A NOVEL SIMILARITY ASSESSMENT FOR REMOTE SENSING IMAGES VIA FAST ASSOCIATION RULE MINING

Jun Liu [a], Kai Chen [a,*], Ping Liu [a], Jing Qian [a], Huijuan Chen [a]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 518055, Shenzhen China - (jun.liu, kai.chen, liuping, jing.qian, hj.chen)@siat.ac.cn

Commission II, WG II/3

KEY WORDS: Similarity Assessment, Fast Association Rule Mining, Multi-Dimension Data Cube, Remote Sensing Image, Image Retrieval

ABSTRACT:

Similarity assessment is the fundamentally important to various remote sensing applications such as image classification, image retrieval and so on. The objective of similarity assessment is to automatically distinguish differences between images and identify the contents of an image. Unlike the existing feature-based or object-based methods, we concern more about the deep level pattern of image content. The association rule mining is capable to find out the potential patterns of image, hence in this paper, a fast association rule mining algorithm is proposed and the similarity is represented by rules. More specifically, the proposed approach consist of the following steps: firstly, the gray level of image is compressed using linear segmentation to avoid interference of details and reduce the computation amount; then the compressed gray values between pixels are collected to generate the transaction sets which are transformed into the proposed multi-dimension data cube structure; the association rules are then fast mined based on multi-dimension data cube; finally the mined rules are represented as a vector and similarity assessment is achieved by vector comparison using first order approximation of Kullback-Leibler divergence. Experimental results indicate that the proposed fast association rule mining algorithm is more effective than the widely used Apriori method. The remote sensing image retrieval experiments using various images for example, QuickBird, WorldView-2, based on the existing and proposed similarity assessment show that the proposed method can provide higher retrieval precision.

1. INTRODUCTION

1.1 General Instructions

Image retrieval is a procedure finding images or sub-images that are similar to the input image from image dataset or a big image. The breadth of remote sensing image is large, the content is massive and complex, the spatial resolution spans various scales, and the phenomenon of “same object with different spectrum” and “different object with same spectrum” generally exists. These features bring huge difficulties to remote sensing image retrieval. The early metadata and text retrieval is far less to meet the current demands of remote sensing image analysis, and content-based image retrieval (CBIR) becomes the hot spot (Wang, 2008).

The CBIR technology is composite information retrieval, image processing, computer version, machine learning, and describes the content of image using the visual features extracted from images. The image retrieval is to find images from dataset with specific features or similar contents (Li, 2007). The current CBIR has made huge progress in retrieval using visual features. However, with the development of imaging technology, CBIR technology, with low level features such as color, texture and shape at the core, suffers from high feature dimension, incomplete content description, low accuracy, lacking of regularities, semantic gap (Do, 2002). The introduction of human visual attention model for remote sensing image retrieval provides a new way of thinking, to guide the image retrieval direction more in line with the human visual system. However, the study of cognitive psychology and neurophysiology, which is the basis of visual attention model, is still in the exploratory stage, and semantic cognition still has some limitations (Peleg, 2009). Meanwhile, the effectiveness of the existing models for remote sensing image retrieval has yet to be further verified and perfect.

Data mining is defined as "a non-trivial process of discovering from implicit, previously unknown, and potentially useful information", that is, finding knowledge from the data (Cao, 2011). The process from remote sensing to semantic information, could be seen as a process from image data to geography spatial knowledge. The objects in remote sensing images has a wealth of visual information (such as a significant corner information, arranged in an orderly distribution of texture, bright colors and a variety of features associated with regular spatial pattern, etc.). These can be used as sources of semantic information or knowledge (Ordonez, 2006). Using data mining technology can discovery various patterns and relationships that exist between the image pixels, the image and the auxiliary data, targets in images (Liu, 2011). Based on these patterns and relationships, we can further realize the image content analysis , induction and interpretation, and further discovery mode and knowledge of interest.

Hence, this paper proposes a fast association rule mining algorithm for remote sensing, and based on this algorithm, a novel similarity assessment is proposed for remote sensing image retrieval. The experimental results indicate that the

[*] Corresponding author

This contribution has been peer-reviewed.
doi:10.5194/isprsarchives-XLI-B2-217-2016

proposed method can efficiently improve the precision of image retrieval.

2. RELATED WORK

Association analysis is an important issue in data mining field, which is used to mined the meaningful relationship in massive data and express these relationship using association rules. In remote sensing image processing and information extraction, association rule mining could be used to find the frequent spatial pattern and establish relationship between them and other data.

