DATA MINING AND ITS APPLICATION IN DATABASE CONTENT REFINEMENT

ZHAI Liang[a, b *], TANG Xinming[b, c], WU Lan[d], LI Lin[a], WANG Zhongyuan[a]

[a] School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan, 430079, China - :
zhailiang@126.com, lilin@telecarto.com, wzy95002@163.com
[b] Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping, Chinese Academy of Surveying and Mapping,
16 Beitaiping Road, Haidian District, Beijing, 100039, China -: tang@casm.ac.cn
[c] Institute of Remote Sensing and Geographic Information Systems, Peking University, Beijing, 100871, China
[d] State Bureau of Surveying and Mapping, 9 Sanlihe Road, Haidian District Beijing, 100830, China -: wulan@sbsm.gov.cn

KEYWARDS: Data mining, NFGIS 1:50,000 DLG database, Content Refinement, Clustering Analysis

ABSTRACT:

At present, China has established a series of National Fundamental Geographical Information System (NFGIS) databases which provide a united and authoritative space platform for GIS user communities and play an important role in social development, economy and state safety. However, with the progress in national informatization, NFGIS has been confronted many new requirements from GIS users including its refinement and updating. The thesis mainly concerns content refinement for National Fundamental Geographical Information System (NFGIS) 1:50,000 DLG database by virtue of statistical data mining technology. Firstly data mining methods are discussed, including statistical analysis, generalization-based mining, fuzzy sets methods, rough set theory and cloud theory, etc. Secondly, we propose a methodology employing clustering strategy in database content refinement. This approach is very suitable to explore the survey data and get useful information. Through conducting the user survey which means to collect users' requirements on NFGIS database, such as data content and attributes information, we gathered the survey results and analyzed further by adopting clustering analysis. Generally speaking, there are two categories of methods for clustering: partitioning algorithms and hierarchical algorithms. Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating samples between groups until some equilibrium is attained. In contrast, hierarchical algorithms proceed by combining or dividing existing groups, producing a hierarchical structure displaying the order in which groups are merged or divided. Users from different sectors have different requirements, while we can hypothesize that there are 3 clusters representing different requirement degree respectively: high, medium and low. Thus we take partitioning algorithm. Surely, we may not specify the clustering number and adopt the hierarchical clustering algorithm. We have discussed these two methods and their clustering results. Finally, we got the content refinement plan of 1:50,000 DLG database. Throughout the research, we can reach the following conclusions: (1) Clustering analysis is essential to survey results; (2) Spatial analysis is supplementary to clustering analysis; (3) Using different clustering methods ensures the correctness of clustering results.

1. INTRODUCTION

The term data mining has been mostly used by statisticians, data analysis, and the management information systems (MIS) communities (Usama Fayyad *et al*, 1997). Data mining is a step in the KDD (knowledge discovery in databases) process consisting of applying computational techniques that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (U.Fayyad *et al*, 1996). And it is emerging as a new active area of research which combines methods and tools from the fields of statistics, machine leaning, database management and data visualization (A. Feelders et al, 2000). Data mining techniques have been applied to many real-life applications, and new applications continue to drive research in the area. Many statistical models exist for explaining relationships in a data set or for making predictions: cluster analysis, discriminant analysis and nonparametric regression can be used in many data mining problems (Jonathan R.M.Hosking *et al*, 1997). In this paper, we propose a methodology employing clustering strategy in database content refinement. This approach is very suitable for exploring the survey data and get useful information.

The remainder of the paper is organized as follows. Section 2 gives an overview on data mining techniques where different methods and techniques are discussed. Section 3 presents the application: content refinement for National Fundamental Geographical Information System (NFGIS) databases and shows the clustering algorithm used in the user survey results. Section 4 presents the refinement plan and discusses some further analysis. Section 5 concludes the presented work.

2 DATA MINING TECHNIQUES

Many different methods have been used to perform data mining tasks. These techniques not only require specific types of data structures, but also imply certain types of algorithmic approaches (Margaret H. Dunham, 2003). In the following, we category and describe some typical data mining techniques.

2.1 Statistical Analysis

Statistics is used as the most common approach for analyzing categorical or quantitative data and statistical analysis is a well studied area where exist a large number of algorithms including various optimization techniques. It handles numerical data well

[*] Corresponding author: ZHAI Liang, PhD candidate. His research interests include fundamental GIS, spatio-temporal database and data mining etc.

and usually comes up with realistic models of spatial phenomena. However, it is a kind of technique that can only be used by the experts with a fair amount of domain knowledge and statistical expertise (Jiang Liangxiao *et al*, 2003).

