SOME KEY TECHNIQUES ON UPDATING SPATIAL DATA INFRASTRUCTURE BY SATELLITE REMOTE SENSING IMAGERY

Peijun Du[a] , Yunhao Chen[b] *

[a] Department of Remote Sensing and Geographical Information Science,
China University of Mining and Technology, Xuzhou City, Jiangsu Province 221008, P.R.China dupj@vip.163.com
[b] College of Resources Science, Beijing Normal University, cyh@bnu.edu.cn

KEYWORDS: Information Updating, Spatial Data Infrastructure (SDI), Satellite Remote Sensing, Change Detection, Data Mining, Content-Based RS Image Retrieval

ABSTRACT:

In general only a small part of spatial data in SDI change because of different factors, so rapid and effective updating methods are very vital. With the improvement of spatial resolution of satellite remote sensing (RS) imagery, it is possible to update spatial data infrastructure efficiently by satellite RS images. But the RS data volume is vast, so high processing capacity is required. It is impossible to update spatial data without the support of high performance of RS image management, retrieval, and change detection and pattern discovery. After introducing some background information including significance of SDI updating, feasibility and advantages of satellite RS imagery used to update spatial data, the framework of updating spatial and attribute information based on RS image is proposed. In the process, the integrated processing of raster data and vector data is very important, so some functions of RS image processing and GIS software should be fused. In order to extract the anticipated images from vast image database, effective retrieval technique is vital. Based on content-based image retrieval, content-based RS image retrieval is put forward and some topics including retrieval pattern, useful image content, feature extraction and similarity measure are researched. After related images are retrieved from vast image database, it is necessary to discover those change areas, so change detection from multi-temporal RS images are discussed further. RS image data mining and knowledge discovery is the synergy of Spatial Data Mining (SDM) and Image Data Mining (IDM). Oriented to the demands of SDI updating to intelligent information processing, some primary issues on RSDM are analysed. It is pointed out that updating SDI by satellite RS imagery will be potential advantageous in the future, and all the techniques discussed in this paper including content-based RS image retrieval, change detection, multi-temporal RS image fusion and RS Data Mining and Knowledge Discovery are very important and will play important roles in the future.

1. INTRODUCTION

Both the establishment and updating of spatial data infrastructure (SDI) are important to the efficient applications of SDI, but much attention is paid to the former at present and data updating has not got enough emphasis. After the fundamental SDI is created, it is necessary to update and maintain the database in order to guarantee accuracy and performance. In general only a small part of spatial data in SDI change because of different factors, so rapid and effective updating methods are very vital. With the improvement of spatial resolution of satellite remote sensing (RS) imagery, especially the satellite RS images with 1m even higher resolutions, it is possible to update spatial data infrastructure rapidly by satellite RS images. But the data volume of RS imagery is vast, so high processing capacity is required. It is impossible to update spatial data without the support of high performance of RS image management, retrieval, and change detection and pattern discovery. In this paper some issues on the above topics will be discussed.

2. THE FRAMEWORK OF UPDATING SPATIAL INFORMATION BASED ON RS IMAGERY

Both vector and raster data are used in urban, regional and national spatial data infrastructure (SDI), so it is necessary to process image data with raster format and map data with vector format simultaneously. In this process, some operations including transformation of raster and vector, multi-scale transformation, generalization and other are involved. Although some available RS and GIS software are powerful, their foci are different. RS image processing software emphasizes image transformation, classification, information extraction and so on, and their abilities to process vector data and geographic information are insufficient. In addition, it necessary to design and develop such geographic information updating system that is efficient to renew information with the support of artificial intelligence and knowledge engineering. Unlike to conventional RS image processing system, this system aims at updating geographic information in SDI by high-resolution RS image; so not all common image-processing functions are required and the images that are processed by other RS system can be used as the input of this system, and the foci of this system are change detect, image analysis, transformation of raster and vector data, automatic reasoning and information updating.

[*] Corresponding author: 86-10-58806098, cyh@bnu.edu.cn

This system should have friendly interfaces with RS and GIS systems and powerful functions of information processing based on the integration of RS and GIS. Some main functions of this system include image data entry, image pre-processing, data link with GIS, historical information retrieval, change modes reasoning, creation and maintenance of knowledge base, automatic change detect, fuzzy processing, segmentation, information fusion, map editing and automatic updating. In other word, it looks like a middle-ware between RS and GIS and updates information in GIS by processing RS image.

The working flow of this system can be described as follows. The former images of a given area are retrieved from image database by taking the imported image as mask, and then multi-temporal information fusion and change detect are used to extract the change areas. After that the change areas are processed guiding by related knowledge and rules and those reliable change regions are determined. The change regions are cut by a mask and then classification and analysis are conducted to the image block in order to get the raster and vector image about those regions. Finally the map databases are opened and such operations as map segmentation, map editing, automatic generation of polygon, reconstruction of topological relationship and map merging, and then metadata and corresponding temporal records are updated according to the requirements of SDI and related processes.

So some issues about RS image fusion, change detect, image retrieval and knowledge discovery play important roles to update information in SDI by RS images, and that will be the topic of this paper.

3. CONTENT-BASED RS IMAGE RETRIEVAL

With the rapid development of Earth Observation System (EOS), the spatial resolution, temporal resolution and spectral resolution of Remote Sensing (RS) image have been improved greatly. Many regions have accumulated vast RS information and formed multi-temporal, multi-resolution, multi-scale, multi-platform, multi-spectral or hyperspectral RS information pyramid. After vast RS information was acquired, how to manage and use them effectively became an important problem on RS information science and RS applications. Among all RS information processing schemes, how to retrieve and get necessary, interested and effective information oriented to a given task is the basis of all other applications. That is to retrieve images from RS image library (or RS image information system). At present, studies on RS imagery retrieval are still in its starting period, so the achievements and experiences in general image retrieval should be introduced and applied. Content-based image retrieval (CBIR) is one of the hot issues in image retrieval, and it can be used to RS image retrieval. That lead to a new area: content-based RS imagery retrieval.

