Association for Information Systems
AIS Electronic Library (AISeL)

ICEB 2003 Proceedings | International Conference on Electronic Business (ICEB)

Winter 12-9-2003

An Adaptive Two-Phase Spatial Association Rules for RSI Data Mining

Chin-Feng Lee

Mei-Hsiu Chen

Follow this and additional works at: https://aisel.aisnet.org/iceb2003

This material is brought to you by the International Conference on Electronic Business (ICEB) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICEB 2003 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

An Adaptive Two-Phase Spatial Association Rules for RSI Data Mining

Chin-Feng Lee, Mei-Hsiu Chen

Email : {lcf, s9014624} @cyut.edu.tw
Department of Information Management
Chaoyang University of Technology
168, Gifeng E.Rd., Wufeng, Taichung County, Taiwan 413, ROC

Abstract

Spatial association rule mining is a kind of spatial data mining to carry some interesting and implicit knowledge about spatial associations from spatial databases. Moreover, many excellent studies on Remote Sensed Image (RSI) have been conducted for potential relationships of crop yield. Lee and Chen [13] proposed a two-phase algorithm by creating Histogram Generators for fast generating coarse-grained spatial association rules, and further mining the fine-grained spatial association rules w.r.t the coarse-grained frequently patterns obtained in the first phase. However during the image processing, partitioning the image into parts can improve its efficiency; to this point, the concept of image blocking is incorporated onto the coarse-grained two-phase data mining of spatial association rules. Therefore, an adaptive two-phase spatial association rules mining method is proposed in this paper to improve two-phase method in terms of efficiency. The proposed adaptive method conducts the idea of partition on an image for efficiently quantizing out non-frequent patterns and thus facilitate two-phase process. Such adaptive two-phase approache saves much computations and will be shown by lots of experimental results in the paper.

Keywords: Spatial Database, Data Mining, Spatial Association Rules, Remote Sensed Image

1. Introduction

Those spatial data, from satellite photos or automated data gathering tool, such as remotely sensed images, local population, and what not have many widely-used sides of application, in many of today' research fields. When storing large amount of either spatial or non-spatial data in the spatial database, the past data gathering tool always emphasizes on how to improve on searching spatial data as well as storage system. However, there is very likely some knowledge or expertise potentially resided in this large amount of acquired data that is not discovered. In this case, it's very important at this moment to find an effective method to acquire this undiscovered knowledge or expertise from such large amount of spatial data and to further analyze for it to be applied on decisions of spatial applications. With the help of spatial data mining, more interesting and useful information are discovered [2][3][4][5][6][7][8] [9][11][12] [13][14][15]. This has, in turn, brought up a new,

very popular field for research today.

The methods in [3][4][5][13][15] for data mining on Remotely Sensed Images (RSI). In [3], the method is to take advantage of the association between reflectance intensities on RSI and crop yield through association rule. But if applying data mining simply on discovering this knowledge from large amount of remotely sensed images, it does not sound very economical considering the time spent and efforts put in. So, designing an effective data mining method is very worthwhile for in-depth research.

More details on methods will be discussed in Section 2. Section 3 presents a method for adaptive two-phase spatial association rules mining. Section 4 verifies the validity of our proposed method through experimentation and discussion. Last section is our conclusion.

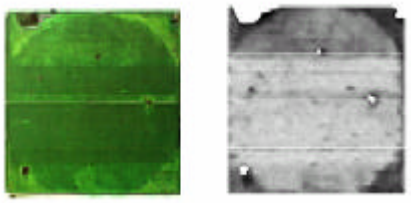
2. Literature Reviews

2.1 Association Rules

An association rule [1] is "$X \rightarrow Y, (s, c)$" where X and Y are itemsets with X being the antecedent and Y being the consequent of association rules. In example of the association between population and crop yield, s denotes the degree of support in which it represents the probability that X and Y occurs simultaneously, and c denotes the degree of confidence in which it represents the probability of Y occurrence when X occurs. Therefore, $c = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$. In some reference, support count $= s \times |D|$, where $|D|$ is the count in data table. Throughout our paper, the support count ($count$) will be taken as the count of frequent itemsets. Decision makers can pre-define a minimum support count and minimum confidence in order for the elimination of infrequent itemsets and the creation of association rules through frequent itemsets.