Clustering analysis is a kind of statistical analysis and used to identify clusters embedded in the data, where a cluster is a collection of data objects that are 'similar' to one another. It can be expressed by distance functions, specified by users or experts. A good clustering method produces high quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high. For example, one may cluster the parcels according to their land use, cover, soil type, ownership and geographical locations. Section 3.1.2 will further illustrate this method.

2.2 Generalization-based Mining

Data and objects in databases often contain detailed information at primitive concept levels (Ming-Syan Chen *et al*, 1996). It is often desirable to summarize a large set of data and present it at a high concept level. For example, one may like to summarize the detailed traffic information at different time in a day and shows its general traffic pattern. This requires generalization-based mining, which first abstracts a large set of relevant data from a low concept level relatively high ones and then perform knowledge extraction on the generalized data.

2.3 Fuzzy Sets Method

Fuzzy sets have been used in many computer science and database areas. In the classification problem, all records in a database are assigned to one of the predefined classification areas. A common approach to solving classification problem is to assign a set membership function to each record for each class. The record is then assigned to the class that has the highest membership function value. Similarly, fuzzy sets may be used to describe other data mining functions. Association rules are generated given a confidence value that indicates the degree to which it holds in the entire database. This can be thought of as a membership function (Margaret H. Dunham, 2003).

2.4 Rough Set Method

Rough set theory was first proposed by the Polish scientist Z.Pawlak in 1982. It is a kind of approach to knowledge-based decision support and has been widely used in uncertain information classification and knowledge discovery. And of course, it can be applied in data mining. It provides a new way to attribute information analysis in GIS, like attribute consistency, importance, dependence and classification.

2.5 Cloud Theory

Cloud theory is a new theory in dealing with uncertainties consists of cloud model, reasoning under uncertainty and cloud transform. This theory combines the randomness and fuzziness, so it compensates the inhere limitation of membership function which is the basis for rough set theory and makes it available to join quality and quantity together in data mining.

Besides those above, there are many other methods used in data mining, like raster based analysis, decision tree, genetic algorithms, visual data mining modelling and artificial neural networks, etc. These methods have always been used together in knowledge discovery.

3. CASE STUDY: DATABASE CONTENT REFINEMENT

Since the 90th last century, China has established a series of NFGIS databases which provide a united and authoritative space platform for GIS user communities and play an important role in social development, economy and state safety. These databases include Digital Linear Graphics (DLG) database, Digital Elevation Model (DEM) database, Digital Orthography Model (DOM) database, Digital Raster Graphics (DRG) database, etc. See Figure1. Together with the provincial databases, they have made up of the national fundamental geo-spatial framework (Li Jingwei, 2003; Wang Donghua, 2003).

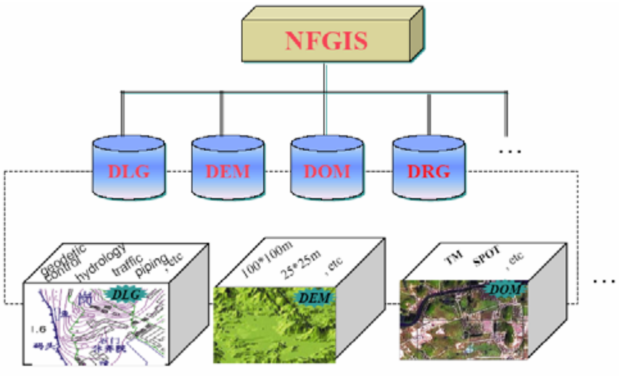

Figure1. NFGIS databases

However, with the rapidly progress made in economy, many geo-spatial information users have promoted many new and higher demands on national fundamental geo-spatial information and most of these are concerning NFGIS databases refinement, which should answer these questions beforehand, such as which kind of dataset or database is most valuable to the users. For example: what kinds of dataset or database are being used? Who are the users or potential ones (Steven M.Frank, 1995)? And, are the current databases satisfied to the users? What features (attributes) should the database include? Among these questions, the last one is of great importance. We adopt the methodology of user survey with questionnaires to fulfil these aims.

