



Privacy Preservation in Spatial Database

Ranu Sahu¹& Raghvendra Kumar²

^{1,2}Dept. of Computer Science & Engineering, LNCT Group of College, Jabalpur, MP, India
Ranu_sahu_7182@yahoo.com; raghvendraagrawal7@gmail.com

Abstract:

The majority of existing algorithms do not warrant the elimination of all well known geographic dependences. The result is that the same associations represented in geographic database schemas are extracted by spatial association rule mining algorithms and presented to the user. The problem of mining spatial association rules from geographic databases requires at least three main steps: compute spatial relationships, generate frequent patterns, and extract association rules. The first step is the most effort demanding and time consuming task in the rule mining process, but has received little attention in the literature. The second and third steps have been considered the main problem in transactional association rule mining and have been addressed as two different problems: frequent pattern mining and association rule mining. Well known geographic dependences which generate well known patterns may appear in the three main steps of the spatial association rule mining process. Aiming to eliminate well known dependences and generate more interesting patterns, this thesis presents a framework with three main methods for mining frequent geographic patterns using knowledge constraints. Semantic knowledge is used to avoid the generation of patterns that are previously known as non-interesting. The first method reduces the input problem, and all well known dependences that can be eliminated without losing information are removed in data preprocessing. The second method eliminates combinations of pairs of geographic objects with dependences, during the frequent set generation. A third method presents a new approach to generate the global relationship between the attributes without disclosing their private information to other reason in the geographic areas.

Keywords: Data Mining; Distributed Data Mining; Association Rule Mining; Spatial Data Mining; Spatial Association Rules; Weka Tool.

1. Introduction

Due to the increased demand for knowledge discovery in all industrial domains, it is necessary to store all the raw data and to provide useful patterns with respect to the user needs. Generally, the storage of all raw data will be done in a database maintained by respective organizations. Data mining techniques are available to retrieve useful information from large database. Prediction and description are the two fundamental goals of data mining. To fulfill these goals many data mining techniques exist such as association rules, classification, clustering and so on. Among these, association rule has wide applications to discover interesting relationship among attributes in large databases. Association rule mining is used to find the rules which satisfy the user specified minimum support and minimum confidence. In the process of finding association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets.

2. Spatial Data Mining

Spatial describes how objects fit together in space, either among the planets or down here on earth. Spatial data refers to all types of data objects or elements that are present in a geographical space or horizon. It enables the global finding and locating of individuals or devices anywhere in the world. Spatial data base is a database that is enhanced to store and access spatial data or data that defines a geometric space. These data are often associated with geographic locations and features, or constructed features like cities. Data on spatial databases are stored as coordinates, points, lines, polygons and topology. Some spatial databases handle more complex data like three-dimensional objects, topological coverage and linear networks. Spatial data mining is the application of data mining to spatial models. In spatial data mining, analysts use geographical or spatial information to produce business intelligence or other results. This requires specific techniques and resources to get the geographical data into relevant and useful formats.

3. Proposed Work

Due to the increased demand for knowledge discovery in all industrial domains, it is necessary to store all the raw spatial data and to provide useful patterns with respect to the user needs. Generally, the storage of all raw data will be done in a database maintained by concerned organizations. Data mining techniques are available to retrieve useful information from large spatial database. Prediction and description are the two fundamental goals of data mining. To fulfill these goals many spatial data mining techniques exist such as association rules, classification, clustering and so on. Among these, association rule has wide applications to discover interesting relationships among attributes in large spatial databases. Spatial association rule mining is used to find the rules which satisfy the user specified minimum support and minimum confidence. In the process of finding association rules, the set of frequent item sets are computed as the first step and then association rules are generated based on these frequent item sets.

In this process, we divided the entire region into the three different regions and each having their spatial database SDB₁, SDB₂, ..., SDB_n and their own key values SK₁, SK₂, ..., SK_n, or select N number of regions each having their own database SDB₁, SDB₂, ..., SDB_n. Each region calculates their frequent items set and support value.