3.1 Basic Concepts

Early image retrieval was similar to text retrieval and was realized by remarks and notes to image, so all images should be remarked firstly and then retrieval can be done through matching remark text. This type of retrieval have some disadvantages: (1) the work of image remarking is very large; (2) it is difficult to select effective remark text that can represent abundant information in image; and (3) the retrieval efficiency is very low. So CBIR (Content-Based Image Retrieval) was paid more attention and got wide applications. Nowadays related studies are focused on medical image retrieval, scanned image retrieval in digital library and trademark image retrieval, and so on. Some software used to image retrieval was already developed successfully, such as QBIC developed by IBM company, Chablo by Berkley University and Photobook by MIT. In China, studies about CBIR were started from 1990s and many achievements were got. Especially wavelet, fractal, ANN and other new theories and techniques have been used to image retrieval. It should be pointed out that the major features used to image retrieval were colour histogram, shape and texture at present because the retrieval objects were small digital images such as medical images, trademark images and others.

Similar to general image, RS image retrieval can be realized by different methods. Early retrieval is realized by image remarks and descriptions, that is text-based retrieval. Content-based retrieval is the hot issue in recent years, and semanteme based retrieval is also emphasized. Figure 1 is the level and framework of RS image retrieval. Content-based RS image retrieval is the process of retrieving and getting interested images from RS image library by the visual features and quantitative index based on image content and features. Figure 2 is the framework of content-based RS image retrieval, and Figure 3 is a typical processing flow of content-based RS image retrieval.

RS image retrieval system can provide users with interactive and visual interface and analyse the retrieval conditions and patterns further, and then those images satisfying the retrieval conditions according to certain index and similarity measurement rules will be given to users.

In content-based RS image retrieval, some important problems include selection of retrieval rules, effective feature extraction and combination, retrieval process and efficiency optimization and so on. Among those analysis to useful features and retrieval patterns are the most important.

3.2 Retrieval Modes

According to current RS information application requirement and characteristics, the main modes used to retrieve image from RS image library can be listed as follows.

3.2.1 Retrieval based on Mask (or Example):The mask is provided as retrieval demonstration and related indexes and parameters are computed by feature of mask, and then those features are compared with corresponding feature of image library by similarity measurement, finally the interesting images can be got by the index of matched

3.2.2 Retrieval based on Attribute: In this mode, the retrieval condition is some attribute or content description about target object, such as vegetation index, gray and other spectral attributes. Here, the relationship between attribute and feature should be created and then corresponding image content can be used to retrieve and match.

3.2.3 Retrieval based on Description Information: Sometimes only some qualitative and quantitative information about aim image is given and used to retrieve. For example, the capture time, region, sensor and some qualitative are given, and those conditions will be used as retrieval parameters.

3.2.4 Retrieval based on Semanteme: Semantic information also can be used as retrieval conditions. The key is how to form the relation between semantic information and RS images, especially how to transform the semantic information to quantitative description about RS images. The studies on this field are few, and much work should be done in the near future.

3.2.5 Synthetic Retrieval Mode: Each retrieval mode can adapt different conditions, and different modes should be integrated to form synthetic retrieval mode in order to solve some difficult tasks.

3.3 Contents Useful in RS Image Retrieval

Color, shape and texture are used mainly in general image retrieval, and they are still useful in RS image retrieval. On the other hand, some contents characterized by RS image include spectral feature, spatial attribute and relationship and metadata must me emphasized and used.

3.3.1 Color: Color is a very important feature in aerial RS image and other single band image. Histogram is the major tool to express color feature, and histogram feature computation and matching can be used to realize image retrieval. The major advantage of color feature is its invariant with rotation and scale. The primary idea of color-based retrieval is to transfer the difference in image to discrimination of histogram, so image retrieval will become histogram matching.

RGB (Red, Green and Blue) color system is usually used to express colorful image. When it is used to image retrieval, the most universal measurement standard is to compute the distance between two feature vectors : (μ_R，μ_G，μ_B) representing the two images respectively in RBG space, and different distances can be used. By comparison, it is proposed that accumulated histogram is superior to general method, and weighted distance is not more precise than Euclidean distance, but central moment is more advantageous in algorithm, efficiency and precision.

3.3.2 Shape: Shape is also important image content used to retrieval. Compared with colour feature that is sensitive to illumination and other external factors, shape is the inherent property of spatial entity and is not affected by other conditions, but it has some shortages in rotation-invariant, displacement-invariant and scale-invariant. Despite those disadvantages, shape is very useful in RS image retrieval because the shape of ground entity can represent features in the image effectively.

In RS image, shape can be viewed as an effective retrieval index, especially to some linear entities (for example, river, road) and objects with regular shape (for example, building, farmland). The basic idea is: (1) to set up the index file by vector data structure according to a certain shape index after image segmentation or edge extraction, and (2) to extract the same shape index from retrieval template, and (3) to match the shape index of retrieval template and images in library by similarity measurement, and (4) to get the retrieval results.

3.3.3 Texture: Texture is used widely in RS image identification and classification, and is one of the typical features of different ground entities. Compared with general images, texture of RS image is more complex and abundant, and is sensitive to different entities. Usually texture is taken as an aided feature of classification, and also used in image segmentation and typical information extraction and so on. Of course, texture features can be used effectively in RS image description and measurement in image retrieval.

Many methods to compute and analyze texture features in RS image have been proposed, including gray co-occurrence matrix, markov random field, Gabor transformation, fractal, wavelet, texture unit and others. Among all methods, gray co-occurrence matrix, fractal and wavelet methods are used more widely.

The process of texture-based RS image retrieval is: (1) to compute texture feature and texture image segmentation, and (2) to establish the index, and (3) to compute the texture feature of retrieval template, and (4) similarity measurement and retrieval.

3.3.4 Spectral Feature: Spectral characteristic is the physical basis of RS, both reflectance and absorption feature can be used and the former will be taken as an example as follows. Different objects have distinct albedo, so different radiation energy will be received by RS sensor and that will leads to the difference of gray among different ground objects in image. The spectral characteristics of ground object can be expressed by spectral curve. Generally, spectral feature can be used to describe and distinguish the ground objects, so it can be used to retrieve RS image also. Spectrum-based retrieval is the characteristic of RS image, and it can reflect object features from spectral dimensional that is not used in common images, so it can be used to multi-spectral and hyperspectral RS image retrieval. The spectral relationship can be both linear combination and non-linear transformation of different band images, and its basis is spectra feature mining in RS image.