This survey aims at gaining a thorough understanding of the national fundamental geo-spatial data application in sectors such as hydrology, agriculture, forest, transportation, economic statistics, land, marine, environment, meteorology, cultural relic, civil, scientific surveying, geology, mine, seismology, and surveying and mapping, etc, and identifying the features and attributes mostly needed by them. The questionnaire is designed according to the National 1:50,000 Relief Map Specification (China). Features having great significance, like boundary, double lane railway, perennial river, important building, etc. are not involved in the questionnaire in order to decrease the number of problems. The respondents are only needed to check what they need.

The questionnaire generally including two parts: The 1st part is the background information, including some common questions, such as the respondent's name, organization, contact information; purpose of using national fundamental geo-spatial information; data precision and updating. The 2nd part includes geodetic control, hydrology, residential area and building, traffic, piping, relief and soil texture and vegetation. Here are two examples:

(1) Factory Building includes those features:

□transformer substation

□sewage treatment works

□industrial well (oil/gas/salt…)

□liquid/gas storage equipment

□tower building (distill tower, chimney, water tower, watch tower…)

□strip mine, excavate field

□saltern, *please select what you need.*

(2) Please select the attribute of swamp that you or your organization has used. If the given choices are not complete, please add.

□nonuse

□passable / not passable

□water depth

□depth of ooze layer

□undefined boundary

□others_______________

3.1 Survey Results Summary

This survey employs two kinds of strategies: Focus Group and Web-Survey (Ke Huixin, 2001). The questionnaire has been issued on the Website: http://lab.casm.ac.cn/ and http://www.csi.gov.cn/ investigate/gisurvey.asp. From August, 1st to October, 1st, 2004, we have received 295 feedback and 274 valid. Response rate is 93%.

In this questionnaire, there are 108 features and attributes (each feature or attribute is regarded as a case or sample). Table 1 is the summary of data content response.

Residential Area and Building	**Feature**	**Agree**	**Total**
House	Shed	122	205
	Cave	60	205
	Mongolian Tent	47	205
Population	Population	121	205

Table 1 Summary of the Data Content Response (part)

Through summing up the responses, we can discover that users from different departments have different requirements and application of fundamental geo-spatial data and draw some qualitative conclusions. For example, Geological sectors use relief maps in geology investigations and filed work, and they hope that the NFGIS 1:50,000 DLG database could include all relief features and attributes. They are in great need of independent features for the purpose of across river, looking for water, residence, identifying orientation, like ‘footbridge’, ‘ferry’, ‘spring’ and ‘well’; Departments of hydrology use relief maps in water resources planning and drainage area measurement, so they hope the database could contain all kinds of rivers and water resource facilities including banks and dams. At the same time, transformer substation and power line are needed for the sake of electricity; as to seismic departments, they use geo-spatial information in identifying seismic centre, fault location and monitoring station. So they need residential area with population, hydrology, railway, highway, piping, power line, spring (especial ascending spring), earth flow, earth slide and volcano; City and country planning departments concerns more about road, contour, residential area, river, power line and land us, etc.

3.2 Survey Results Analysis

3.2.1 Selection of Clustering Methods

In order to make the NFGIS 1:50,000 database content refinement plan, summing up the responses and getting some qualitative conclusions are far from enough, since each department hopes that the database could include the features or attributes having close relation with their daily work and specialty. Therefore, we should take the users’ needs into account as a whole and select features or attributes according to requirement degree. One alternative is clustering method: features and attributes having same degree may be clustered.

Clustering analysis is the searching for groups (clusters) in the data, in such a way that samples belonging to the same cluster resemble each other, whereas samples in different clusters are dissimilar. Generally speaking, there are two categories of methods for clustering (Wulan, 2002; Zhang Guojiang, 2002; Theodore P. Beauchaine, 2002):

(1) Partitioning Algorithms. A partitioning algorithm describes a method that divides the data set into k clusters, where the integer k needs to be specified by the user. Typically, the user runs the algorithm for a range of k-values. Algorithms of this type include k-means, partition around medoids, fuzzy clustering etc. K-means algorithm is one of partitioning algorithms.

(2) Hierarchical Algorithms. A hierarchical algorithm describes a method yielding an entire hierarchy of clusters for the given data set. Agglomerative methods start with the situation where each object in the data set forms its own little cluster, and then successively merges clusters until only one large cluster remains which is the whole data set. Divisive methods start by considering the whole data set as one cluster, and then split up clusters until each object is separate.