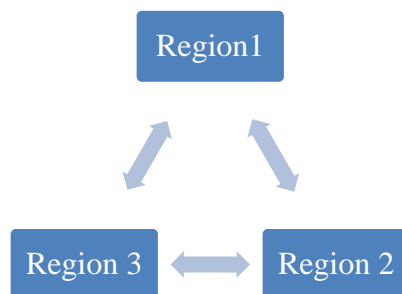


Figure 5.1: Communication among three regions and DM

Each region is arranged in a ring architecture then find the partial support. Now the region 1 sends their Partial Support (PS) value to region 2 and region 2 sends their value to region 3 and this process continues.

till region n and after that region n send their value to region 1. region 1 subtract all the Random number value from the Partial Support value and calculate their actual support, now region 1 broadcast the actual support value to the entire region present in the distributed environment.

○ **Proposed Algorithm:**

Encryption Process

Step1: Take the Spatial Database

Step2: Convert into the horizontally partitioned distributed database (N Number of datasets)

Step3: Calculate the support count of each database with the help of Weka Tool.

Step4: Calculate the support and confidence.

$$\text{Support} = \left(\frac{X_{UY}}{T(\text{total number of transaction})} \right)$$

$$\text{Confidence} = \text{Prob}\left(\frac{X_{UY}}{X}\right)$$

Step5: Calculate partial support and partial confidence.

$$\text{Partial Support (PS)} = X. \text{Support} - DB \times \text{Minimum Support}$$

$$\text{Partial Confidence (PC)} = X. \text{Confidence} - DB \times \text{Minimum Confidence}$$

Step7: Add their own private key in all partial support and partial confidence.

$$\text{Partial Support (PS)} = X. \text{support} - DB \times \text{minimum support} + \text{Key}$$

$$\text{Partial Confidence (PC)} = X. \text{Confidence} - DB \times \text{Minimum Confidence} + \text{Key}$$

Step8: Divided the partial support and partial confidence into the three different values.

Step9: Converted partial support, partial confidence and partial lift values into the ASCII value and compute the matrix Y.

Step10: Take the transpose of the matrix (Y^T).

Step11: Convert ASCII code matrix (Y^T) into the binary format.

Step12: Consider our own secret key(X matrix)

Step 13: Covert the X matrix into binary format

Step 14: Perform Exclusive-or between X and Y.

Step15: The resultant matrix is the encrypted format of plain text stored into the associative memory.

Step16: The resultant matrix is sanded to the protocol initiator Server.

Decryption Process

Step 1: Consider the resultant matrix M

Step 2: Compute transpose of M matrix as MT matrix

Step 3: Convert matrix MT into binary format

Step 4: Consider our own private key X

Step 5: Covert matrix X into binary number format

Step 6: Perform exclusive-or operation between MT and X

Step 7: The resultant matrix is converted to the ASCII code and finally we have the original text.

Step8: After receiving all the original values from the different database, the protocol initiator takes the step for data analysis by calculating Global support and confidence.

Step9: After that the protocol initiator broadcast the results to all the database server admin presents in the distributed environments.

Flow Chart:

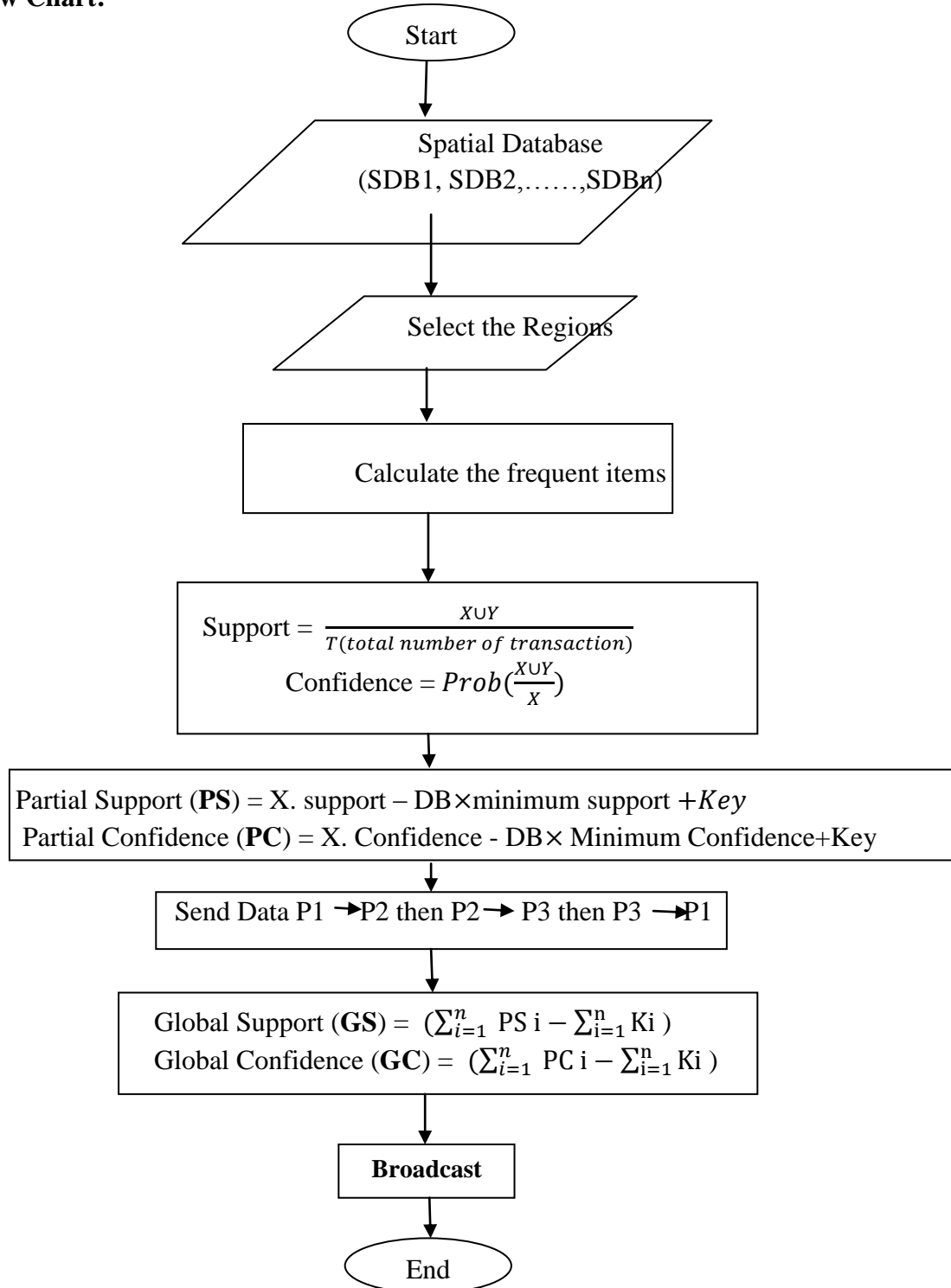


Figure: Flow Chart

Design & Implementation

In this thesis, we considered the spatial database, spatial database represented in table 1, table2 and table 3, which is partitioned into the homogeneous environment, having same number of attributes.

Table1: Spatial Database for Region 1

TID	REFRENCE CITY	SPATIAL OBJECTS						
		A	B	C	D	E	F	G
1	MUMBAI	1	1	1	1	1	0	1
2	CHENNAI	1	0	1	0	1	0	1
3	BENGALURU	1	1	1	1	0	1	0
4	HYDERABAD	1	0	0	0	1	0	0
5	KOLKATA	0	1	0	1	0	1	0
6	RAIPUR	0	0	1	0	0	0	1
7	CHANDIGARH	1	1	1	0	0	1	0
8	INDORE	1	0	1	0	1	0	1
9	COIMBATORE	0	1	0	1	0	1	0
10	PUNE	1	1	1	1	1	0	1

