Mining Rare Itemsets on Multi-Level Hierarchies Database

Taweechai Ouypornkochagorn

***Abstract*—Mining multi-level association rule is one of the most important fields in data mining to exploit hidden knowledge in large database which has extraordinary relations as hierarchies. Unfortunately, with enormous mining outcomes from multi-level hierarchies database show unexpected characteristic that many itemsets, especially leaf level itemsets, are neglected. In this paper, I propose SML algorithm to extract the rare itemsets in any levels of concept in relational database. Statistic data has been used to evaluate and the new proposed information, RareSupp, is used for representing the interestingness that applicable in real-world applications.**

***Index Terms*— Multi-level hierarchies, rare itemset.**

I. Introduction

Association rule mining is the one of the most important issue in hidden knowledge discovery introduced by [14] that it evaluates correlative relationship among a large set of data items. Through the extraction of knowledge, the uncovered knowledge can be applied for decision making process and developing strategies in various real-world applications [6], additionally, adapted for other mining techniques such as classification [1].

Nevertheless, many databases in several applications in real do not form their data structure in simple format. A lot of data items do not have only one level of abstraction, but they have multi-level hierarchies which results various patterns of derived rules or itemsets. These itemsets are formed between primitive level items and more conceptual level of abstraction items. This behavior is very attractive to users that are able to gain the advantages from shallow and deep information [4], [5] and many researches [2], [4], [7]-[12], [17] had been proposed to serve this purpose.

Outcomes of multi-level hierarchies mining are very interesting topic for discussion. Rule generating characteristic provides a lot of high support itemsets at conceptual level or parent node of hierarchy, but provides few numbers of itemsets at specific level or leaf node of hierarchy. Let see the example in Table 1 and Fig. 1 which show the small part of sample in census database from www.ics.uci.edu. The conceptual itemset {Farms Operators, US} (parent-to-parent node) has 50% support, but only 25% on specific itemset {Farmer, Ohio} (leaf-to-leaf node). Both are interesting in different view but unfortunately, with the large difference in support value, many times specific itemsets are dropped from user interesting by threshold. So, these interesting itemsets which were dropped inadvertently by evaluation processes will be rare itemset to found.

Taweechai Ouypornkochagorn is with Faculty of Engineering at Si Racha, Kasetsart University Si Racha Campus, Chonbury, Thailand (e-mail: sfengtwo@src.ku.ac.th)

Rare itemsets [19], [20] are the pattern in database that does not occur frequently. Those patterns may be unexpected important phenomena or perhaps are infrequent patterns which are very interesting. The general levelwise pattern mining via traditional Apriori algorithm is the simple way to evaluate but it has the several problems in practical. Apriori can be adjusted its threshold to cover almost all of rare itemsets but the very large portions of its result are not interesting, especially on multi-level hierarchy. The other ways to retrieve interesting itemsets, including with rare itemsets, had been proposed by several researches both Apriori style and new methods. Notice from those researches, Apriori tackles many problems on rare itemsets mining e.g. the large number of frequent itemsets, the evaluation time performance and so on.

In this paper, I propose the new approach for mining rare itemsets from multi-level hierarchies database, based on classical Apriori algorithm improvement. I apply the normal distribution curve to preserve rare itemsets, and propose additional information value for itemset's rareness and interestingness representation.

TABLE I
Sample of Census Database

(Omitted)	Sex	Occupation	Place of Birth	(Omitted)
…	M	Farmers	Ohio	…
…	F	Chemical	Thailand	…
…	M	Horticulturist	New York	…
…	M	Petrolium	Japan	…