The difference between these two algorithms is the following: Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating samples between groups until some equilibrium is attained. In contrast, hierarchical algorithms proceed by combining or dividing existing groups, producing a hierarchical structure displaying the order in which groups are merged or divided.

Users from different sectors have different requirements, while we can hypothesize that there are 3 clusters representing different requirement degree respectively: high, medium and low. Thus we take partitioning algorithm.

Surely, we may not specify the clustering number and adopt the hierarchical clustering algorithm. In the following, we will discuss the two methods and their clustering results.

3.2.2 Distribution of Survey Data

Figure2 shows the distribution of the survey results after normalization. It‘s the distribution grouped by the feature class. There are 7 feature classes, each of which contains several “features and attributes”. The value of vertical axis denotes the requirement degree: the percentage of the ‘Agree / Total’ of each feature. And the value of horizontal axis denotes ‘geodetic control’, ‘hydrology’, ‘residential area and building’, ‘traffic’, ‘piping’, ‘relief and soil texture’ and ‘vegetation’ respectively.

For example, in geodetic control, there are ‘monument point’ and ‘astronomical point’, the former has higher requirement degree.

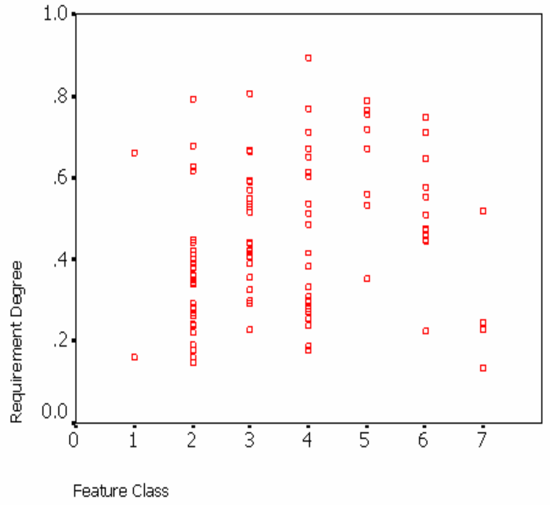

Figure2. Data Distribution Based On the Feature Class

3.3 Clustering Analysis

3.3.1 Clustering on Survey Data: Based on Partitioning Algorithm

Among all partitioning algorithms, k-means algorithm is most widely used. It is applied in distinguishing relative homogeneous sample set. Since the sample set has some similarity in this survey---the samples are all geographical features and they have close characters. So k-means algorithm is applied here. In (1), it partitions N data points into K disjoint subsets (clusters) R_i containing N_j data points so as to minimize the sum-of-squares criterion.

$$J = \sum_{i=1}^{k} \sum_{\overline{X}_j \in R_i} \| \overline{X}_j - \overline{W}_i \|^2 \qquad (1)$$

Where $\overline{X}_j$ is a vector representing the jth data point and $\overline{W}_i$ is the geometric centroid of the data points in R_i. This algorithm initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The geometric centroid's position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

SPSS for windows software（version 10.0，SPSS Inc.）is employed here for analysis. At the beginning, the SPSS software selects the centres (geometric centroids) randomly. And after iteration process the centres are changed. Table 2 shows the centres after clustering process. The values of centres represent the means of the samples in every class. According to the values of centres, we can easily find that cluster 1, 3, 2 correspond to high, medium, low requirement degree respectively.

Table 3 illustrates the Euclidean distances between final cluster centres. The distance between cluster 1 and cluster 2 is bigger than the one between cluster 1 and cluster 3, because cluster 1 and cluster 2 represent high and low requirement degree respectively. And from the semantic view, ‘high’ and ‘low’ have a big difference.

Table 4 tells the number of cases (samples) in each final cluster.

	Cluster		
	1	2	3
Agree Percent	.6925	.2573	.4637

Table 2 Final Cluster Centres

Cluster	1	2	3
1		.616	.324
2	.616		.292
3	.324	.292	

Table 3 Distances between Final Cluster Centres

Cluster	1	26
	2	41
	3	41
Valid		108

Table 4 Number of Cases in each Cluster

We need further analysis to see whether cluster number—3 is suitable. If not, we would have to cluster again. Through analyzing distances of cases from its classification cluster centre, we can tell whether the cluster number—3 is suitable. If many cases were seriously out of centre, then the cluster number is not suitable. Figure3 is a box diagram: the black bold line in the centre represents the mean; the rectangle box is the bound of interquarti1e range. We can find that all cases are not out of centre seriously. To some degree, this indicates that cluster number—3 is suitable.