Table2: Spatial Database for Region 2

TID	REFRENCE CITY	SPATIAL OBJECTS						
		A	B	C	D	E	F	G
1	SURAT	1	0	0	1	0	1	1
2	AHMEDABAD	1	1	1	1	1	0	0
3	NASHIK	0	1	1	0	1	0	0
4	KOCHI	1	0	0	1	0	0	1
5	RAJKOT	0	1	1	0	1	0	0
6	GURGAON	1	1	0	1	0	1	1
7	JAIPUR	1	0	0	1	0	0	1
8	BHUBANESHWAR	1	1	1	1	1	0	1
9	VISAKHAPATNAM	1	0	0	1	0	0	1
10	TRIVENDRUM	1	1	1	0	0	0	0

Table3: Spatial Database for Region 3

TID	REFRENCE CITY	SPATIAL OBJECTS						
		A	B	C	D	E	F	G
1	SURAT	1	0	0	1	0	1	1
2	AHMEDABAD	1	1	1	1	1	0	0
3	NASHIK	0	1	1	0	1	0	0
4	KOCHI	1	0	0	1	0	0	1
5	RAJKOT	0	1	1	0	1	0	0
6	GURGAON	1	1	0	1	0	1	1
7	JAIPUR	1	0	0	1	0	0	1
8	BHUBANESHWAR	1	1	1	1	1	0	1
9	VISAKHAPATNAM	1	0	0	1	0	0	1
10	TRIVENDRUM	1	1	1	0	0	0	0

In this use the spatial database for region 1, it analyzed that database with the help of weka tool after applying the apriori algorithm, shows in below figure.

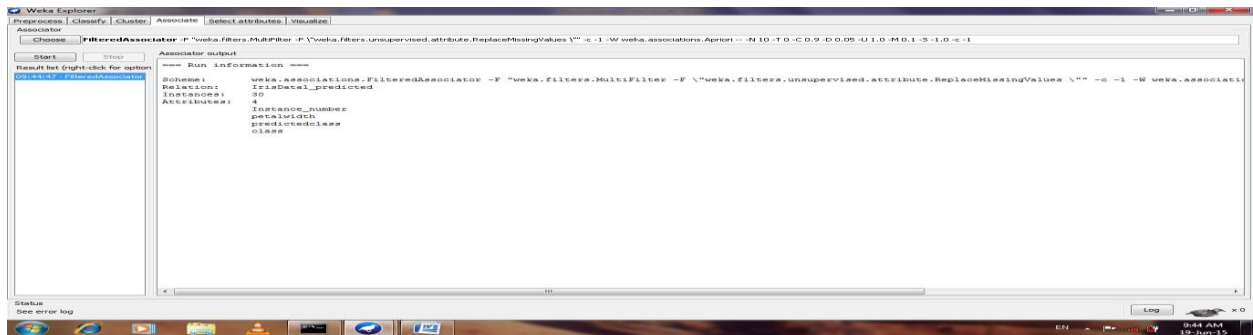


Figure: Shows analyzed dataset for the region 1

In this use the spatial database for region 2, it analyzed that database with the help of weka tool after applying the apriori algorithm, shows in below figure.

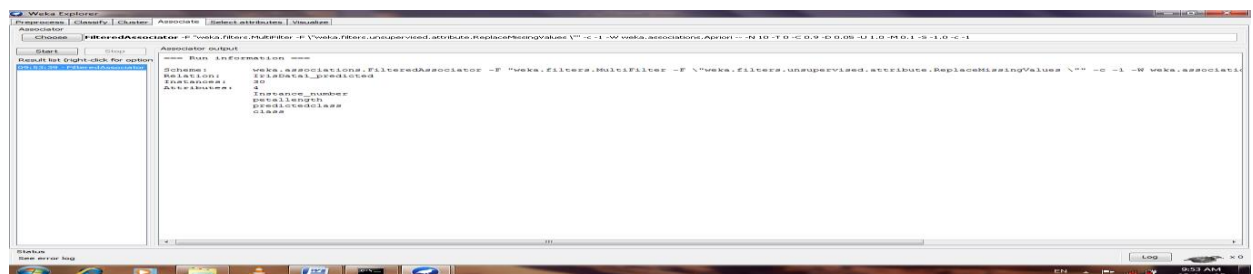


Figure : Shows analyzed dataset for the region 2

In this use the spatial database for region 3, it analyzed that database with the help of weka tool after applying the apriori algorithm, shows in below figure.