Nowadays, there are many association rule mining algorithm, among them the Apriori algorithm proposed by Agrawal is always cited as the most classic one (Agrawal, 1993). Most later algorithms are based on Apriori algorithm. The following are some basic definition in association rule mining based on market basket data:

item: every good in basket is an item.

item set: some good in basket form an item set.

transaction and transaction set: Let $I = \{i_1, i_2, ..., i_n\}$ denotes all the items in market basket, where i_n is an item. A transaction has several items, and all the transactions form a transaction set $T = \{t_1, t_2, ..., t_m\}$, where t_m is a transaction.
k-item: It’s a set that has k items.

frequent k-item: It’s k-item whose support is bigger than minimum support.

support: It’s the number of transaction that has specific items. The support of item X is $\sigma(X)$:

$$\sigma(X) = |\{t_i \mid X \subseteq t_i, t_i \subseteq T\}| \qquad (1)$$

Association rules: It’s an expression like $X \rightarrow Y$, where X and Y are non-intersection items, i.e., $X \cap Y = \varnothing$. The support and confidence are used to measure the intensity of association rule, with the following expressions:

$$support(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \qquad (2)$$

$$confidence(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \qquad (3)$$

It’s necessary to traverse the whole transaction set when calculating the support of every candidate frequent item using Apriori algorithm, and the bigger number of transactions will lead to more computation. Hence, many improved Apriori algorithms focus on how to speed up the generation of frequent items and decrease the number of traversing transaction set.

3. PROPOSED METHOD

As mentioned in previous section, when the transaction set is large, the multiple traversing of transaction set would waste many computation. Therefore, how to decrease the time of traversing transaction set is the key of improving the efficiency of association rule mining. In this paper, the multi-dimension data cube is created to achieve this objective.

3.1 Multi-Dimension Data Cube Model (MDDC)

Every item in an transaction is defined as a dimension, and the range of item is defined as the length of this dimension. For example, for the following transaction set:

Transaction	Item set		
	Item1	Item2	Item3
T1	1	4	3
T2	5	2	4
T3	3	1	2

Table 1. Some transactions in a transaction set

This transaction set has 3 dimensions, i.e., Item1, Item2 and Item3, and the length of these 3 dimensions are 5, 4 and 4 respectively. Hence this transaction set could be represented as a 3-dimension data cube as follows:

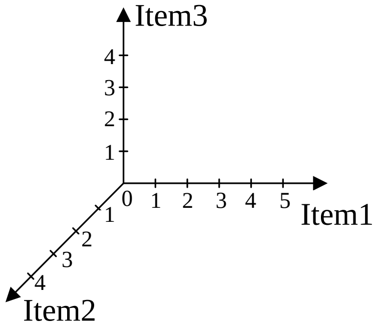


Figure 1. 3-dimension data cube

Every transaction in transaction set is represented using a point in this 3-dimension rectangular coordinate system. Similarly, a transaction set including N items could be represented using N-dimension data cube. The N-dimension data cube could be stored as a N-dimension array, for example, the T1 transaction could be expressed as C[1][4][3]=1. In the procedure of association rule mining, the transaction set is transformed into a MDDC only traversing transaction set one time, and the latter processing will be done using MDDC so as to improve computing efficiency.

3.2 Association Rules Mining Procedure

Following the definition of frequent item in Apriori algorithm, this paper also needs to obtain the k-item frequent set to generate association rules using MDDC.

The specific steps are as follows:

The transaction set is transformed into MDDC after traversing transaction set one time. For simplicity, every transaction consist of three items in this paper. The minimum support and confidence is defined as *min_sup* and *min_cof.*

(1) obtaining 1-item and 1-item frequent

The frequency of every value in 3 dimension are calculated using loop operator of array:

```
for k=1:maxkv
    sup_k=sum(A(k,:,:))/Len;
end
```

where maxkv is the max value in the first dimension. sup_k is the support of every value in the first dimension. sum() is the summation function. A(k,:,:) is the subset that values equal to k in 3-dimension data cube A. Since the length of A is limited and far less than the length of original transaction set, the summation calculation is very fast.

(2) obtaining k-item and k-item frequent set

The k+1-item is obtained by intersection of frequent k-item. Considering the storage of frequent k-item, it’s necessary to verify whether this frequent k-item could be intersected before intersection. If not, the intersection should be gave up so as to improve computation efficiency. In this procedure, there is no need to traverse the original transaction set, and relying on the addressing capability of array, the support could be fast obtained.