3.3.4.1 Feature extraction and analysis in multi-spectral RS image: Spectral feature is used to measure the information among different bands. For example, the most universally used Landsat TM image has seven bands, and each has corresponding application field. On one hand, these bands can be used individually; on the other hand, the combination of different bands can form some new useful features. The most useful multi-spectral feature is vegetation index, and normalized difference of vegetation index (NDVI) is most universally used. NDVI is computed by Equation (1):

$$NDVI = \frac{DN_{Infra} - DN_{\mathrm{Re}\,d}}{DN_{Infra} + DN_{\mathrm{Re}\,d}} \tag{1}$$

Here, DN_{Infra} is the gray of infrared band (the fourth band of TM image)，DN_{Red} is the gray of red band (the third band of TM image).

Similar to NDVI, NDBI (Normalized Difference of Building Index) used to extract construction from TM image was proposed, and it is computed by Equation (2):

$$\mathrm{NDBI} = (DN_{TM5}+ DN_{TM4}) /(DN_{TM5}+ DN_{TM4}) \tag{2}$$

In addition, Zhou proposed the band combination used to extract water from TM image. We can describe it with NDWI (Normalized Difference of Water Index) like NDVI and NDBI. It is computed from Equation (3):

$$\mathrm{NDWI} = (DN_{TM2}+ DN_{TM3}) - (DN_{TM4}+ DN_{TM5}) \tag{3}$$

Only for water pixels NDWI is larger than 0, and other ground with NDVI less than 0.

Other spectral features were be used in some studies, for example, The square root of sum of gray square in all bands, multi-band texture and so on. All those features can be used to multi-spectral RS image retrieval according to specific conditions.

3.3.4.2 Spectral Feature Extraction and Analysis in Hyperspectral RS Image Retrieval: Only some limited bands were used in multi-spectral RS information. It expressed the spectral features of ground objects only by some wide and discrete bands and couldn't express spectral features fully. But hyperspectral RS can express ground object by spatial-spectral dimensional information using tenths or hundreds of divided spectral and continuous narrow bands. In hyperspectral RS image, spectral vector of pixel can be used directly, and its key is similarity measurement of spectral vectors. Spectral angle (SA) and Spectral Information Divergence (SID) are used widely. SA can be computed by Equation (4).

$$\cos\alpha = \frac{\vec{A}\bullet\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i A_i}\ \sqrt{\sum_{i=1}^{N} B_i B_i}} \qquad (4)$$

Here, N is the amount of band, $\vec{A}$ = (A_1, A_2,······, A_N) and $\vec{B}$ =(B_1, B_2,······, B_N) are two pixel spectral vector respectively, their elements A_i、B_i is the albedo of the *i*th band, and α is spectral angle. In practice, spectral angle, α isn't necessary to compute, and its cosine can be used directly. The more similar of the two vectors, the nearer to 1 cos α is.
SID can be computed by Equation (5):

$$SID\ (\vec{A},\vec{B}) = D(\vec{A}\|\vec{B}) + D(\vec{B}\|\vec{A}) \qquad (5)$$

Here,

$$D(\vec{A}\|\vec{B}) = \sum_{i=1}^{N} p_i \log(p_i / q_i),$$

$$D(\vec{B}\|\vec{A}) = \sum_{i=1}^{N} q_i \log(q_i / p_i)$$

$$p_i = A_i / \sum_{i=1}^{N} A_i, \qquad q_i = B_i / \sum_{i=1}^{N} B_i$$

It proved that SAM and SID are more useful and effective to measure the similarity between two spectral vectors in hyperspectral RS images than correlation coefficient and Euclidian distance, so they are used more frequently to similarity measurement in RS image retrieval.

3.3.5 Spatial Relationship and Spatial Feature: RS image is a special kind of image with precise spatial position and complex spatial relationship, and those relationship and features should be used RS image retrieval. For example, spatial inclusion and adjacency can be used to retrieve interesting images or blocks, and that is similar to extended SQL query or GeoSQL in GIS.

Here, spatial relationship predications are the key to describe retrieval conditions. Some useful spatial predications include CROSS, INSIDE, CONTAIN, ADJACENT, BUFFER, CORRELATION and so on. Those predications can be combined with some other conditions to realize RS image retrieval based on spatial relationship. In some simple spatial retrieval, spatial location (coordinate) can be used to retrieval directly. But in more occasions, spatial relationship should be combined with other condition.

3.3.6 Metadata: Metadata is "data about data". It plays important roles in spatial information system, data warehouse to manage information system, control data quality and convenience user visit.

In retrieval, metadata can play the role of "filter". That means some images that don't belong to retrieval scope obviously will be abandoned according to metadata firstly so the retrieval scope will be reduced and efficiency will be improved. In some conditions, metadata can be used as retrieval conditions directly.

3.3.7 Knowledge and Rules: RS information processing based on domain knowledge, establishment of RS image knowledge library and expert system and RS image data mining are some hot issues in RS information science recently. Correspondingly, knowledge and rules are possible to be viewed as the advanced expression form of image content in RS image retrieval. For example, in studies on Land use/cover change (LUCC), sometimes it is necessary to retrieve those images reflecting land use change information and trends, so some knowledge and rules about land use change must be proposed and connected with specific image feature, and then the can be used to retrieve useful images. Because the establishment of knowledge and rule library and reasoning mechanism is complex and difficult, this problem should be researched deeply in the future.

All above features can be used in content-based RS image retrieval, and how the efficiency and performance of different features are and how to combine those features and optimize retrieval flow should be determined according to specific goals and practical situation.

4. MULTI-TEMPORAL RS INFORMATION FUSION

The fusion of multi-source and multi-temporal RS information is one important aspect of RS information processing, and it consists of multi-source RS information fusion for improving classification accuracy and multi-temporal RS information fusion for dynamic monitoring in general. The key of satellite RS information used to dynamic monitoring lies in change detect or extraction. Some commonly used change detect algorithms include image computation, difference in VIs, change vector analysis, PCA approach, spectral features variation and comparison of classification results, but there is not a united framework for all algorithms. Based on the theoretical basis, principles of algorithms, operation objects and application modes, it is proposed that all algorithms can be categorized into four groups: image computation, image transformation, classification statistics and model-based analysis in this paper, and each group includes some common approaches.