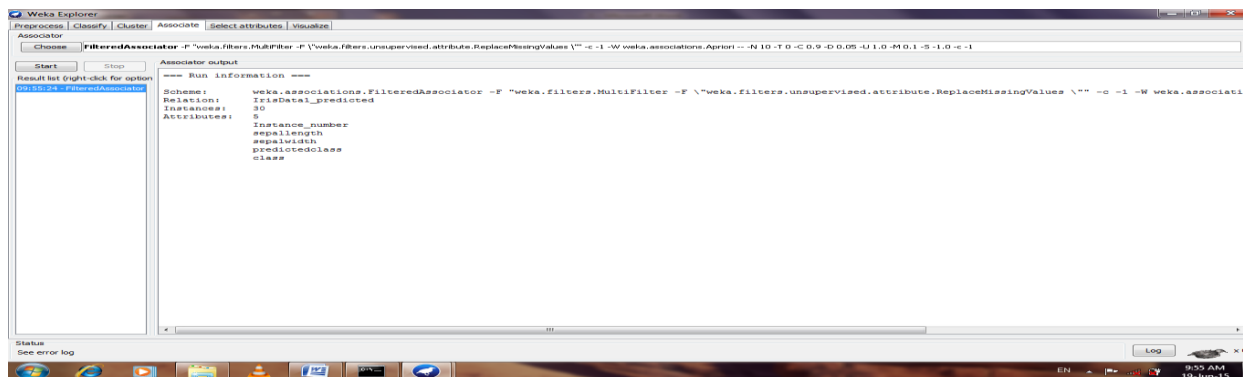


Figure: Shows analyzed dataset for the region 3

In this thesis, take the spatial database from three different regions, for analyzing the spatial database, consider the minimum support value is 40% and for providing the highest privacy to spatial database, it consider three different key numbers, whose value prospectively 1, 2 and 3. First each region calculated their support count by using the Apriori algorithm, then after that each region calculated their partial support by using the formula then after that added the key number and sanded to the next region head presented in the distributed environment, than after that the protocol initiator region will calculated their actual support, and broadcast it to all the region.

Region 1:

For the attribute AB, Key = 1, Support=0.40, Confidence=0.9

$$\text{Partial Support (PS)} = \text{Mod} (X. \text{ support} - \text{DB} \times \text{Minimum Support} + \text{Key}) = 0.4 - (10 \times 0.4) + 1 = 2.6$$

$$\text{Partial Confidence (PC)} = \text{Mod} (X. \text{ Confidence} - \text{DB} \times \text{Minimum Confidence} + \text{Key}) = 0.9 - 10 \times 0.6 + 1 = 4.1$$

Divided the partial support and confidence value into the three different values

$$PS_1=1, PS_2=1, PS_3=0.60$$

$$PC_1=2, PC_2=2, PC_3=0.1$$

Converted partial support and partial confidence values into the ASCII value and compute the matrix Y.

$$Y = \begin{bmatrix} SOH & SOH & NULL \end{bmatrix}$$

Take the transpose of the matrix (Y^T)

$$Y^T = \begin{bmatrix} SOH \\ SOH \\ NULL \end{bmatrix}$$

ASCII code matrix (Y^T) into the binary format

$$Y^T = \begin{bmatrix} 001 \\ 001 \\ 000 \end{bmatrix}$$

Consider our own secret key(X matrix)

$$X_1 = \begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix}$$

Covert the X matrix into binary format

$$X_1 = \begin{bmatrix} 010 \\ 100 \\ 001 \end{bmatrix}$$

Perform Exclusive-or between X and Y

$$Z_1 = X \text{ Ex-OR } Y^T = \begin{bmatrix} 001 \\ 001 \\ 000 \end{bmatrix} \text{ Ex - OR } \begin{bmatrix} 010 \\ 100 \\ 001 \end{bmatrix} = \begin{bmatrix} 011 \\ 101 \\ 001 \end{bmatrix}$$