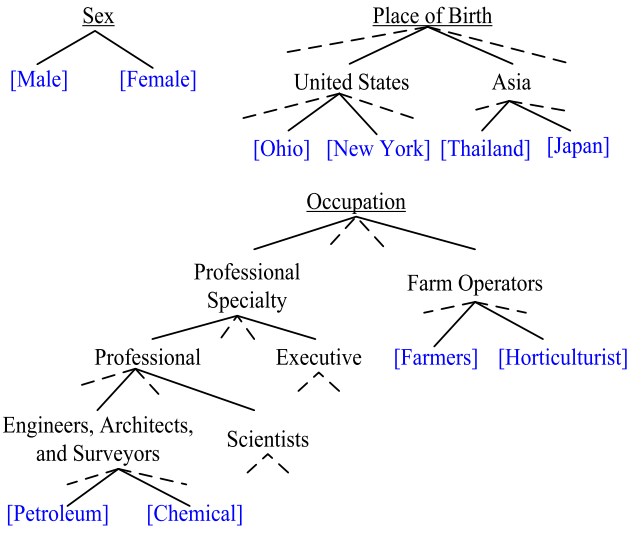

Fig. 1. Hierarchies of Census data

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. Problem Statement

Mining rare itemsets via Apriori algorithm is the efforts to find frequent rare patterns which seem useful. However, in real applications which have multiple hierarchies with different tree structure and associated attributes, confront with difficult task to search useful rare patterns. I summarized the problems as following:

A. Structural problems

Several information domains are composed on a database and formed as hierarchies. They are different pattern and tree's height that cause to hardly extract frequent itemsets both non-cross and cross level of abstraction itemsets.

B. Distribution of support value problems

According to Apriori algorithm, the extremely number of itemsets may be generated in primitive level, and it is more seriously when multiple conceptual levels have been included. The most popular tool to filter interesting itemsets is user minimum support; however it still has many concerns. In Table 2, I examine the extracted items which have 1 itemset size on 3 popular databases, Northwind (from www.microsoft.com), Adult and Census1990 (from www.ics.uci.edu). Obviously, I found that the 10% support values of each databases locate on very different percentile cause from dissimilar data characteristic. So, even if the minimum supports are the same, you cannot expect the number of results. It could be the strictly threshold as Adult database, or can be liberally threshold with a lot of itemsets return as Northwind and Census database. Or briefly say, at the same minimum support value, Adult database was resulted with the group of more interesting itemsets than averagely lower interesting itemsets of Northwind and Census database outcomes. This implies that you cannot use minimum support to clearly indicate the interestingness of itemsets. Moreover, from difference of statistic data, a little change of user minimum support may be trivial or crucial thing to interest on different databases. For example from Table 2, when changing from 10% to 11% minimum supports makes the increasing of percentile with 3.68%, 2.20% and 2.31% in Northwind, Adult and Census database respectively. This means, major effect is happened on Northwind database but little effect is occurred on Adult and Census database.

The distribution of support values among different size of itemsets is also the essential subject to be considered. Referring Table 1 and Fig. 1, focusing on Place of Birth and Occupation field, probability of {Horticulturist} and {Ohio} that can found on table are 1/4, when probability of {Horticulturist, Ohio} is the multiplication between probability of {Horticulturist} and {Ohio} or equal 1/16 (unfortunately, this itemset does not happen on Table 1). Similar with cross level itemsets: {Farm Operators, Ohio} and {Farm Operations, US} are 1/8 and 1/4 respectively. It means leaf-to-leaf 2-itemset has very low probability to occur; and definitely hardly to pass with user minimum support. In this point, you have to realize that the leaf-to-leaf itemsets which are neglected by user minimum support are not interesting, but they are neglected because they have high probability to be neglected. I called the interesting itemsets which are neglected causes from their probability are "rare itemsets" which are our objective to evaluate them in this paper.

C. Effectiveness of existing thresholds problems

Continue from two previous problems, I doubt with the effectiveness of the popular threshold, minimum support. Can it help to evaluate interesting itemsets or rare itemsets for all levels effectively? Does it indicate the interestingness or only show about the confident degree of itemsets?

III. Mining Rare Itemset on Multi-Level Hierarchies with SML-Apriori Algorithm and RareSupp Information

In this paper, I propose new way to solve 3 problems above by following:

A. Structural problems – Simplify it

Many researches [2], [4], [7]-[12], [17] proposed the solutions for this problems by complex coding hierarchy. In this paper, I use the simple solution by propagating the conceptual items in hierarchies to database and I tend them as same as primitive items. The example of propagation was shown in Table 3. This method will simplify the complexity of hierarchies' problems. After propagating, I can apply the basic principle of Apriori for finding association rules. Note that, in the step of Apriori, I add candidate generating constraint that 2 items which have same root ancestor do not have to generate next size of candidate itemset. For example, {Chemical} and {Engineers} will not need for generating candidate {Chemical, Engineers} because they have same root {Professional Specialty}.