4.1. Image Computation

Suppose that F_i is the image of the *i*-th period, and the change ΔF_{ij} between two periods can be detected by image computation. It can be demonstrated by Equation (6):

$$\Delta F_{ij}=F_j-F_i \tag{6}$$

Fi can be expressed by gray of pixels, or statistical parameters of texture. But the grays of the same land cover on two periods are different in general because of different image capturing time, climatic condition, zenith and other factors. Although some corrections can be used to reduce the differences, the precision is not very high. Usually a threshold is selected to decide the change area because some given laws exist in the gray change of multi-temporal RS images. In order to overcome the influences of gray, the variation or difference of Vegetation Index (VI) is used in some studies. It is suitable for the change of vegetation, farmland, and other green land covers, but it can't be used to another change, such as the change from rural land to urban land or traffic land.

4.2 Image Transformation

Some image transformation algorithms that can be used to dynamic monitoring include principal component analysis (PCA, or K-L transformation), K-T transformation and frequency transformation (e.g, Fast Fourier Transformation, FFT and wavelet transformation) and canonical analysis (CA transformation), and PCA is used widely at present. The primary idea of multi-temporal RS information processing is to emphasize the main information (especially, change information) by image transformation. It consists of three transformation modes: PCA to image differences, differences of PCs and PCA to multi-temporal images. In PCA to image differences approach, the differences of two images is got at first and then KL transformation is used to those difference images, so the first component will hold the main difference information image, or change information. Differences of PCs is realized by computing the difference of corresponding PCs of two images, especially the differences of the two first PCs are used to detect change information. In PCA to multi-temporal images, a mixed data set is created by mixing image of every band of two periods and KL transformation is conducted to this mixed data set, so the former two PCs are about some same information and the change information can be detected by the third and fourth PCs.

4.3 Classification Statistics

The qualitative, quantitative, and location detection and studies can be realized by the statistical analysis of multi-temporal RS classification results. Three methods are used to this level usually.

The first is encoding-based analysis, in which the class of each pixel on two periods are combined to form a new code and this code is compared with the pre-known characteristic codes of TSE, and then those pixels matching the change codes are extracted, so the qualitative, quantitative, and locational analysis can be implemented by the codes since the change type, spatial location can be got and summarized.

The second method is the analysis based on change matrix. The change matrix is derived from the area change of different land use types in two periods, and it can be used to analyze the change status and its driving mechanism and trends.

The third method is change detection with support of GIS. Much background information in GIS can be used as ancillary data of TSE monitoring. The main factors of TSE can be captured by regional investigation and statistical analysis and those factors can be used to determine those sensitive regions to change and create their buffers. Multi-temporal RS information of the possible change regions is used to further classification and analysis. The spatial analysis modules and data in GIS are used in this method, so the data volume of RS data is reduced, and the efficiency and accuracy can be improved to a great extent. But GIS is required at first in this mode, and that is impractical to many mining areas at present.

4.4 Model based Analysis

Some specific models which use RS data (gray, texture, VIs, and so on) as the their parameters are used to express evolution in model-based analysis so that the change detection and statistical analysis can be done. RS Information Model (RSIM) and other models driven by RS data are very effective to geographical process. But model generation involves geoscience's mechanism, spectral signatures, parameter identification, experiment and analysis, so it is a very complicated process.

5. RS DATA MINING AND KNOWLEDGE DISCOVERY

Data Mining (DM) and Knowledge Discovery in Database (KDD) has become an advanced information technology since 1990s, and it is the integration of multi-discipline like machine learning, artificial intelligence, database, statistics, scientific computation visualization and others. DM aims at discovering those hidden, unknown and useful information and knowledge from vast, incomplete, noisy and fuzzy data and providing intelligent and automatic means to data processing and understanding. In general, KDD is viewed as a complicated procedure including data selection, preprocessing, data transformation, data mining, knowledge expression and explanation, knowledge management and application, and DM is the most important step. In recent years, DM in relational database and transaction database has developed quickly and already got some applications in different domains. At the same time, text data mining, multimedia data mining (especially image data mining) and spatial data mining (SDM) have become new hot issues.

5.1 State-of-the-Art of Related Studies

RS image is both a kind of image information and a type of spatial data. So Spatial Data Mining (SDM) and Image Data

Mining (IDM) are two fields with close link to RS data mining. In this section some research progress will be introduced.

5.1.1 Studies on SDM: SDM is one of the hot issues of spatial information science in recent years, especially DM and KDD in GIS database has got many achievements. Researchers in China have done more work in this field. Li Deren proposed knowledge discovery in GIS database and gave deep studies on the theories and techniques of SDM. Di Kaichang gave systematic and overall studies on spatial data mining and knowledge discovery (SDMKD) including its framework, processing flow, main algorithms and some practical applications. In addition, researchers in CAS have done some studies in this field. By SDM knowledge and rules about spatial association, feature, classification, clustering, patterns and trends can be discovered and used to understand spatial data, find relations between spatial and attribute data, build spatial knowledge library, optimize query and spatial database, and acquire simple and effective features. The main SDM algorithms include probability theory, D-S theory, spatial statistics, deduction, clustering, spatial analysis, fuzzy set, cloud theory, rough sets, ANN, Genetic Algorithm (GA), visualization, decision tree and so on. But there are still many key theoretical and technical problems should be solved including multi-source spatial data cleaning, data mining taking uncertainty into account, incremental mining, raster and vector data integrated mining, multi-resolution and multi-level mining, parallel mining, spatial data mining primitives and data mining on distributed and network database.

5.1.2 Progress of RS Image Data Mining (RSIDM): The difference between RS Image Data Mining (DM) and general image analysis is that the latter pays more attention on some specific patterns and characteristics and belongs to scope of pattern recognition, but the former emphasizes on discovering those rules and knowledge by analyzing all significant and meaningful patterns in image sets without any pre-known conditions. It can be known the two operations can cooperate and promote each other. General image processing and analysis techniques can be used to process those images involved in DM, for example, image segmentation, feature extraction, and so on, and some methods (for example, clustering) can be used as DM methods directly, and DM can provide image processing with rules and knowledge that are important to intelligent information processing.

