[l] := \mathrm{U}_k L[l, k]$;
(16) End

III. Experimental Results and Analysis

The dataset named MeSH® used in this research work from real world domain. This dataset is available from NLM's PubMed database [8]. The details of this dataset are described in table 1.

S.No	Dataset Name	Dataset Size	No. of Transactions	No. of Items
1.	MeSH-A	8MB	9,885	1800
2.	MeSH-C	10MB	10,000	830
3.	MeSH-D	12MB	10,000	1200

Table 1. Details of dataset used

Summary of Results using Min_Support and No. of Frequent Itemsets Generated Factor are given below:-
On the basis of Min_support and No. of frequent itemsets generated, we have drawn the following graphs for analyzing the results. The graphs for levels 1, 2 and 3 are shown in figure 1, figure 2 and figure 3 respectively.
Parameters: Max_Support 90 %, Delta 0.5

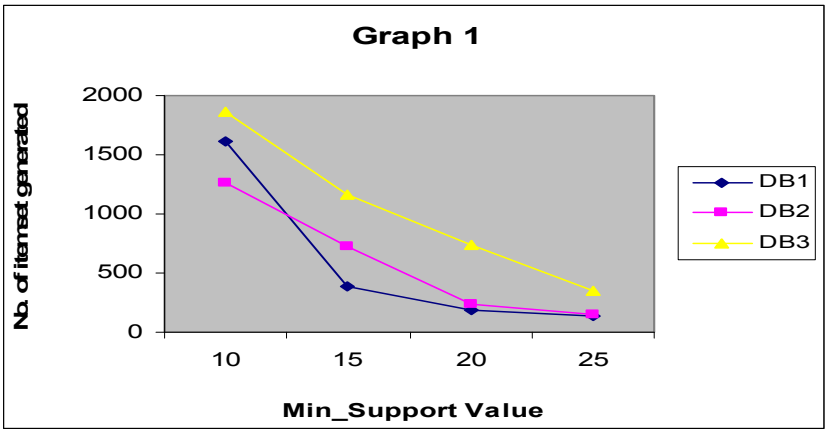

Figure 1. Min_Sup Vs itemsets generated (Level-1)

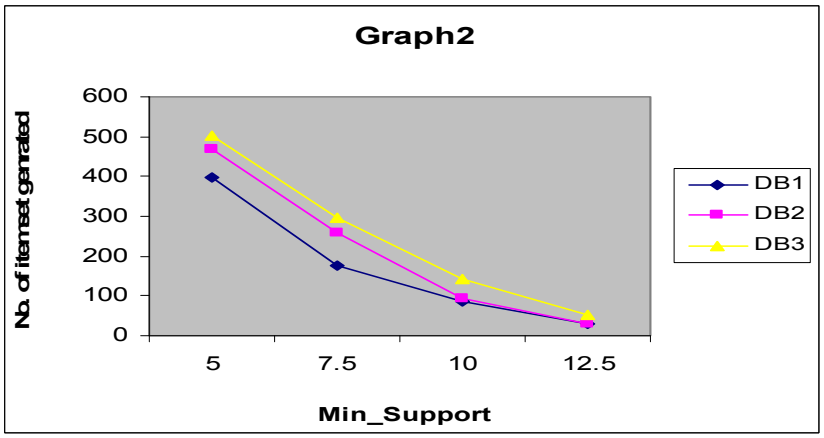

Figure 2. Min_Sup Vs itemsets generated (Level-2)