B. Distribution of support value problems – Solve by normal distribution curve

I found that statistic data of support values should be considered in frequent itemsets evaluation. Therefore I propose the new threshold, a percentile value (α), for generating multiple minimum supports for every group of same itemset size. Each of minimum support values depends

TABLE II
Percentile of 10% Support on Various Databases

Database	#Record/ #Item	Average	SD	Percentile of 10%Sup
Northwind (Microsoft)	2,155 /202	85.35	154.31	80.05%
Adult (UCI)	32,561 /104	2,817.78	5,876.24	52.97%
Census 1990 (UCI)	2,458,285 /2,170	27,252.45	171,994.75	89.81%

TABLE III
Example of Census Database Propagation

	Sex	Occupation	Place of Birth	
…	M	[Farmers], Farm Operators	[Ohio], United States	…
…	F	[Chemical], Engineers, Professional, Professional Specialty.	[Thailand], Asia	…
…	M	[Petrolium], Engineers, Professional, Professional Specialty.	[New York], United States	…
…	M	[Horticulturist], Farm Operators	[Japan], Asia	…

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

on statistic data of support values of itemsets which have same itemset size. Percentile value (α) is the statistic measurement for data positioning among their members, generally based on normal distribution curve. So, I propose to change the way from user specific minimum support (the old way) to selecting the significant itemsets by user specific percentile based on their statistic data, for example, changing from 10% user minimum support to 80% percentile value for "**each groups of same size of itemsets**". From this point, all of significant itemsets in each size will be evaluated. That means rare itemsets will have more opportunity to be chosen and they can be used for next candidate generating. Note that, many of itemsets which have high support value will be dropped because that high value may seem too small when comparing with other itemsets on same itemset size, or it means they are no significance.

Definition 1: the minimum support of *k*-itemsets candidates: MS_k is defined as below where MSS_k is minimum support that calculates from user percentile value (α) and *LS* is user specific least support.

$$MS_k = \begin{cases} MSS_k & If\ MSS_k > LS; \\ LS & Otherwise; \end{cases} \quad (1)$$

Definition 2: the minimum support of *k*-itemsets: MSS_k is the generated support threshold which its value depends on statistic data of support values of all candidates in same itemset size, based on Normal Distribution Curve. MSS_k is the support value which located on percentile, calculated from (2), and that percentile have to equal with user specific percentile α (practically, Standard Curve Statistical Table is easier to use). *Z-score* getting from (2) will be used in (3) to find out MSS_k. Remark that μ is Mean and σ is Standard Deviation.

$$Percentile\ (z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\ dz \quad (2)$$

$$MSS = z \times \sigma \quad + \mu \quad (3)$$

Notice from Definition 1, least support (LS) is the one of two user specific values which refers to the lowest minimum support which should satisfy to become a frequent itemsets and prevents to create uninteresting concepts when standard deviation of its level nears zero. Least support is used together with user specific percentile (α) which illustrates in Definition 2. Percentile (α) is general statistic measurement that can reflect to significant position of data. In our opinion, percentile is clearly meaningful more than minimum support so much.

Integrating with the offering above, I propose Statistic Multi-Level Apriori (SML-Apriori) as shown in Fig. 2. First of all, database must be prepared by hierarchies propagating process, likes the example in Table 3. This database will be conducted together with the two user specific thresholds: percentile (α) and LS for our algorithm. Notice that, SML-Apriori modifies traditional Apriori by adding step in line 9-12. I can explain that statistic values will be calculated in each sizes of itemset, that shown in line 9-10. Then at line 11, MSS_k will be returned from function get_MSS(), related to Definition 2. Finally at line 12, the true minimum support gets from the largest value between MSS_k and LS. After all frequent itemsets are evaluated, all of them will be calculated for RareSupp information (it will be described next in definition 5), but I do not show in figure.

C. Effectiveness of existing thresholds problems – Offer new information values, rareness and RareSupp

I found that minimum support is still the suitable indicator to measure confidential degree of itemsets, but confidential degree does not indicate the interestingness degree or measuring rareness of itemset. Root level itemsets which have high support value such as {US, Farm Operators} with 50% on Table 1 probably seem less interesting than {Ohio, Farmers} with 25%. So, {Ohio, Farmers} is the rare itemset which has more opportunity to drop by user minimum support. In this paper, I separated thresholds of confidence from interestingness. In aspect of confidence, I have proposed in Definition 1 and 2, but in term of interestingness and rareness, it shown in Definition 3 -5.