Nowadays studies on RS Data Mining are focused on machine learning, model of DM system, typical applications and other fields.

On machine learning field, Daniel Charlebois discussed machine learning in RS analysis and applied it to expert system; Michael Schroder analyzed the interactive learning in RS image library and applied it to content-based image retrieval; Mihai Datcu combined RS data interpretation and image data mining based on multi-level Bayesian learning model and applied it to processing of SAR, STRM and other RS data.

On RS data mining model and framework fields, Leen-Kiat et al proposed a common model used to RS image data mining: DIMUS (Data Investigation Model for Unsupervised Segmentation) that consisted of checking tools, clue generator, classifier, adjustor and mapping tools, and this model was used effectively in vegetation mapping by NOAA-AVHRR image, land use analyzing by TM image and sea-ice image analyzing by ERS-1 SAR and RADARSAT image.

On typical applications field, Aldridge used RS-GKDD (Rough Set Based Geographic Knowledge Induction) approach with support of Rough Set theory to analyze the landslide images and find relationships among elevation, rock properties and landslide; LasKaris investigated DM techniques in multi-source image analysis based on Hopfield neural network; George Brannon Smith extracted the rules describing direction relations of entities from raster mage by fuzzy data mining methods; Sara McCaslin researched knowledge discovery from multi-spectral satellite images by fuzzy neural network and extracted related classification rules; Eklund extracted knowledge about soil salinity from TM image and geographical data by C4.5 algorithm; Huang extracted knowledge about wetland classification from GIS data and SPOT images by C4.5 algorithm; and so on. All those work has given useful tests to RS image data mining and some new theories including rough set, fuzzy theory, neural network and others have been used.

Chinese researchers have also given studies on RS data mining in recent years. He Guojin investigated the theoretical methods of satellite RS data mining and knowledge discovery and proposed that RS data mining include such steps as image selection, pre-processing, feature analysis, information identification and knowledge interpretation; Zhou Chenghu put forward a inhabitant land automatic extraction algorithms based on knowledge discovery; Di Kaichang researched RS image classification methods based on inductive learning; Li Deren applied data mining to object identification and automatic classification in novel RS images; Liu Weiguo discovered knowledge from GIS database and used them to vegetation in RS image.

In addition, JPL has developed an applied image data mining system and used it to satellite detection. GeoMiner was also a spatial data mining system and can be used by RSIDM.

5.2 Some Problems should be solved

From above sum-up and analysis, it can be known that many significant knowledge and rules that are useful to image understanding and applications can be discovered from RSIL by DM technique. But current studies are focused on applying DM methods in relational database to RS image and only some experimental data are used in those studies, so related studies should be researched more deeply. Nowadays, at least the following problems about RS image data mining should be researched and solved.

(1) **The framework of DM in RS image library is not constructed.** Although some thoughts and methods of DM have been used in RS image processing, it still belongs to the category of conventional methods. The systematic and overall framework of DM in RS image library wasn't established, and some related problems like data preprocessing, organization and DM procedures weren't researched. It is known that knowledge including association rules between RS image information and geo-activities and objects, RS information characteristics of ground object, classification and clustering rules, dynamic evolution rules of spatial process can be discovered, but the properties of each kind of knowledge, demands RS image to DM algorithms and operations and other similar problems have not been paid enough attention so far. Obviously those

problems should be solved at first in RS image data mining and its applications.

(2) **How to select, design and optimize effective algorithms for RSDM.** Although some algorithms including probability, D-S theory, Artificial Neural Network (ANN) and others have been used to RS data mining in some studies, that is only some direct application and simple improvement and optimization to conventional algorithms, and the characteristics of RS information such as complexity, uncertainty, fuzziness aren't considered. Those characteristics create some problems to the use of conventional method. On the other hand, RS image is multi-temporal, multi-scale, multi-resolution, multi-platform, so it is difficult to realize RSDM only by current algorithms. For example, multi-temporal RS data requires corresponding incremental and dynamic DM, multi-resolution RS data requires multi-scale DM, and RS data mining with support of GIS requires DM to both vector and raster data structures simultaneously. All those problems require effective algorithms, so it is necessary to design, develop and optimize related algorithms by different methods.

(3) **Knowledge assessment, analysis and management.** RSDM can discover many rules and knowledge. Are all those rules useful or reliable, and how to assess and manage them? How to solve and adjust the contradictions among different rules? All these questions should be answered after DM is done. Only those rules and knowledge that are correct and effective can be used in further tasks. But there are few studies on knowledge assessment, analysis and management at present, and much work should be done on this topic.

(4) **Integration of DM and intelligent information processing.** The ultimate aim of DM is to apply knowledge to image processing and interpretation. In the past DM and image processing are independent and isolated, but there are many links between the two operations. In the future DM should be combined with image processing from some aspects like data preprocessing, image segmentation, feature extraction and so on. So it is necessary to research the approaches of combining RS data mining and intelligent information processing.

(5) **Integration of RS image library, DM tools with spatial data warehouse, digital city and NSDI.** RS data mining is not only an independent tool for image processing, but also a tool or service that can be used and shared by different users on a local area net or internet. That means DM should be incorporated into the framework of digital city, spatial information grid and spatial data infrastructure.

In order to solve those problems and advance the development of RSDM, we would like to give some studies in the following sections.

5.3 The Framework of RSDM

In order to make full use of vast RS image and information in image database, it is necessary to transform data to knowledge that hidden behind original data. RSDM is the integration of DM, SDM and RS image processing. Its goal is to discover the potential, useful and significant knowledge and rules from vast RS images by DM processing and to provide theoretical and technical support to intelligent RS information processing.

Based on the primary framework of Data Mining and properties of RS image processing, applications and image organization, we proposed a framework of RSDM that is demonstrated by Figure.4. The framework divides DM into four steps: preprocessing, data transformation and integration, DM operations, knowledge explanation and assessment, and both data flow and major operations in each step are proposed in this framework.