2.2 Spatial Data Mining on RSI

Images on RSI can be parted into a couple of different types such as TM, SPOT, AVHRR, and TIFF. Take TM image on RSI as example, TM consists of seven bands which are B for Blue, G for Green, R for Red, RIR for Reflective-Infrared, MIR for Mid-Infrared, TIR for Thermal-Infrared, and MIR2 for Mid-Infrared2 [3][4][5][13][15]. Each band contains a relative reflectance intensity value in the range 0 to 255 for each pixel. One remotely sensed image is associated with yield, in the way which that one yield image is a map that defines agricultural yield standards. It is often shown in color or gray-scaled image. Shown in Figure 1 (a) is a remotely sensed image and Figure 1 (b) is a yield image. Incorporating the association between remotely sensed image and yield, data mining is to excavate the useful expertise that may help the experts or agribusiness people to improve crop cultivation.



(a) (b)

Figure 1: Remotely sensed image and Yield map

2.3 Two-Phase Spatial Association Rules Mining Method

Two-phase association rule architecture for RSI data mining is proposed by Lee and Chen [13]. First, an image from RSI database is acquired. Do a color count and analysis of variance on this image. According to the analysis of image from RSI database, the Histogram Generator (HG) from the user-predefined intervals is generated. HG looks for the most representative characteristic value in an image to quickly find the coarse-grained association rules. From these coarse-grained association rules, fine-grained association rules are found. This method is primarily divided into four steps. First step is color count and analysis of variance on remotely sensed images. Second step is to generate Histogram Generator according to user-predefined intervals. Third step is to mine out the coarse-grained association rules according the user generated HG through algorithm of association rules. Fourth step is to mine out the fine-grained association rules according to the coarse-grained association rules.

3. Adaptive Two-Phase Spatial Association Rules Mining Method

The two-phase data mining of spatial association rules in Section 2.2 can improve the Apriori method in terms of its efficiency. However during the image processing, partitioning the image into parts can improve its efficiency; to this point, the concept of image blocking is incorporated onto the coarse-grained two-phase data mining of spatial association rules. Image blocking performs the mining on each of the disjoining blocks partitioned from an image. The motive of taking on the image blocking is due to the inter-pixel redundancy in image data [16]. That means, the occurrences of neighboring image points redundancy is quite high; in other words, there is a higher possibility of more frequent itemsets. Therefore, the adaptive two-phase data mining of spatial association rules is to eliminate the blocks which do not produce frequent itemsets to improve its efficiency through image blocking.

3.1. Flowchart of the Adaptive Two-phase Data Mining of Spatial Association Rules

Figure 2 is the flowchart of adaptive two-phase spatial association rules for RSI data mining. Exploration of the frequent itemsets in coarse-grained association rules is made in four main steps. First step is to partition an image. Second step is to produce local frequent itemsets and record local non-frequent itemsets at the same time. Third step is to group each local frequent itemsets into global candidate itemsets. Fourth step is to add together the support count in global candidate itemsets as well as in non-frequent itemsets to be the final support count in global candidate itemsets. When bigger than the minimum support count, a global frequent itemsets is produced. Next sub-section will put more emphasis on each of the four steps.