(3) generating association rules

For every frequent item, all the subset could be generated and a rule could be obtained by the following judgment:

$$\frac{support(l)}{support(s)} \geq min\_cof \tag{4}$$

where *l* is a frequent item, *s* is a subset of this frequent item. If this condition is satisfied, the follow rule is outputted:

$$s \Rightarrow l-s, support(l), \frac{support(l)}{support(s)} \tag{5}$$

where l-s is the left subset that removing s from l, $support(l)$ and $\frac{support(l)}{support(s)}$ are the support and confidence of this rule.

In the above association rule mining procedure, the original transaction set is traversed only one time. Although the MDDC is traversed many times when calculating frequent items and association rules, the time consuming of traversing MDDC is far less than traversing original transaction set. Hence the proposed method is obviously much faster than Apriori algorithm.

3.3 Similarity Assessment

For a three-band image, at every pixel, the pixel values are [a, b, c]. If the support and confidence are not considered, the following 12 association rules will be obtained:

a=>b	a=>c	b=>c
b=>a	c=>a	c=>b
ab=>c	ac=>b	bc=>a
a=>bc	b=>ac	c=>ab

Table 2. All the 12 association rules

Since the former 6 rules only involve the relationship between two pixels, they are not capable enough to describe the content of an image, and will increase the calculation amount of association rule mining and similarity assessment, hence in this paper, only the last 6 rules are used.

The gray level range of a band of an remote sensing image is 0 to 255, hence in order to decrease the dimension of multi-dimension data cube, the gray level of every band should be compressed. The simplest linear compression is used in this paper. Assuming the gray level of compressed image is G, then the gray value of compressed image is:

$$g' = ceil(\frac{g+1}{256} * G) \tag{6}$$

where g is the current gray value, ceil is a round up function. The gray level of 0 to 255 will be compressed to 1 to G.

The association rules are first mined from an image using the MDDC, the all the association rules are transformed to rule vector which could represents the content of this image. The generation of rule vector is as follows:

Firstly, all the six rules are numbered, for example, rule ab=>c is numbered as n=1, and c=>ab is numbered as n=6. Then the basic vector of a rule is *a*G*G+b*G+c*, and the final vector coordinate of this rule is *(a*G*G+b*G+c)*6+n*, with the vector value of sup*conf. Once all the vector of rules are calculated, the content of this image could be described using the vector histogram. Finally the similarity between two images could achieved by calculating the distance of two rule vector histogram using the first order approximate distance of Kullback-Leibler divergence:

$$dis = \sum_{i=1}^{N} \frac{(r1(i) - r2(i))^2}{r1(i) + r2(i)} \tag{7}$$

where *r1* and *r2* are two rule vector histogram.

4. EXPERIMENTS AND DISCUSSION

4.1 Experiments for mining efficiency

For a 300 pixels by 300 pixels image with three bands, the compressed gray level is set as 8, then every transaction has 3 items, and the number of transactions is 300*300= 90000. In order to test the transaction number influence on algorithm performance, the number of transactions is increased from 30000 to 90000. The minimum support is increased from 0.1% to 0.5%, and the minimum confidence is 0.2. The relationship between time consuming and support change are calculated. The testing environment is notebook computer, windows 7 operation system, 4G RAM, i5 CPU with 2.5GHz, and Matlab R2010b. The results are shown in figure 2, where “3w” denotes that the transaction number is 30000, and the meaning of other legends are similar to “3w”.

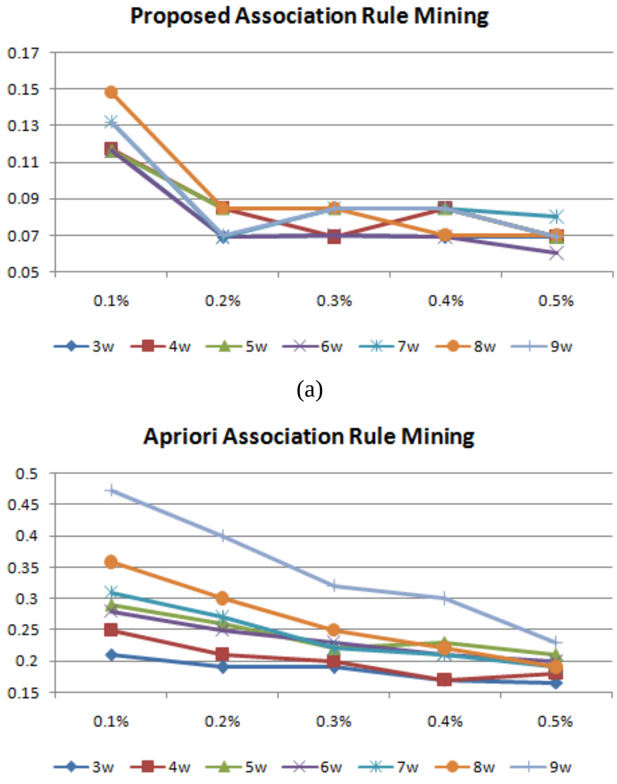
Figure 2 shows the images before and after gray level compression. For better visual effect, the compressed image is stretched to 0-255 using linear transformation.