Definition 3: the rareness of k-itemsets is probability of occurrence that is defined below, where *k* is size of itemset, *l* is itemset and *c* is item contained in *l*.

$SML-Apriori(D', \alpha, LS)$
(1) $L_1 = find\_frequent\_1-itemsets(D)$;
(2) $for(k = 2; L_{k-1} \neq \phi; k++)\{$
(3) $\quad C_k = Candidate\_Gen(L_{k-1})$;
(4) $\quad for\ each\ transaction\ t \in D\{$
(5) $\quad\quad C_t = subset(C_k, t)$;
(6) $\quad\quad for\ each\ candidate\ c \in C_t$
(7) $\quad\quad\quad c.count++$;
(8) $\quad\}$
(9) $\quad \sigma_k = get\_stddev(C_k | c.count > 0, c \in C_k))$;
(10) $\quad \mu_k = get\_mean(C_k | c.count > 0, c \in C_k))$;
(11) $\quad MSS_k = get\_MSS(\alpha, \sigma, \mu)$;
(12) $\quad MS_k = \begin{cases} MSS_k & If\ MSS_k > LS; \\ LS & Otherwise; \end{cases}$
(13) $\quad L_k = c \in C_k | c.count > MS_k$;
(14) $\quad return\ L = \cup_k L_k$;
(15) $\}$

Notation:
$D' = Database\ which\ propagated\ by\ its\ hierarchies$
$\alpha = User\ specific\ percentile$
$LS = Least\ support$
$L_k = Frequent\ itemset\ with\ k-items$
$C_k = Candidate\ itemset\ with\ k-items$
$MSS_k = Minimum\ support\ by\ significant\ of\ k-itemsets$
$MS_k = Minimum\ support\ of\ k-itemsets$

Fig. 2. Psuedo code of SML-Apriori

TABLE IV
EXAMPLE OF RARESUPP CALCULATION

Itemset	Support	Rareness	RareSupp
{US, Farm Operators}	2/4 = 50%	1-(1/2×1/2) = 75%	50%×75% =37.5%
{Ohio, Farmers}	1/4 =25%	1-(1/4×1/4) = 93.75%	25%×93.75% =23.44%

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$Rareness(l) = 1 - \prod_{i}^{k} Support(c_i) \quad (4)$$

Definition 4: the minimum rareness of k-itemsets is the user specific value that used to choose rare itemsets.

Definition 5: the RareSupp of k-itemsets is support weighting information that is defined as below:

$$RareSupp(l) = Support(l) \times Rareness(l) \quad (5)$$

From definition 3, I propose the way to measure rareness of itemset and propose the choosing way for rare itemset in definition 4, but I realize that support of itemset must be considered as well. So, RareSupp in definition 5 was proposed for that point as additional information to user.

RareSupp calculation example was shown in Table 4. Remark that, RareSupp is merely additional information for sorting itemsets by interestingness. It does not apply for itemsets filtering.

IV. Experimental Results

The experiments test on Census1990 database (U.S. Department of Commerce Bureau of Census) from (www.ics.uci.edu), by selecting 15 attributes which 5 hierarchies embedded, and random for 5% sample from original data or 122,914 records. This sample was propagated with abstract items contained in hierarchies, which consists of totally 1,885 items, 537 items are parent node in hierarchy trees. After running SML-Apriori with user percentile 60%-80%, I found that the outcomes contained with leaf-to-leaf, leaf-to-other parent, or parent-to-other parent items that I expected.