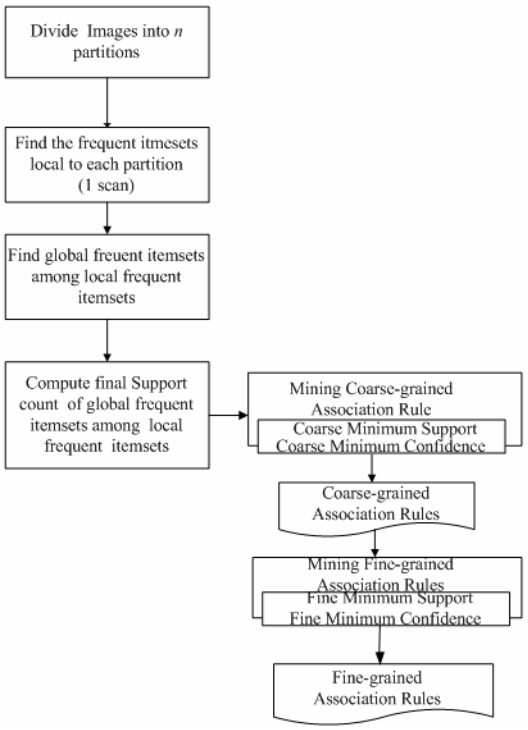


Figure 2. Adaptive two-phase association rule flowchart for RSI data mining

3.2 Procedures of the adaptive two-phase data mining of spatial association rules

Step 1: Partition the image

According to the mined coarse-grained association rules by Histogram Generator proposed by Lee and Chen [5], T^{HG} is the data table through HGs quantization of T^{origin}, which incorporates the RSI and yield images into one data table based on their positions of coordinates. Perform data mining on T^{HG} to explore the coarse-grained association rules. First is to partition into n number of disjoining blocks { $P^1, P^2, \ldots, P^n$ }. Divide the minimum support count by the count in data table T^{HG}. Multiply what's left in the previous calculation by the count in P^i to obtain the minimum support count in each block ($s_p^c = \frac{s^c}{|T^{HG}|} \times |P^i|$), where s_p^c is the minimum support count in the block, $|T^{HG}|$ is the count of data table T^{HG}, s^c is the minimum support count and $|P^i|$ is the count in the block.

Example 3.1: Table 1 is an image of 8×2 pixels, after which the data table T^{HG} is made from quantization by four HGs on bands R, G, B and Yield, respectively. The image is partitioned into four small blocks (n=4) in which block P^1 consists of {1, 2, 3, 4}, block P^2 of {5, 6, 7, 8}, block P^3 of {9, 10, 11, 12}, and block P^4 of {13, 14, 15, 16}. Whereas minimum support count (s^c) is 10, the minimum support count in each block is $s_p^c = \frac{s^c}{|T^{HG}|} \times P^1 = \frac{s^c}{|T^{HG}|} \times P^2 = \frac{s^c}{|T^{HG}|} \times P^3 = \frac{s^c}{|T^{HG}|} \times P^4 = \frac{10}{16} \times 4 = 2.5$.

Table 1. Four partitioned blocks 8*2 image data table

Partition	Id	coordinate	R	G	B	Yield
P^1	1	0,0	3	3	3	3
	2	0,1	0	0	0	3
	3	1,0	3	3	3	3
	4	1,1	3	3	3	3
P^2	5	2,0	3	3	3	3
	6	2,1	0	0	0	3
	7	3,0	0	0	0	3
	8	3,1	1	1	1	3
P^3	9	4,0	3	3	3	3

	10	4,1	3	3	3	3
	11	5,0	3	3	3	3
	12	5,1	3	3	3	3
P^4	13	6,0	3	3	3	3
	14	6,1	3	3	3	3
	15	7,0	3	3	3	3
	16	7,1	3	3	3	3

Step 2: Exploration of local frequent itemsets in each block

Step2.1: Calculate local frequent 1-itemsets (F_1^i) for each P^i. And each item in local frequent 1-itemsets is represented as *I*($band_j^{S_a}$, $Yield^{S_a'}$, *count)* because the itemsets in F_1^i can be represented as an association rule that consists of a antecedent and a consequent. For the association rule by the itemsets *I*, its antecedent is represented as *I*.$band_j^{S_a}$ and consequent as *I*.$Yield^{S_a'}$. *I.count* is the support count of the itemset. For example, *I*(R^0,$Yield^1$, 5) represents the color value of R as 0, the color value of Yield as 1 and the support count equals to 5.

On the other hand, record all the local non-frequent 1-itemsets that are smaller than the minimum support count in the block onto NF_1^i. Similarly, the format of local non-frequent 1-itemsets is *I*($band_j^{S_a}$, $Yield^{S_a'}$, *count)*, where *I*. $band_j^{S_a}$ is represented as the antecedent, *I*.$Yield^{S_a'}$ as the consequent and *I.count* as the support count of the itemsets.

Step2.2: Assume $\theta_1 \in F_{\ell-1}^i$ and $\theta_2 \in F_{\ell-1}^i$ where $q_1 \neq q_2$. To produce local frequent ℓ-itemsets (F_ℓ^i), first step has to consider whether or not $\theta_1.Yield^{S_a'}$ equals to $\theta_2.Yield^{S_a'}$. And if they equal to each other, they are joined together to produce local candidate ℓ-itemsets (C_ℓ^i). In local candidate ℓ-itemsets, each format of the itemsets is *I*($band_1^{S_{a1}}$, $band_2^{S_{a2}}$, …, $band_{\ell-2}^{S_{a\ell-2}}$, $band_{\ell-1}^{S_b}$, $band_\ell^{S_g}$, $Yield^{S_a'}$, *count*), of which *I*($band_1^{S_{a1}}$, $band_2^{S_{a2}}$, …, $band_{\ell-2}^{S_{a\ell-2}}$, $band_{\ell-1}^{S_b}$, $band_\ell^{S_g}$, $Yield^{S_a'}$, *count*) is the antecedent, *I*. $Yield^{S_a'}$ is the consequent and *count* is the support count. The support count. is defined as min(q_1.*count*, q_2.*count*) of these established association rules in the local candidate itemsets. When the support count is larger than the minimum support count in the block, local frequent ℓ-itemsets (F_ℓ^i) will be produced. local frequent ℓ-itemsets are formatted as *I*($band_1^{S_a}$, $band_2^{S_a}$, …, $band_\ell^{S_a}$, $Yield^{S_a'}$, *count*). Similarly, *I*. $band_1^{S_a}$, *I*. $band_2^{S_a}$, …, *I*. $band_\ell^{S_a}$ is represented as the antecedent, *I*.$Yield^{S_a'}$ as the consequent and *I.count* as the support count of the itemsets. For example, q_1=(R^0, G^0, $Yield^1$,4) $\in F_2^i$, q_2=(R^0, B^0, $Yield^1$, 3) $\in F_2^i$. When $q_1.Yield^1 = q_2.Yield^1$, *I*(R^0,G^0,B^0,$Yield^1$, 3) will be produced and support count=min(4, 3); again when the support count is larger than the