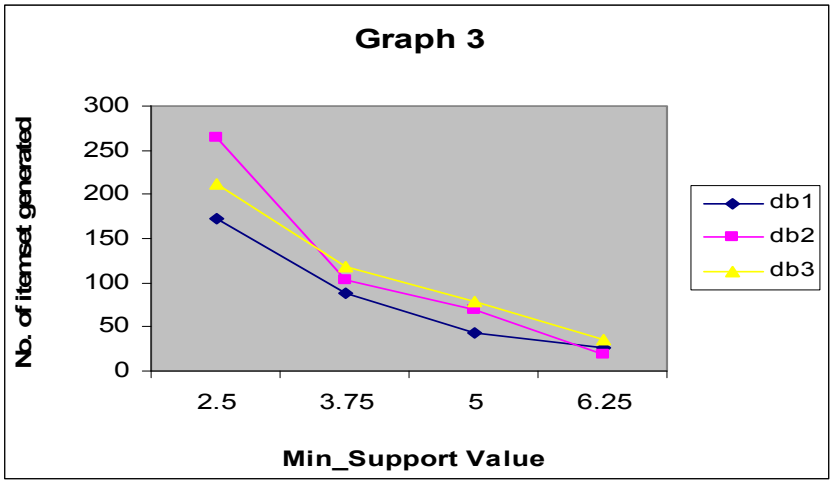

Figure 3. Min_Sup Vs itemsets generated (Level-3)

It is clear from above results that Min_Support is decreasing by factor delta(0.5) and as Min_support decreases at lower levels we find very specific information. The execution time of the algorithm is variable for different datasets with a variation in Min_Support. The time for different frequent item set mining algorithms depends a lot on the structure of the data set. The mining of multiple-level rules can provide more specific information for the users at lower levels and enhance the flexibility and power of data mining systems.

On the basis of Max_support and No. of frequent itemsets generated we have drawn graphs for analyzing the results. The graphs named graph 1, graph 2 and graph3 are shown in figure 4, figure 5 and figure 6 respectively.
Parameters : Min_Support 10 %, Delta 0.5

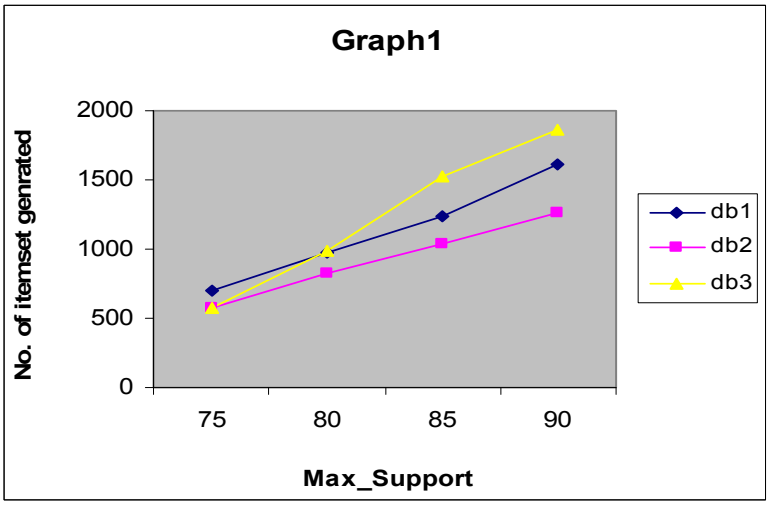

Figure 4. Max_Sup Vs itemsets generated (Level-1)

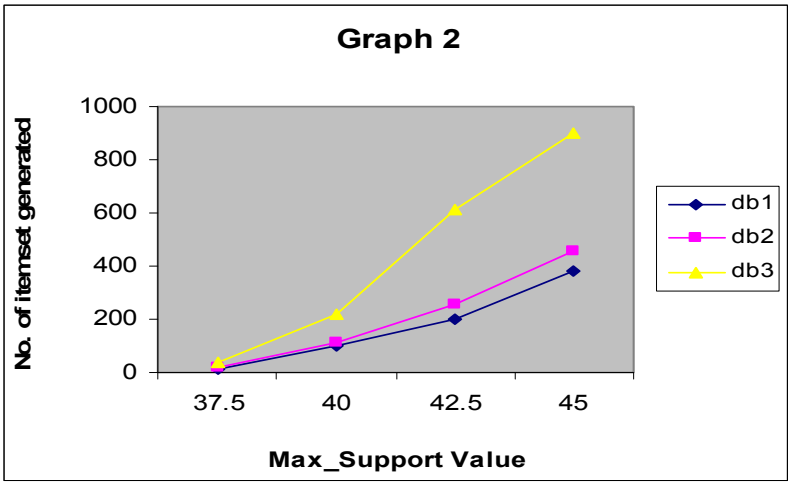

Figure 5. Max_Sup Vs itemsets generated (Level-2)