This contribution has been peer-reviewed.
doi:10.5194/isprsarchives-XLI-B2-217-2016

(a) (b)

Figure 2. Images before and after gray level compression. (a) before (b) after

The performances of proposed method and Apriori algorithm are shown in figure 3. The x axis is support and the y axis is time consuming. The time consuming includes the time from reading image to outputting association rules.



(a)

(b)

Figure 3. Forest retrieval results for WorldView-2 image dataset (a) proposed method (b) DT-CWT method

Overall, the time consuming of Apriori algorithm is 2-3 times that of proposed method. The time will decrease with the increase of minimum support because the number of frequent items will decrease. During the whole mining procedure, the original transaction set is traversed only one time, and the transaction set is transformed into multi-dimension data cube, hence the whole time consuming is very low because that the traversing efficiency of data cube is higher than that of transaction set. When the minimum support increases to some degree, the time consuming will retain stable with some small changes. Therefore, for a large image dataset, the proposed mining algorithm could obviously improve the association mining efficiency.

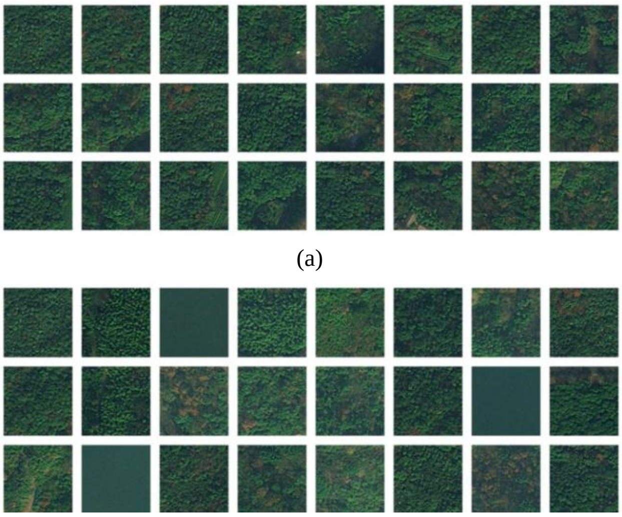
4.2 Experiments for image retrieval based on QuickBird and WorldView-2 image datasets

The QuickBird and WorldView-2 image databases were used to test and compare the performance of proposed method and the other traditional retrieval methods including histogram matching, Gabor wavelet, DT-CWT and color moment. Four classes i.e., settlement and open forest for QuickBird, forest and water for WorldView-2, were used. For the returned result of every input image, the top 8, 16, 24, 32, 40, 48, 56 and 64 images were chosen to count how many images were right results, and this ratio named average precision, was considered as the final performance of all methods with the following expression:

$$c = \frac{1}{M} \sum_{i=1}^{M} \frac{N_{correct\_i}}{N_{sum\_i}} \tag{8}$$

where M is the number of input images, N_{sum_i} and $N_{correct\_i}$ are the total number and correct number of returned images for the i-th input image respectively. The bigger c shows the better retrieval performance for a method.

Figure 4 show the top 24 retrieval results of forest for WorldView-2 image dataset using the proposed method and DT-CWT method. We can see that three returned images from DT-CWT results are obviously incorrect while for the proposed method, these 24 returned images are all correct.