minimum support count in the block, the global frequent itemsets of $I(R^0, G^0, B^0, Yield^1, 3)$ will be produced. The color value of R is 0, of G is 0, of B is 0, of Yield is 1, and the support count is equal to 3.

On the other hand, record all the local non-frequent ℓ-itemsets that are smaller than the minimum support count in the block onto NF_1^i. Similarly, the format of local non-frequent ℓ-itemsets is $I(band_1^{S_a}, band_2^{S_a}, \ldots, band_\ell^{S_a}, Yield^{S_a'}, count)$, where $I.(band_1^{S_a}, band_2^{S_a}, \ldots, band_\ell^{S_a})$ is represented as the antecedent, $I.Yield^{S_a'}$ as the consequent and *I.count* as the support count of the itemsets.

Example3.2: Take Table 1 for example, the block P^3 in which local frequent 1-itemsets are produced, (R^3, $Yield^3$, 4), (G^3, $Yield^3$, 4) and (B^3, $Yield^3$, 4) are also produced. Furthermore, the local frequent 2- itemsets are (R^3, G^3, $Yield^3$, 4) where count=4 obtained from min(4, 4).

On the other hand, no local frequent itemsets are produced in the block P^2 because the support count of the itemsets is smaller than the minimum support count in the block. Our method records the itemsets onto the local non-frequent itemsets so as to produce the local non-frequent itemsets such as (R^3, $Yield^3$, 1), (G^3, $Yield^3$, 1), and (R^3, G^3, $Yield^3$, 1).

Step 3: Composition of a global candidate itemsets from each local frequent itemsets

First step is to incorporate the local frequent ℓ-itemsets $F_\ell^1, F_\ell^2, \ldots, F_\ell^n$ in each block into a global candidate ℓ-itemsets (C_ℓ). Second is to calculate the support count in C_ℓ; that is, $C_\ell.count = F_\ell^1.count + F_\ell^2.count + \ldots + F_\ell^n.count$.

Example 3.3: In Table 1, the local frequent 3-itemsets in P^1 are (R^3, G^3, B^3, $Yield^3$, 3), and the local frequent 3-itemsets in P^3 and P^4 are (R^3, G^3, B^3, $Yield^3$, 4). Incorporation of both comes the global candidate 3-itemsets which are (R^3, G^3, B^3, $Yield^3$, 11), where count=(3+4+4)=11.

Step 4: Calculation of the final support counts in global candidate itemsets for establishing the global frequent itemsets

Adding together the established global candidate itemsets $C_1, C_2, \ldots, C_\ell$ in Step 3 and the support count in each local non-frequent itemsets (NF_ℓ^i) yields $C_\ell.count = C_\ell.count + NF_\ell^i.count$, and the final support count in the global candidate itemsets can be produced. If the support count is larger than the minimum support count (s^c), then they are the global frequent itemsets.

Example 3.4: Again take Table 1 as an example, the global candidate itemsets (R^3, G^3, B^3, $Yield^3$, 11) are produced in Step 3. There are local non-frequent itemsets such as (R^3, G^3, B^3, $Yield^3$, 1) in the P^2. Thus the final global candidate itemsets that are produced are (R^3, G^3, B^3, $Yield^3$,

12), where count=(11+1)=12 that is larger than the minimum support count 10. So the global frequent itemsets (R^3, G^3, B^3, $Yield^3$, 12) are produced.