Region 2:

For the attribute AB, Key = 1, Support=0.40, Confidence=0.8

Partial Support (PS) = Mod (X. support - DB× Minimum Support + Key) =0.4-(10×0.4) + 2=3.6

Partial Confidence (PC) = Mod (X. Confidence - DB×Minimum Confidence +Key) =0.8-10*0.6+2=3.2

Divided the partial support and confidence value into the three different values

$PS_1=1, PS_2=1, PS_3=1.6$

$PC_1=1, PC_2=1, PC_3=1.2$

Converted partial support and partial confidence values into the ASCII value and compute the matrix Y.

$Y = [SOH \ SOH \ SOH]$

Take the transpose of the matrix (Y^T)

$$Y^T = \begin{bmatrix} SOH \\ SOH \\ SOH \end{bmatrix}$$

ASCII code matrix (Y^T) into the binary format

$$Y^T = \begin{bmatrix} 001 \\ 001 \\ 001 \end{bmatrix}$$

Consider our own secret key(X matrix)

$$X_2 = \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix}$$

Convert the X matrix into binary format

$$X_2 = \begin{bmatrix} 011 \\ 100 \\ 001 \end{bmatrix}$$

Perform Exclusive-or between X and Y

$$Z_2 = X \text{ Ex-OR } Y^T = \begin{bmatrix} 001 \\ 001 \\ 001 \end{bmatrix} \text{ Ex - OR } \begin{bmatrix} 011 \\ 100 \\ 001 \end{bmatrix} = \begin{bmatrix} 010 \\ 101 \\ 000 \end{bmatrix}$$

Region 3:

For the attribute AB, Key = 1, Support=0.40, Confidence=0.8

Partial Support (PS) = Mod (X. support - DB× Minimum Support + Key)=0.4-(10×0.4) + 3=0.6

Partial Confidence (PC) = Mod (X. Confidence - DB×Minimum Confidence + Key) =0.8-10*0.6+3=4.1

Divided the partial support and confidence value into the three different values

PS₁=0.0, PS₂=0.0, PS₃=0.6

PC₁=2, PC₂=2, PC₃=0.1

Converted partial support and partial confidence values into the ASCII value and compute the matrix Y.

Y = [NULL NULL NULL]

Take the transpose of the matrix (Y^T)

$$Y^T = \begin{bmatrix} NULL \\ NULL \\ NULL \end{bmatrix}$$

ASCII code matrix (Y^T) into the binary format

$$Y^T = \begin{bmatrix} 000 \\ 000 \\ 000 \end{bmatrix}$$

Consider our own secret key(X matrix)

$$X_3 = \begin{bmatrix} 1 \\ 5 \\ 1 \end{bmatrix}$$

Covert the X matrix into binary format

$$X_3 = \begin{bmatrix} 001 \\ 101 \\ 001 \end{bmatrix}$$

Perform Exclusive-or between X and Y

$$Z_3 = X \text{ Ex-OR } Y^T = \begin{bmatrix} 001 \\ 101 \\ 001 \end{bmatrix} \text{ Ex - OR } \begin{bmatrix} 000 \\ 000 \\ 000 \end{bmatrix} = \begin{bmatrix} 001 \\ 101 \\ 001 \end{bmatrix}$$

After the encryption, now the entire region sends their encrypted value with the key value, to the protocol initiator region, then the protocol initiator region decrypted that value by using some of the decryption steps that shown in the above algorithm. Perform the Ex-OR operation between the resulting matrix Z_1 Ex-OR X_1 , Z_2 Ex-OR X_2 and Z_3 Ex-OR X_3 . Then we have the matrix M_1 , M_2 and M_3 , after that for calculating the resulting matrix M, perform the Ex-OR operation between M_1 , M_2 and M_3 .

$$M = M_1 \text{ Ex-OR } M_2 \text{ Ex-OR } M_3 = \begin{bmatrix} 000 \\ 101 \\ 000 \end{bmatrix}$$

Then take the transpose of the resulting matrix

$$M^T = [000 \ 101 \ 000] = 0.5$$

After taking the transpose converted into the ASCII value and then, the value of the resulting matrix M is same as the global support value, If the global support value greater than zero then it means that the, attribute value, that has been taken is globally frequent attribute, it may be locally infrequent attribute. So here, this thesis the calculated value of global support is greater than zero and it accepted as globally frequent item.