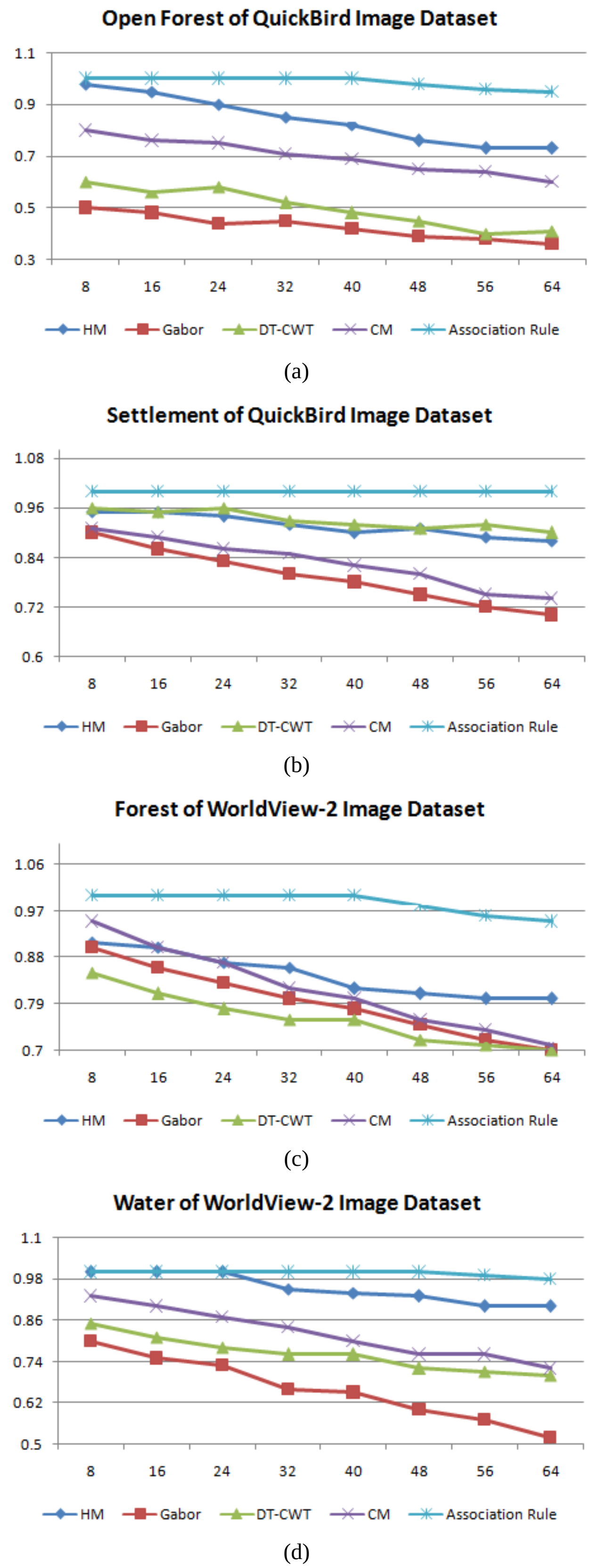


(a)

(b)

Figure 4. Forest retrieval results for WorldView-2 image dataset (a) proposed method (b) DT-CWT method

Figure 5 are the objective evaluation results using average precision index of all employed retrieval methods. It's obvious that the proposed method can provide higher precision for all classes and show robustness to some extent. The performance of other methods are various according to different classes.



(a)

(b)

(c)

(d)

Figure 5. Objective evaluation results using average precision index. (a) and (b) are retrieval results of open forest and settlement from QuickBird image dataset, while (c) and (d) are the results from WorldView-2 image dataset

5. CONCLUSION

Association rules can represent the deep pattern and relationship of the content of images. This paper proposed a fast association rule mining algorithm using multi-dimension data cube to decrease the traversing time of original transaction set, and then proposed a novel similarity assessment for image retrieval. The experimental results using QuickBird and WorldView-2 images indicate that the proposed method can provide results with higher precision compared with traditional CBIR methods. The future work will focus on finding more features and optimizing the multi-dimension data cube model.

ACKNOWLEDGEMENTS

This work is jointly supported by the National Natural Science Foundation program (No. 41301403, 41471340) and the Basic Research Program of Shenzhen (No. JCYJ20150831194441446, JCYJ20150630114942312 and JCYJ20150630114942260).

REFERENCES

Wang Z. Z., Yong J. H., 2008. Texture Analysis and Classification With Linear Regression Model Based on Wavelet Transform, *IEEE Transactions on Image Processing*, 17, pp. 1421 – 1430.

Li Y.K., TIMO R.B.. 2007. Semantic-sensitive satellite image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 45 (4), pp. 853-860.

Do M. N., Vetterli M., 2002. Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Transactions on Image Processing*, pp.146 – 158.

Peleg S., Naor J., Hartley R., Avnir D., 2009. Multiple Resolution Texture Analysis and Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, pp. 518 – 523.

Cao L.B., Zhang H.F., Zhao Y.C., et al. 2011. Combined Mining: Discovering Informative Knowledge in Complex Data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*: Cybemetics, 41(3), pp. 699-712.

Ordonez C., 2006. Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine*, 10, pp.34 – 343.

Liu J., Shao Z.F., 2011. Texture image retrieval based on LogPolar transform and association rules mining. *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp.952 – 956.

Agrawal R., Imielinski T., Swami A.. 1993. Mining association rule between sets of items in large databases. *ACM SIGMOD*. pp. 207-216.