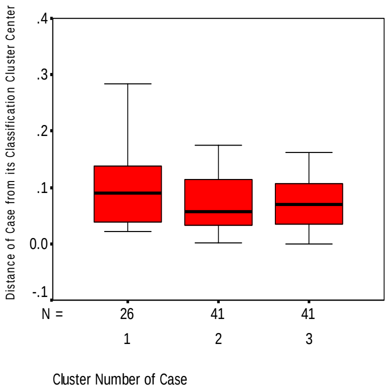

Figure3. Distance of Case from its Classification Cluster Centre

3.3.2 Clustering on Survey Data: Based on Hierarchical Algorithm

The basic process of hierarchical clustering:

Step 1: Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.

Step 2: Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
Step 3: Compute distances (similarities) between the new cluster and each of the old clusters.
Repeat steps 2 and 3 until all items are clustered into a single cluster.
Step 3 can be done in different ways, which is what distinguishes average-link clustering from single-link and complete-link.

The following is the details of average-link, single-link and complete-link clustering.

- Average-link clustering:

$$d(R,Q)=\frac{1}{|R||Q|}\sum_{i\in R,j\in Q} d(i,j) \qquad (2)$$

In (2), R and Q represent two different sets, d(i, j) is the distance between the two clusters. It satisfies d (i, i) = 0; d (i, j) ⩾ 0; d (i, j) = d (j, i). The following is same.

- Single-link clustering:

$$d(R,Q)=\min_{i\in R,j\in Q} d(j,j) \qquad (3)$$

- Complete-link clustering:

$$d(R,Q)=\max_{i\in R,j\in Q} d(j,j) \qquad (4)$$

We adopt average-link clustering and get the dendrogram, see Figure4. From it, we can easily find that: if the clustering number is 4, then there would be one single sample (alley) in a class. This is unreasonable. So clustering number should no more 4. If it were 2, then there would be 2 types of semantics: need and not need, which is also unreasonable. So clustering number should be 3.

4. CONTENT REFINEMENT PLAN

4.1 Plan 1: Based on Partitioning Algorithm

4.1.1 Alternatives of Different Selections

All of the cases (108) belong to different requirement degree: high, medium and low.

Alternative 1: contains the ‘high’ (requirement degree) features only, 26 samples;

Alternative 2: contains the ‘high’ and ‘medium’ features, 67 samples;

Alternative 3: contains all the features, 108 samples.

Three alternatives are obtained through the above clustering. At the same time, we noticed that many features or attributes have been omitted for the simplicity of the questionnaire. So we should add such features or attributes.

4.1.2 Comparison

Alternative 1 is the intersection of the 3 alternatives, including the minimum features and attributes. And they are the most important features and attributes which can be used as the geographic references.