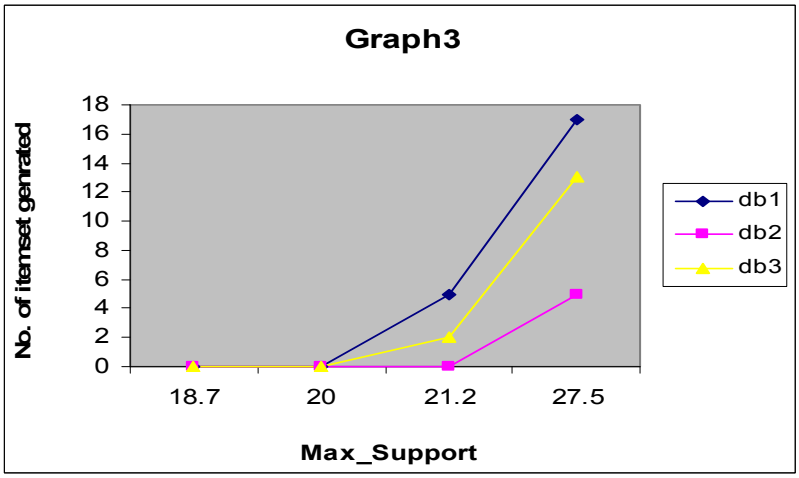

Figure 6. Max_Sup Vs itemsets generated (Level-3)

It is clear from above results that Max_Support is decreasing by factor delta (0.5) an no. of frequent itemsets is decreases as Max_support decreases at each level. The Multiple-level association rules can be discovered efficiently from large databases.

IV.CONCLUSION

This study demonstrates that mining multiple-level knowledge is both practical and desirable. This work has successfully discovered multiple-level association rules using MLT2 algorithm. The association rules discovered provide more specific information for the users at multiple levels of abstraction. Our algorithm has efficiently discovered Multiple-level association rules from three datasets (MeSH-A, MeSH-C, MeSH-D) from NLM's PubMed database. We have noticed that the execution time of the algorithm depends on the size and complexity of concept hierarchy discovered and hence it is variable for different datasets. This algorithm discovers association rules for successive levels making use of rules already discovered for upper levels of concept hierarchy. Number of association rules discovered depends on value of parameters at each level like support, confidence, and lift. This

work is contribution towards representing knowledge at multiple-levels in the form of association rules that enhances the comprehensibility of the results for the users.

V. REFERENCE

[1] Fu, Yongjlan, "Data Mining: Tasks, techniques and applications," IEEE. Potentials, 1997, pp.18-20.

[2] Choonho Kim and Juntae Kim, "A Recommendation Algorithm Using Multi-Level Association Rules," in Proc. WIC International Conference on Web Intelligence, Oct 2003, pp. 524-527.

[3] N.Rajkumar, M.R.Karthik and S.N.Sivanandam, "Fast Algorithm for Mining Multilevel Association Rules," 2003 IEEE, pp. 152-155.

[4] Jiawei Han and Micheline Kamber, "Data Mining: Concept and Tech-niques," 2nd Edition, 2001, pp.6-9.

[5] Janas and J.M., "A Priori Algorithm for Mining Multidimensional Association Rules," in Proc. 25th International Conference on Information Technology Interfaces IT1, June 2003, pp. 193-198.

[6] Jiawei Han and Yongjian Fu, "Discovery of Multiple-Level Association Rules from Large Databases," in Proc. 21st VLDB Conference, 1995.

[7] A. Siddiqui, D. Prohaska A. Ghosh S. Anamanamuri, "Implementation of Multiple-Level Association Rule Mining in Weka ," June 2005.

[8] http:/www.nlm.nih.gov/mesh

[9] Jiawei Han and Yongjian Fu., *"Discovery of Multiple-Level Association Rules from Large Databases"*. Proceeding in IEEE Trans. on Knowledge and Data Eng. Vol. 11 No. 5, pp 798-804, 1999.