For users, the processing flow of DM can be illustrated by Figure 5. Users interacts with the Graphic User Interface (GUI) and detail mining procedure will be guided by DM engine, finally the discovered knowledge will be return to users for further uses.

5.3.1 Knowledge that can be discovered: According to our studies, the knowledge that can be discovered from RSIL can be categorized into five types, including spectral knowledge, spatial rules, dynamic (temporal) knowledge, abnormity and synthetic rules.

5.3.1.1 Spectral Knowledge: Spectral signatures of ground object are the basis of RS information acquisition and processing. Much knowledge about a certain ground object like spectral relationship among different bands, effective band combinations, sensitive bands, NDVI and NDBI can be discovered. That knowledge will be used to further image classification and information extraction effectively. Because signature is the major difference between RS image and other images, spectral knowledge is the most important and useful knowledge for intelligent RS information processing.

5.3.1.2 Spatial Rules: As a type of image data with spatial attributes and spatial relationship, RS images can record information about land use and land cover from both spatial dimension and spectral dimension. So DM to RS images can discover complex spatial rules and knowledge similar to SDM such as general geometric knowledge, spatial distribution rules, spatial association rules, spatial aggregation rules, spatial discrimination rules and spatial evolution trends.

5.3.1.3 Dynamic Evolution Knowledge: Dynamic monitoring is an important application field of RS information. By compared to RS information of different times, the change trends can be analyzed from spatial and spectral dimension and further diagnosis and analysis to changes can be investigated and that is useful for decision-making. Temporal mining to RS images can be used to dynamic evolution effectively.

5.3.1.4Abnormity Pattern Extraction : In some applications like crop growth monitoring, disaster analysis and precision agriculture, abnormity pattern extraction is used frequently. Because abnormity discovery and isolated point identification is an important of DM applications, knowledge about abnormity patterns is also useful in RS image data mining.

5.3.1.5Synthetic knowledge and rules about Geo-activities and entities: On many cases, RS images should be processed and analyzed with support of geographic information and background data, especially GIS are playing more and more important roles in RS information processing, and both data in GIS database and spatial analysis functions of GIS are used. So DM to RS images and other auxiliary data is more useful than single DM to RS images because it can reveal the rules between RS image and related geographical information, analyze geo-activities from mechanics, dynamics and more deep level. Especially with the development of RS information model, RS-based Geoscience studies and RS-driven spatial models, intelligent and automatic analysis and applications to those problems all require vast knowledge, and DM can solve them to a great extent. Here, DM to raster and vector data simultaneously is necessary.

5.3.2 Some Key Techniques: Based on the current situation and existed problems of RSDM, it is necessary to give more studies on the following techniques and problems.

5.3.2.1 Establishment and Realization of DM Framework: We will propose a primary framework of RSDM in this paper, but that is only a whole and rough model, and the specific model and procedures should be established based on related requirements and DM operations. Especially, how to integrated DM with RS image data model, metadata, organization and management scheme and properties of RS images is very important.

5.3.2.2 Data Preprocessing: Data preprocessing plays important roles in DM. For RSDM, data preprocessing includes two aspects. One is traditional RS data preprocessing such as geometric correction, registration, filtering and enhancement; and the other is preprocessing used for DM such as data cleaning up, data transformation, data extraction and data integration. The detail tasks of preprocessing in DM mainly include: (1) to retrieve and extract the images for DM; (2) to organize those images used for DM by a given data model; and (3) to process the images, for example, image segmentation, edge extraction, filtering and enhancement, and so on.

The DM task can be assigned by three means generally: (1) DM by given images or data source; (2) DM by given objects; and (3) DM by given features. In the first case, images can be got by metadata and used to further processing, but in the latter two cases, content-based image retrieval is necessary to get the images used to DM. Fast and effective indexing scheme is necessary because of complexity of RS image retrieval, and two commonly used methods are dimensionality reduction and indexing to high-dimensional data. In addition, feature extraction from images is important to further processing and it can be done by image preprocessing.

RS image data model is another key problems. At present three approaches are used in image library: (1) RS images are saved as files; (2) RS images are saved as binary large object in database; and (3) RS images are saved as objects in OODBMS. Data model should be determined when preprocessing is done.

5.3.2.3 DM Modes and Approaches: Based on analysis to RS image library, the following topics should be emphasized in data mining in RS image library: DM to data with uncertainty, parallel, distributed and web data mining, integration of information fusion and data mining, incremental mining, interactive and visual mining, and so on.

- **DM to data with uncertainty**. Uncertainty is one of the important properties of RS image. Although some measurements may be used to reduce and eliminate uncertainty before images are input into image library, it is necessary to take uncertainty into account in data mining. In general two approaches can be used, one is to do some specific preprocessing, and the other is to design robust DM algorithms oriented to uncertain and fuzzy data and avoid generating wrong and insincere knowledge and rules because of uncertainty in original data.

- **Parallel, distributed and web Data Mining**. With the development and establishment of National Spatial Data Infrastructure (NSDI), RS image library may be composed of distributed database on network, so it is necessary to do distributed and web Data Mining. In addition, parallel DM is very important to vast volume RS data.

- **Integration of information fusion and DM**. Fusion is an important approach in multi-source, multi-scale and multi-temporal RS information processing, and it can play important roles in RS data mining also. The relationship between information fusion and data mining can be analyzed from three levels. The first is fusion used as preprocessing mean of DM, in which information fusion is done at first and then DM to fused image is dome. Because fusion information is used to DM some important features and information may be neglected. The second is fusion as a simple DM technique or as an important step in DM. The third level is the whole integration of information fusion and DM, in which both fusion-based DM algorithms and mining-based information fusion algorithms should be put forward.

- **Incremental Mining**. Unlike general database with stable data, data in image library increases quickly, and that will bring forward new demands to DM. Incremental DM will combine DM with image increasing, adjust and conduct data mining procedure with accumulation of images. On one hand, knowledge will be renewed and discovered knowledge before will be corrected and improved; on the other hand, some new knowledge many be discovered with data increasing.

- **Interactive and visual data mining.** Man-machine interaction can optimize processing procedure and enhance efficiency; especially for vast RS information mining, interactive and visual operations are very important to avoid and reduce appearance of unreliable and uncertain knowledge.