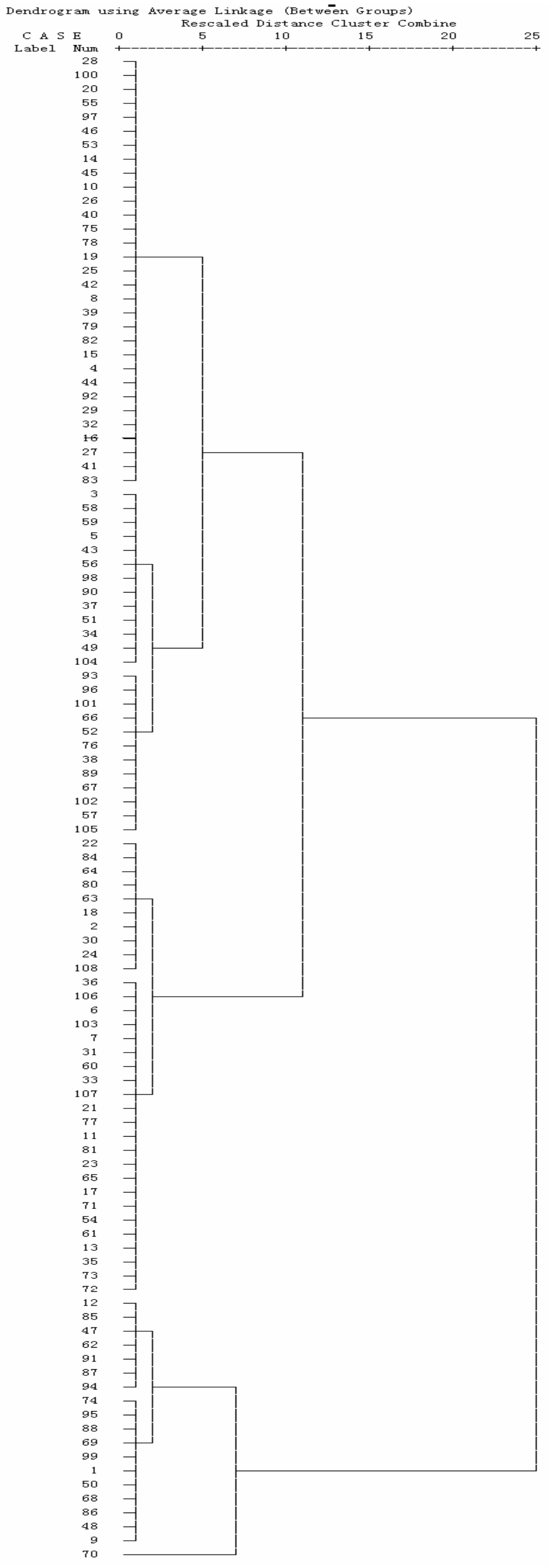

Figure4. Dendrogram Using Average Linkage between Groups

Alternative 3 is the most detailed, but it is not fit for the database refinement plan. As discussed above, each department hopes the database could include the features or attributes having close relation with their daily work and specialty. It is impossible to satisfy the needs of each department for the sake of economy in China, because the data collection, updating, distribution and maintenance will cost enormously. It is recommended that alternative 3 be used as the reference for national relief data classification.

The content in alternative 2 can satisfy the needs of geology, hydrology, transportation and seismology departments by and large. For example, geology departments need 'footbridge', 'ferry', 'spring' and 'well'; seismic departments need residential area with population, hydrology, railway, highway, piping, power line, spring (especial ascending spring), earth flow, earth slide and volcano; City and country planning departments concerns more about road, contour, residential area, river, power line and land us, etc. Those requirements are all reflected in alternative 2.

4.1.3 Further Analysis

Through the comparison above, alternative 2 appears to be the basis for NFGIS 1:50,000 database content refinement. However, clustering analysis is conducted according to users' requirement degree on each feature or attribute and the integrity and relevancy among features or attributes are neglected. The rationality of alternative 2 should be further discussed in virtue of spatial analysis.

Spatial analysis is statistical description or explanation of either locational or attribute information or both (Guo Renzhong, 2001; Wang Jiayao, 2001). Further analysis involves three facets including integrity of features, logical consistency and attribute complexity consistency.

(1) Integrity of features analysis

Some features and attributes are both rejected in alternative 1 and 2. But a few important features should be added, such as 'reef' which has ownership; fixed or seasonal Mongolian tents are significant to civil administration or planning departments. They all should be added.

(2) Logical consistency of features analysis

'Altitude annotation' and 'water depth annotation' have the same importance, but in alternative 2 'water depth annotation' is missed. It should be added. If not, the database would include 'altitude annotation' without 'water depth annotation'. This could cause logical confusion.

(3) Attribute complexity consistency analysis

As to the attributes of 'river', 'high watermark' and 'water line' are always used together; as to 'dam/bank', 'safety line of flood control' and 'warning line' are also used together; but in alternative 2, they were missed. Similarly, the attribute of 'highway'—'constructional materials' was missed. Since these missing attributes information may break attribute complexity consistency and they are not difficult to collect, they should be added.

4.2 Plan 2: Based on Hierarchical Algorithm

Similarly, we can find that cluster 3, 1, 2 correspond to high, medium, low requirement degree respectively and get 3 alternatives:

Alternative 1: The plan contains the 'high' (requirement degree) features only, 19 samples;

Alternative 2: The plan contains the 'high' and 'medium' features, 75 samples;

Alternative 3: The plan contains all the features, 108 samples.