4. Experimental results of adaptive two-phase data mining of spatial association rules

The experiments conduct on the time-related ratio between the adaptive two-phase data mining, the two-phase data mining of spatial association rules and the Apriori.

Figure 3 shows the time-related ratio between the adaptive two-phase data mining which incorporates the Histogram Generator count that is 128 as well as blocking and the Apriori on five 50,000 pixels images under different color counts. Figure 3 also shows the time ratio between the two-phase data mining which incorporates the Histogram Generator count that is 128 but excludes the blocking and the Apriori. When the time ratio is less than 1, it implies that both the two-phase data mining and the adaptive two-phase data mining are effective. From the Figure 3, the adaptive two-phase data mining yields the time ratio that is significantly less than the two-phase data mining; in other words, the adaptive two-phase data mining can improve the two-phase data mining of spatial association rules on its effectiveness.

Figure 4 is the time-related ratio between the adaptive two-phase data mining which incorporates the Histogram Generator count that is 128 as well as blocking and the Apriori. It also shows the time ratio between the two-phase data mining which incorporates the Histogram Generator count that is 128 but excludes the blocking and the Apriori. When the time ratio is less than 1, it implies that the adaptive two-phase data mining is very effective. From the Figure 4, the adaptive two-phase data mining requires significantly less amount of time; in other words, image blocking in the first and coarse-grained association rules in the later are able to enhance the two-phase data mining in its effectiveness

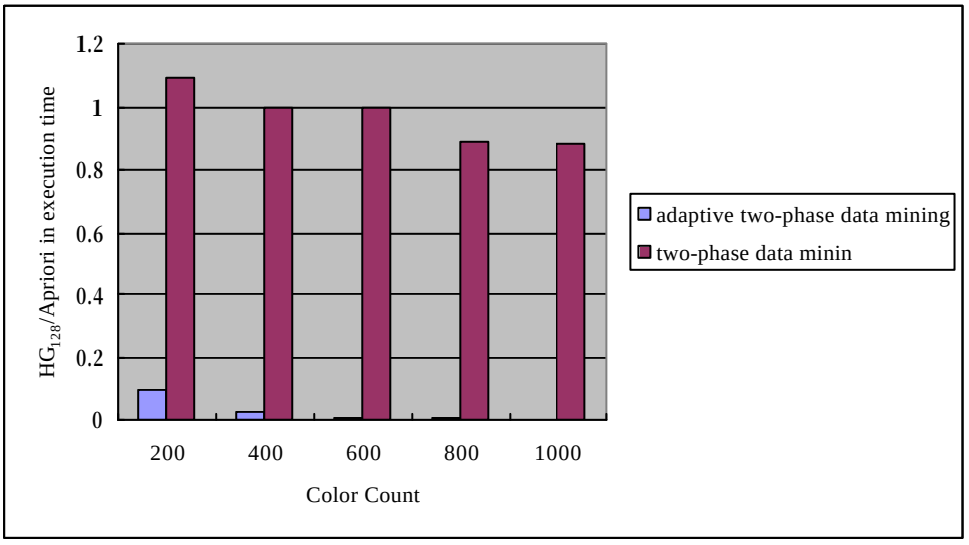


Figure 3. Execution time comparison between 128 HGs w.r.t. Partition (adaptive two-phase data mining) and 128 HGs (two-phase data mining) for some color count

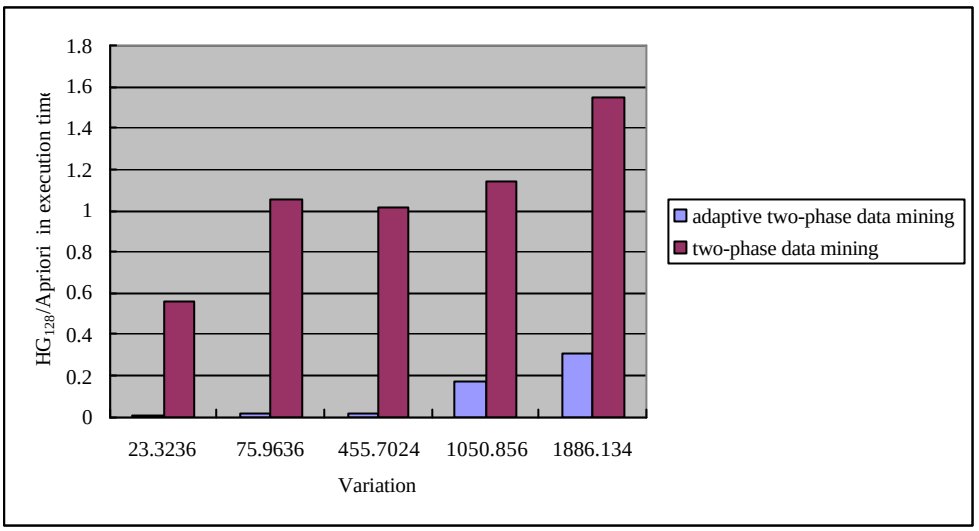


Figure 4. Execution time comparison between 128 HGs w.r.t. Partition (adaptive two-phase data mining) and 128 HGs (two-phase data mining) for some variation