5.3.2.4 Knowledge Representation, Assessment and Management: Many rules and pattern can be discovered by RSDM, and the further key problem is how to express, assess, manage and use those rules and knowledge, especially how to assure the reliability and applicability of discovered knowledge and avoid wrong knowledge is very important.

5.4 Some Applicable Algorithms

RS image DM is not only the simple applications of DM techniques to RS images, but has more complicated and profound theoretical, technical and application framework. In order to meet the demands of RSDM, it is necessary to put forward some new algorithms. In addition, the data sets used to RSDM include single-source RS images, multi-source and multi-scale images, multi-temporal images, combination of RS images and background information, web image library and others, so it is very important to analyze every case and design corresponding algorithms. At present some algorithms that can be used to DM in RS image library include image processing and pattern recognition, clustering, visualization, fuzzy theory, rough set theory, D-S theory and so on, in addition associate rule mining, decision tree and spatial analysis can be used to the integrated mining of RS image and background information.

5.4.1 Clustering: Clustering can segment RS image information into different kinds from spatial dimension by a given segmentation algorithm and similarity measure approach, and then those hidden rules and knowledge can be discovered by further processing to each kind. In general clustering should be combined with other algorithms such as inductive learning, statistics and decision tree because clustering can only realize aggregation to original data and properties, features of each class should be used to discover detail rules.

5.4.2 Artificial Neural Network (ANN): Some Artificial Neural Network including BPNN, RBF, ART and Hopfield neural network have been used to DM, also ANN got wide applications in RS image classification, and their integration will promote the development of ANN used to RSDM. For vast RS information, ANN has its advantages over traditional methods. Taking application of BPNN as an example, after samples of each class are selected and input into BPNN, the learning procedure won't stop until the ending condition is satisfied. After computation ends, weight among neural units in adjacent layers can be explained according to practical details and used as knowledge to further image processing.

5.4.3 Association Rules: As a widely used DM approach, association rule mining can be used to RSDM conveniently. On one hand, some traditional algorithms such as Apriori, Sampling and DIC algorithms can be used directly or by some improvements; on the other hand, some new algorithms should be proposed according to characteristics of RS information. Association rules in RS images include association between different entities on image, association between different bands of a type of object, association between different features such as spectral feature and texture of a kind of object and association between properties of ground object and its features on image. After related association rules are discovered, it is very significant for further processes such as image classification, RS inversion, and computation of biological, chemical and physical parameters.

5.4.4 Decision Tree: Decision tree is used to RS image classification and information extraction effectively. But how to establish a decision tree is a very important problem. Nowadays the commonly used method is to establish it by analysis and experience of domain experts, and that is very complicated and time-consuming, and sometimes not reliable and applicable, so more effective and reliable methods should be put forward. DM can be used to this task. The traditional decision tree learning algorithms including CLS and ID3 can be used to establish decision tree from RS images.

Many traditional methods like pattern recognition, statistics, evidence theory, spatial analysis, probability theory and some novel theories like rough sets, fuzzy theory, genetic algorithm (GA), support vector machine (SVM) and cloud theory can be used to RSDM, but more work should be done on their realizations and applications.

6. CONCLUSIONS AND PROSPECTS

Although high-resolution satellite RS images have provided geographical information updating with powerful and abundant information sources, it is necessary to give further studies to related theories, methods and software systems, especially the intersection and integration of multi-disciplinary theories and techniques are required. Oriented to the requirements of updating geographic information in SDI using high resolution RS images, a special software system and its processing flow is proposed in this paper, and some key topics including content-based RS image retrieval, change detect, information fusion and RS Data Mining and Knowledge Discovery are discussed in this paper. It is only a primary study on the topic, and the further studies will be focused on the implementation and application of this system, and the design and experiment of related algorithms, so the updating of geographic information by RS image will be realized, and the SDIs on different levels will be updated rapidly and they will promote the applications of RS information.

REFERENCES

Arnold W.M. Smeulders, Marcel Worring, Simone Santini et al. 2000, Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions On Pattern Analysis And Machine Intelligence. 22(12), pp 1349-1380.

Bo XiaoChen, Liu Jianping. 1999,Color histogram based on image retrieval. Journal of Image and Graphics, 4(1), pp: 33-37.

Datcu, M.; Seidel, K. 1999, Query by image content and information mining. IEEE Geoscience and Remote Sensing Symposium, IGARSS '99 Proceedings. Vol.2, pp 1335 –1337.

Dobson J.E. Commentary: 1993, A conceptual framework for integrating RS, GIS and geography. PE&RS, 59(10), pp:1491-1496.

Gong Jianya. F 2001, undamentals of Geographical Information System. Science Publishing House.

Huang Xianglin, Shen Lansun. 2002,Research on content-based image retrieval techniques. Acta Electronica Sinica, 30(7), pp: 1065-1071.

John G.Lyon, Ding Yuan, Ross S.Lunetta et al. 1998, A change detection experiment using vegetation indices. PE&RS, 64(2), pp:143-150.

Kruse.F. A. and A. B. Lefkoff. Analysis of Spectral Data of Manmade Materials, Military Targets, and Background Using an Expert System Based Approach. www.google.com.

Kwarteng .A.Y and P.S.chavez. 1998, Change detection study of Kuwait city and environments using multi-temporal landsat TM data. Int. J. of RS, 19(9), pp: 1651-1662.

Lambin E F. 1994, Change-vector analysis in multitemporal space: a tool to detect and categorize land-cover change processes using high temporal resolution satellite data. Remote sensing of environment, 48(2), pp:231-244.

Li Xiangyang, Zhuang Yueting and Pan Yunhe. 2001, The technique and system of content-based image retrieval. Journal of Computer Research and Development. 38(3), pp: 344-354.

Liu Zhongwei, Zhang Yujin. 2000, A comparative and analysis study of ten color feature-based image retrieval algorithms. Signal Processing, 16(1), pp:79-84.

Luo Rui, Zhang Yongsheng and Fan YongHong. 2002,Research on content-based image retrieval in remote sensing image library. Journal of Remote Sensing. 6(1), pp: 24-29.