4.3 Discussion

Through comparing plan 1 and plan 2, we detect that there are 13 features or attributes(water level, high water level, river width/depth, length of dam or bank, shanty, population, etc) having different clustering results. After a discussion from experts and a careful comparison, we recommend the alternative 2 in plan 1 with some modification for the final content refinement plan. The details of it can be found in the NFGIS Database Refinement Report.

5. CONCLUSIONS

In this paper, we have talked about data mining approaches in details and applied clustering algorithm in NFGIS database content refinement. Throughout the research, we can reach the following conclusions:

(1) Clustering analysis is essential to survey results. How can we get the useful information from the survey results? It is not enough to summarize it only and get some qualitative conclusions. We need further analysis by adopting data mining method: clustering analysis.

(2) Spatial analysis is supplementary to clustering analysis. Clustering analysis is conducted according to users' requirement degree on each feature or attribute, so the integrity and relevancy among features or attributes are neglected. The rationality of the alternatives should be further discussed in virtue of spatial analysis including integrity of features analysis, logical consistency analysis and attribute complexity consistency analysis. After spatial analysis, more features and attributes are added.

(3) Using different clustering methods ensures the correctness of clustering results. In the above, we adopt both partitioning and hierarchical clustering algorithm, which to some degree helps us find the correct and reasonable clustering results.

Furthermore, we can integrate application model and establish the testing database according to the proposed content refinement plan.

ACKNOWLEDGEMENTS

The authors thank Miss GAO Xiaoming (Chinese Academy of Surveying and Mapping, China) for her help in collecting questionnaire feedbacks. Financial support for this research was provided by the National Surveying and Mapping Fund (No. 1469990324221).

REFERENCES

A. Feelders, H. Daniels, M. Holsheimer, 2000. Methodological and practical aspects of data mining. Information & Management, 37, pp. 271-281

Guo Renzhong, 2001. Spatial Analysis. Beijing: Higher Education Press , pp. 4-5

Jiang Liangxiao, Cai Zhihua, 2003. Review and Prospect of Spatial Data Mining. Computer Engineering, 29, pp. 9-10

Jonathan R.M.Hosking, Edwin P.D.Pednault, Madhu Sudan, 1997. A statistical Perspective on Data Mining. Future Generation Computer Systems,13, pp. 117-134

Ke Huixin, 2001. Web-Survey Methodologies. Modern Media, 4, pp. 80-84

Li Jingwei, 2003. NFGIS 1: 50,000 Database Establishment and Integration. in Digital China Geo-spatial Fundamental Framework, edited by Chen Jun and Wu Lun. Beijing: Science Press, pp. 162-168

Margaret H. Dunham, 2003. Data Mining Introductory and Advanced Topics. Beijing: TsingHua University Press, pp. 23-24

Ming-Syan Chen, Jiawei Han, Philip S. Yu, 1996. Data Mining: An Overview from a Database Perspective. IEEE Transaction on Knowledge and Data Engineering, 8(6) , pp. 869-883

Steven M.Frank, Michael F.GoodChild, Harlan J.Onstrud, Jeffrey K.Pinto, 1995. Framework Data Sets for the NSDI. http://www.ncgia.ucsb.edu/Publications/Tech_Reports/95/95-1.pdf (accessed 20 May. 2003)

Theodore P. Beauchaine, Robert J. Beauchaine III, 2002. A Comparison of Maximum Covariance and K-Means Cluster Analysis in Classifying Cases into Known Taxon Groups. Psychological Methods, 7, pp. 245-261

U.Fayyad, G.Piatetsky-Shapiro, P.Smyth, R.Uthurusamy, 1996. Advances in Knowledge Discovery and Data Mining. Massachusetts: MIT Press, pp.2-4

Usama Fayyad, Paul Stolorz, 1997. Data mining and KDD: Promise and Challenges. Future Generation Computer Systems, 13, pp. 99-115

Wang Donghua, 2003. NFGIS 1: 50,000 DEM Database Establishment. in Digital China Geo-spatial Fundamental Framework, edited by Chen Jun and Wu Lun. Beijing: Science Press, pp. 148-161

Wang Jiayao, 2001. Spatial Information System Theory. Beijing: Science Press, pp. 262-264

Wulan, 2002. Methodology for Selection of Framework Data-Case Study for NSDI in China (Master Thesis), ITC , pp. 10-15

Zhang Guojiang, 2002. Applications of Soft Computing and Data Mining in Electrical Load Forecasting (Doctor Dissertation), Zhejiang University, pp. 37-42