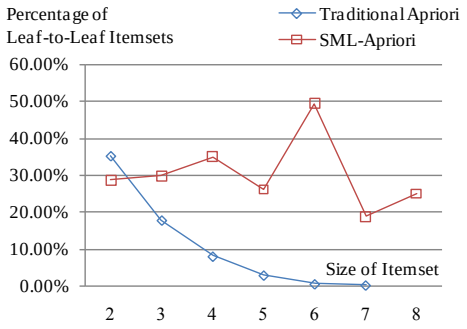

Fig. 3. Rare itemset exploration performance at each itemset sizes

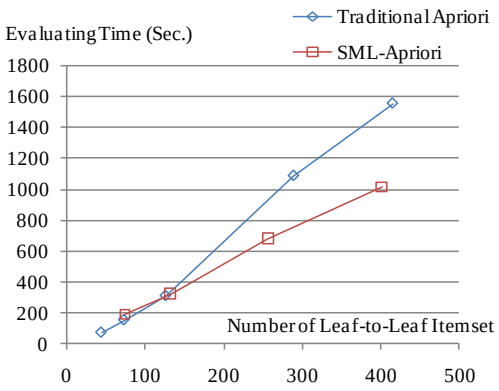

Fig. 4. Evaluating time performance at various number of outcomes

Next, I roughly examine the rare itemsets by looking for average percentages of leaf-to-leaf itemsets to all of frequent itemsets in each sizes of itemset, and then comparing with traditional Apriori with minimum support 30%-50%. The results were shown in Fig. 3. Apparently, SML-Apriori can explore rare itemsets more efficient, almost all of itemset sizes with 30.39% average whereas performance of traditional Apriori declines at large sizes with 10.65% average.

Evaluating time performances were shown in Fig. 4. Our experiments run on CPU Intel Core2 Duo 2.80GHz with 4G RAM. At the approximately same leaf-to-leaf itemsets, our algorithm, SML-Apriori, takes approximately equal evaluating time at the low number of leaf-to-leaf itemsets, but apparently faster about 26%-32% at high number of leaf-to-leaf itemsets.

Turning to interestingness measurement and rareness measurement, I make the experiment on knowledge representing between SML-Apriori with RareSupp information and support value on traditional Apriori that results in Table 5-6. The outcomes were sorted and only

TABLE V
Top 5 of Itemsets by SML-Appriori, with RareSupp

No	Itemset	Supp/ Rareness	RareSupp
1	Occupation: Group of Legislator Industry: Agricultural production Place of Birth: Group of Alabama Ability to Speak: Speaks Only Eng Year of Entry to US: Born in the US Place of Work: NA Means of Transportation to Work: NA	35.01%/ 96.39%	33.74%
2	Occupation: Legislator (Others same as 1 excepted Occupation)	35.01%/ 96.39%	33.74%
3	Industry: Group of Agricultural production (Others same as 1 excepted Industry)	35.01%/ 96.31%	33.71%
4	Occupation: Legislator Industry: Group of Agricultural production (Others same as 1 excepted Occupation and Industry)	35.01%/ 96.31%	33.71%
5	Industry: Group of Agricalture, Forestry, and Fisheries (Others same as 1 excepted Industry)	35.01%/ 96.21%	33.67%

Abstraction item is labeled prefix "Group of". We neglect the item Language Spoken at Home: "Not in Universe" from the results.

TABLE VI
Top 5 of Itemsets by Traditional Appriori

No	Itemset	Supp/ Rareness	RareSupp
1	Place of Birth: Group of Alabama Year of Entry to US: Born in the US	91.32%/ 16.60%	15.16%
2	Place of Birth: Group of Alabama Year of Entry to US: Born in the US Race: Group of American Indian	91.32%/ 16.60%	15.16%
3	Place of Birth: Group of Alabama Race: Group of American Indian	91.32%/ 8.68%	7.92%
4	Year of Entry to US: Born in the US Race: Group of American Indian	91.32%/ 8.68%	7.92%
5	Race: Group of American Indian Ability to Speak: Speaks Only Eng	87.62%/ 12.38%	10.84%

Abstraction item is labeled prefix "Group of". We neglect the item Language Spoken at Home: "Not in Universe" from the results.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

shown for top 5, most interesting to less. I found that even if the itemsets from traditional Apriori have very high support values, but they look like unuseful. In the opposite way of SML-Apriori with RareSupp information, RareSupp can help us to sort the interesting and rare itemsets more effective. In our some top of experimental results, from our approach, you can see the people who born in Alabama mostly are legislator in agricultural industry whereas you only get something that everyone known like the people born in Alabama are born in US, from traditional Apriori. So, our proposed information value will encourage user to realize it with real-world applications.

V. Conclusion

In this paper, I demonstrate the new algorithm, SML-Apriori, to mine multi-level rare itemsets with statistic technique. I also propose methods to serve multi-hierarchy which generally embedded in several attributes in real-world databases and raise the new information value, RareSupp, for measuring interestingness of itemsets. With our approach, I can explore non-cross and cross level rare itemsets in any sizes of itemset. Additionally, our chosen threshold, user percentile (α), is bearable with the changing to any databases and it is the one of good meaningful thresholds. The experiments show that SML-Apriori and RareSupp have very good outcomes and seem more realize in real-world applications. Anyway, applying to other style of Apriori is one of my challenges in my future works.