Marchisio, G.B.; Wen-Hao Li; Sannella, M.; et al. 1998, GeoBrowse: an integrated environment for satellite image retrieval and mining. IEEE Geoscience and Remote Sensing Symposium Proceedings, IGARSS '98. Vol.2, pp 669 –673.

MAS .J.F 1999, Monitoring land-cover changes: a comparison of change detection techniques. Int. J. of RS, 20(1), pp: 139-152.

Merrill K.Ricld and Jiajun Liu. 1998, A comparison of four algorithms for change detection in an urban environment. Remote sensing of environment, 63(2),pp: 95-100.

Prakash.A and R.P.Gupta. 1998, Land-use mapping change detection in coal mining area——a case study in the Tharia, Coalfield, India. Int. J. of RS, 19(3), pp:391-410.

Pu Ruiliang, Gong Peng. 2000,Hyperspectral Remote Sensing and its applications. Higher Education Press.

Wang Huifeng, Sun Zhengxing, Wang Jian. 2002, Semantic image retrieval: review and research. Journal of Computer Research and Development, 39(5), pp: 513-523.

Wang Jinnian, Zhang Bing, Liu Jiangui et al. 1999, Hyperspectral data mining-toward target recognition and classification. Journal of Image and Graphics, 4(11), pp:957-964.

Yao Yurong, Zhang Yujin. 2000,Shape-based image retrieval using wavelet and moment. Journal of Image and Graphics, 5(3), pp: 206-210.

Zha Yong, Ni Shaoxiang and Yang Shan. 2003,An effective approach to automatically extract urban land use from TM image. Journal of Remote Sensing, 7(1), pp: 37-41.

Zhang Yujin. 2003, Content-based visual information retrieval. Science Press.

Zhou Chenghu, Luo Jiancheng, Yang Xiaomei et al. 1999, Remote sensing image Geo-understanding and analysis. Science Press.

ACKONWLEDGEMENT

The research is supported by the National Natural Science Foundation of China (No.40401038), the Opening Research Fund Program of the Key Laboratory of Geomatics and Digital Technology, Shandong Province, China (No 2003-03) and the Science Foundation of China University of Mining and Technology (grant No. D 200403)

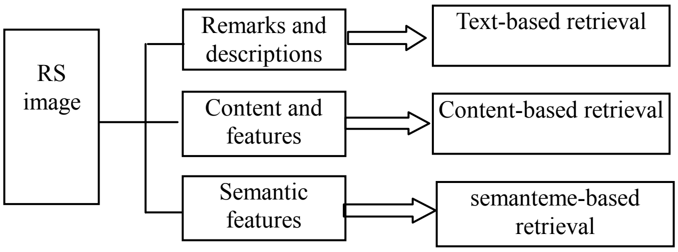

Figure 1. Three levels of RS image retrieval

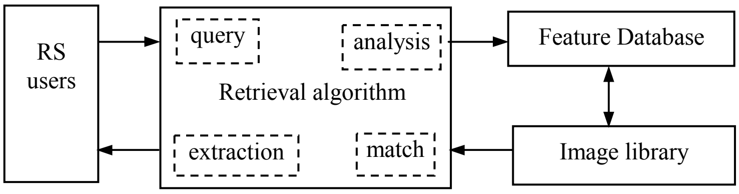

Figure 2. The framework of content-based RS image retrieval

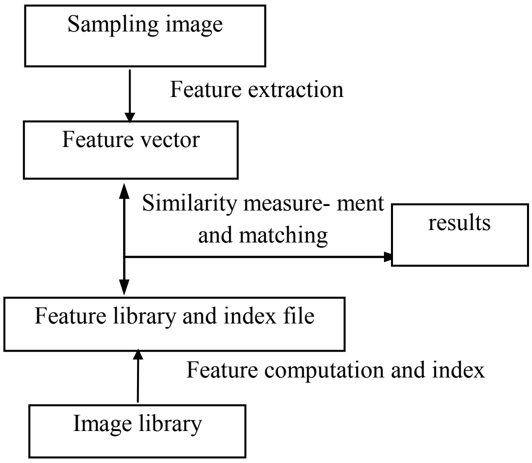
Sampling image
Feature extraction
Feature vector
Similarity measure- ment and matching
results
Feature library and index file
Feature computation and index
Image library

Figure 3. A typical processing flow of content-based RS image retrieval

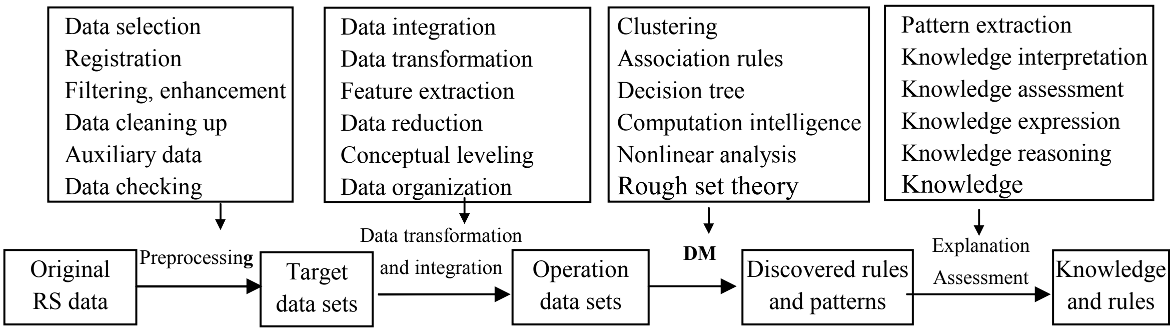

Figure 4. Framework of RSDS

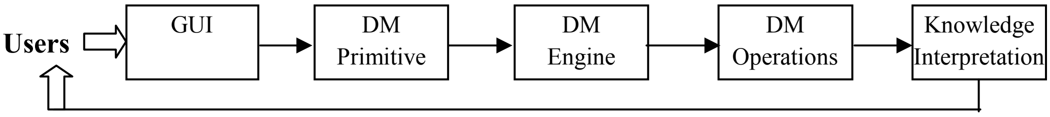

Figure 5. The processing flow of RSDM