5. Conclusions

Our proposed two-phase data mining method in this paper is essentially effective for the applications of data mining on remotely sensed images. Adaptive two-phase spatial association rules mining method conducts the idea of partition on an image for efficiently quantizing out non-frequent patterns and thus facilitate two-phase process.

References

[1] R. Agrawal, and R. Srikant (1994), "Fast Algorithms for Mining Association Rules in Large Database," Conference of Very Large Data Bases, Santiago, Chile, pp. 487-499.

[2] E. Clementini, P. D. Felice, and K. Koperski (2000), "Mining Multiple-Level Spatial Association Rules for Objects with a Broad Boundary," Data and Knowledge Engineering, Vol. 34, pp. 251-270.

[3] A. Denton, W. Perrizo, Q. Ding, and Q. Ding (2002), "Efficient Hierarchical Clustering of Large Data Sets Using P-trees," Proceeding of 15th International Conference on Computer Applications in Industry and Engineering, San Diego, CA, pp. 138-141.

[4] Q. Ding, Q. Ding, and W. Perrizo (2002), "Decision Tree Classification of Spatial Data Streams Using Peano Count Trees," Proceeding of ACM Symposium on Applied Computing, Madrid, Spain, pp. 413-417.

[5] Q. Ding, Q. Ding, and W. Perrizo (2002), "Association Rule Mining on Remotely Sensed Images Using P-trees," Proceedings of PAKDD 2002, pp. 66-79.

[6] M. Ester, A. Frommelt, H. P. Kriegel, and J. Sander (1998), "Algorithms for Characterization and Trend Detection in Spatial Databases," Proceeding of 4th International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA, pp. 44-50.

[7] M. Ester, S. Gundlach, H. P. Kriegel, and J. Sander (2000), "Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS

Support," Data Mining and Knowledge Discovery, Vol. 4, pp. 193-216.

[8] M. Ester, and H. P. Kriegel (1997), "Spatial Data Mining: A Database Approach," Processing 5th Int. Symposium on Large Spatial Databases, Berlin, pp. 47-66.

[9] M. Ester, H. P. Kriegel, and J. Sander (1999), "Knowledge Discovery in Spatial Databases," Conf. of 23rd German on Artificial Intelligence, Bonn, Germany, pp. 61-74.

[10] J. Han, and Y. Fu (1995), "Discovery of Multiple-Level Association Rules from Large Databases," Processing 21th International Conference Very Large Data Bases, pp. 420-431.

[11] K. Koperski, and J. Han (1995), "Discovery of Spatial Association Rules in Geographic Information Databases," Proceeding Fourth Advances in Spatial Databases Symp. Springer, Berlin, pp. 47-66.

[12] K. Koperski, J. Han, and N. Stefanovic (1998), "An Efficient Two-Step Method for Classification of Spatial Data," Proceeding of International Symposium on Spatial Data Handling, Vancouver, BC, Canada, pp. 45-54.

[13] Lee and Chen (2003), "An Efficient Spatial Association Rules Mining Method on Remote Sensed Image," *Processing of* 2003 *International Conference on Information Management* (*ICIM* 2003), Chaiyi, Taiwan, pp. 1144-1151.

[14] W. Lu, J. Han, and B. C. Ooi (1993), "Discovery of General Knowledge in Large Spatial Databases," Proceeding of Far East Workshop on Geographic Information Systems, World Scientific, Singapore, pp. 275-289.

[15] W. Perrizo, Q. Ding, Q. Ding, and A. Roy (2001), "Deriving High Confidence Rules from Spatial Data using Peano Count Trees," Proceedings of International Conference on Web-Age Information Management, Springer-Verlag, Lecture Notes in Computer Science 2118, pp. 91-102.

[16] K. Sayood (1996), "Introduction to Data Compression," Morgan Kauffman Publishers, San Fransisco, CA.