References

[1] B. Liu, W. Hsu and Y. Ma. "Integrating Classification and Assocoiation Rule Mining," In *Proc. KDD-98*, 1998, pp80-86.

[2] D. Xiangjun, Z. Zhiyun, N. Zhendong and J. Qiuting, "Mining Infrequent Itemsets Based on Multiple Level Minimum Supports," *Innovative Computing, Information and Control (ICICIC '07)*, 2007, pp.528.

[3] H. Xiong and P.-N. Vipin Kumar, "Mining strong affinity association patterns in data sets with skewed support distribution," *The Third IEEE International Conference on Data Mining (ICDM' 03)*, 2003, pp. 387-394.

[4] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," In *Proc. of the 21st International Conference on Very Large Databases (VLDB)*, Switzerland, 1995, pp. 420-431.

[5] J. Han and Y. Fu, "Mining Multiple-level Association Rules in Large Databases," *IEEE Transactions on Knowledge and Data Engineering*, 1999 vol.11.

[6] K. Younghee and K. Ungmo, "Mining Multilevel Association Rules on RFID Data," *The 1st Asian Conference on Intelligent Information and Database Systems (ACIIDS' 09)*, 2009, pp. 46-50.

[7] M. Pater, A. Bogan-Marta, C. GyĘrödi, R. GyĘrödi, "Multilevel frequent pattern mining from databases using AFOPT data structure," *The 7th International Conference on Technical Informatics CONTI'2006*, Romania, 2006, pp. 251-256.

[8] M. Pater and D.E. Popescu, "Market-basket problem solved with depth first multi-level apriori mining algorithm," *The 3rd International Workshop on Soft Computing Applications (SOFA '09)*, 2009, pp. 133-138.

[9] M. Pater and D.E. Popescu, "The Benefits of Using Prefix Tree Data Structure in Multi-Level Frequent Pattern Mining," *The 2nd International Workshop on Soft Computing Applications (SOFA '07)*, 2007, pp.179-182.

[10] M. Runying, "Adaptive-FP: An Efficient and Effective Method for Multi-Level and Multi-Dimensional Frequent Pattern Mining," Simon Fraser University, 2001.

[11] N. Rajkumar, M.R. Karthik and S.N. Sivanandam, "Fast Algorithm for Mining Multilevel Association Rules," *IEEE Web Technology and Data Mining*, TENCON, 2003, pp. 687-692.

[12] R. Agrawal, C. Agrawal and V.V.V. Prasad, "Depth first generation of large itemsets for association rules," *IBM Tech. Report*, RC21538, 1999.

[13] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," In *Proceeding of 20th International Conference on VLDB*, 1994, pp.487-499.

[14] R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1993, pp. 207–216.

[15] R. GyĘrödi, C. GyĘrödi, M. Pater, O. Boc and Z. David, "AFOPT Algorithm for multi-level databases," *The 7th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 05)*, ISBN 0-7695-2453-2, Timisoara, 2005, pp. 135-139.

[16] R. GyĘrödi, C. GyĘrödi, M. Pater, O. Boc and Z. David, "FP-Growth algorithm for multi-level databases," *The 15th International Conference on Control Systems and Computer Science (CSCS-15)*, Bucuresti, 2005, pp. 847-952.

[17] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," In *Research Report RJ 9963*, IBM Almaden Research Center, San Jose, California, USA, 1995.

[18] W.J. Frawley, G. Piateetsky-Shapiro and C.J. Matheus, "Knowledge discovery in databases: An overview," In *G. Piateetsky-Shapiro and W.J.Frawley, eds. Knowledge Discovery in Databases*, AAAI/MIT Press, 1991.

[19] M. Adda1, L. Wu and Y. Feng, "Rare Itemset Mining," The *6th International Conference on Machine Learning and Applications*, 2007.

[20] L. Szathmary, P. Valtchev and A. Napoli, "Finding Minimal Rare Itemsets and Rare Association Rules," In *Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (KSEM 2010)*, 2010, pp. 16-27.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)