Conclusion

We have seen that some of the discovered rules actually convey new knowledge, however the search for these “nuggets” requires a lot of tuning and efforts by the data analyst in order to constrain the search space properly and discard most of the obvious or totally useless patterns

hidden in the data. This is typical of exploratory data analysis, and SPADA can be considered one of the most advanced tools that data analysts currently use in their iterative knowledge discovery process.



References

- [1] Boulicaut, J.F.; Jeudy, B. "Mining Free Itemsets Under Constraints". In: International Database Engineering & Applications Symposium, Ideas, 2001, Washington Dc. Proceeding: Ieee Computer Society, 2001. P. 322-329.
- [2] Brin, S.; Motwani, R.; Ullman, J.; Tsur, S. "Dynamic Itemset Counting And Implication Rules For Market Basket Data". In: Acm Sigmod International Conference On Management Of Data, Sigmod, 1997, Tuscon. Proceedings: Acm Press, 1997. P. 255-264.
- [3] Burdick, D.; Calimlim, M.; Gehrke, J. Mafia: "A Maximal Frequent Itemset Algorithm For Transactional Databases". In: Ieee International Conference Data Engineering, Icdede, 17., 2001, Heidelberg. Proceeding: Ieee Computer Society, 2001. P. 443-452.
- [4] Chaves, M. S.; Silva, M. J.; Martins, B. "A Geographic Knowledge Base For Semanticweb Applications". In: Brazilian Symposium On Databases, Sbbd, 20., 2005, Uberlândia. Proceedings: Ufu, 2005a. P.40-54.
- [5] Chaves, M. S.; Silva, M. J.; Martins, B. "Gkb - Geographic Knowledge Base". Lisboa: Di/Fcul, 2005.
- [6] Chifosky, E.J.; Cross, J.H. "Reverse Engineering and Design Recovery": A Taxonomy. Ieee Software, V.7, P.13-17, Jan. 1990.
- [7] Clementini, E.; Di Felice, P.; Van Oostern, P. "A Small Set Of Formal Topological Relationships For End -User Interaction". In: Abel, D; Ooi, B.C. (Ed.). Advances In Spatial Databases: Springer-Verlag, 1993. P. 277-295.
- [8] Clementini, E.; Di Felice, P.; Koperski, K. "Mining Multiple-Level Spatial Association Rules For Objects With A Broad Boundary". Data & Knowledge Engineering, V.34, N.3, P.251-270, Sept. 2000.
- [9] Cockcroft, S. A "Taxonomy Of Spatial Data Integrity Constraints". Geo informatica, V.1, N.4, P.327-343, Dec. 1997.
- [10] Ester, M. Et Al. "Spatial Data Mining: Database Primitives, Algorithms And Efficient Dbms Support". Journal Of Data Mining And Knowledge Discovery, V.4, N.2-3, P.193-216, July 2000.
- [11] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P." From Data Mining To Knowledge Discovery in Databases". Ai Magazine, V.17, N.3. P.37-54, 1996.
- [12] Fukuda, T. Et Al. "Mining Optimized Association Rules For Numeric Attributes" In: Acm Sigmod Symposium On Principles Of Database Systems, Pods, 15., 1996, Montreal. Proceeding: Acm Press, 1996. P.182-191.
- [13] Goethals, B.; E Zaki, M. J. "Advances In Frequent Itemset Mining Implementations": In: Ieee Icdm Workshop On Frequent Itemset Mining Implementations, 2003, Melbourne. Proceeding: Ceur-Ws.Org, 2003.