QUALITATIVE THE SEMATIC FACTORS ON MAP FOR SIMILARITY SPATIAL SEARCHING BASED ON SPATIAL DESCRIPTION

KHƯU Minh Cảnh[1] and LÊ Trung Chơn[2]

Ho Chi Minh University of Technology, Vietnam
[1]Email: kmcanh@yahoo.com
[2]Email: ltchon@hcmut.edu.vn

ABSTRACT

This paper presents some analyses on the sematic factors of the map for similarity searching. The problem is related to cartographic generalization techniques and spatial description. We present qualitative analysis the effects of map making methods based on cartographic generalization. The analysis focuses on the altering of the map and specific the map generalization tools. On the other hand, we plan to build the concept model for simalarity searching which will help to recognize the sub map in a map. The results will be used to improving and orienting the map similarity searching techniques.

1. INTRODUCTION

GIS is a system for acquire, store, analysis, manage and view the geographic. Similar spatial searching is one of the spatial analysis problem. Based on set of spatial descriptions, it is the problem to find the geometric similarity between the set of spatial objects. In the field of regconition, many methods have been developed for the geometric indentical problem. The main thing is how to choice the measurement standards for the comparison. And the results will be the set of spatial objects having similar shape with the input criteria. In spatial data, these criteria will be the set of geometric relations which extracted from the input and will be the keys for the map generalization processes and parallelization all of them.

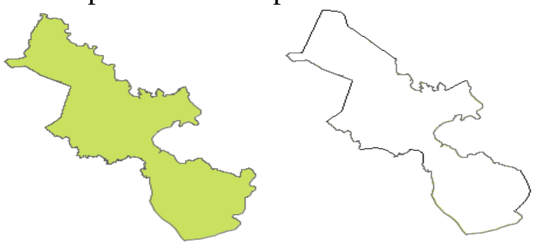

Fig. 1. Two maps can be recognized as similar in some spatial meaning extent.

2. THEORIES AND RELATED RESEARCHES

2.1 Measure theories

By definition, the function *d* for analysis measurement between two mathematics objects must be met three following conditions:

- Non-negative, that is d(A, B) $\geq$ 0. So the equality occurs when two objects is absolutely identical.

- Symmetry, that is d(A,B) = d(B,A).
- Triangle inequality, that is d(A,B) + d(B,C) ≥ d(A,C).

For examples, by definition, Hausdorff distance from set A to set B is defined as the smallest distance from each elements of A to B. And the Hausdorff measurement of two sets is the maximun value of Hausdorff distance from set A to set B and from set B to set A as:

$$d_H(X,Y) = \max\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\}$$

***Equ 1*. Hausdorff distance of 2 set X and Y**

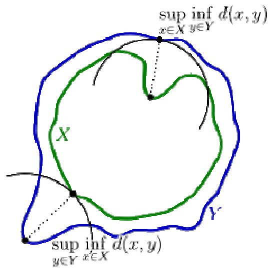

Fig 2. The description of Hausdorff measurement

between two geometry sets [taken from wikipedia].

Many researches have showed that the computation of Hausdorff measurement between two geometry sets with having m and n objects is:

- O((m+n)log(m+n)), for non-intersect lines.
- O((mn)2log3(mn)), for polygon case.

In some projects, the Hausdorff measurement is combined with other methods such as neural network to find patterns. Hausdorff measurement serves as a coarse searching filter to accelerate the speed. Hausdorff method can also find the rotate object from the sample. Study by Maylor Leung and Yu Xiaozhou (reported on April 29th, 2004) has showed that many cases using Hausdorff methods formeasuring the set of points, lines, angles and polygons and they have succeeded in searching geometry in images:

- Hausdorff Distance (HD), introduced by Huttenlocher (1993).
- Modified Hausdorff Distance (MHD).
- Line Segment Hausdorff Distance (LHD1), introduced by Gao and Leung (2003).
- Modified Line Segment Hausdorff Distance (LHD2), introduced by Chen and Leung (2003).
- LHD by Yu and Leung (2003).
- Curve Segment Hausdorff Distance (CHD), introduced by Yu and Cheng (2003).

In addition, others metric measurement methods used to compare two geometric objects are:

- Hamming distance.
- Fréchet distance.
- Minkowsky distance.
- Reflection distance.
- Bottleneck distance.
- Turning function distance.
- Nonlinear Elastic Matching distance.

- And some traditional geometry methods such as: skeleton comparison, quantified by polynomial or string to compare.

2.2 Map generalization and related issues

In theory, the map is characterized by three compositions: the topic, the purpose and the scale. The combination will establish the content, the detail level and the completeness of the map. In particular, the map scale strongly affects the details and the completeness. So, 1-square-kilometer in field will be presented as:

- 1 square meter on 1:1,000 map.
- 1 square decimeter on 1:10,000 map, and so on.

Clearly, not all the elements on the 1:1,000 map can be presented on the smaller-scale maps. Because some factors will not be shown up or will not be shown exactly the real shape in the smaller-scale maps. Here, we get the map making process named map generalization or cartographic generalization. Map generalization is the method to show up the major map features and characteristics for the phenomenon reflected the map topic. Thereby, the map generalization will reduce the complexity of the map while keeping the meaning of the map. In this process, the important objects will be retained and the unsubstantial objects will be left. In reality, the process depends on the map scale and the quality of the database as well as slightly the limitation of presentation system (cited in chapter 5, Digital Cartography book, authors Robert G.Cromley). There are two wide-accepted concepts of map generalization: the first concept by Goodchild (1991) and the second concept by Muller (1991). A common idea is the purpose-oriented map making by some processes. So far, eight generalization techniques have been categorized by Jones (1997): elimination, simplification (or reduction), typification, exaggeration, enhancement, collapse, amalgamation, and displacement and two additional ones have been implemented on currently softwares: smoothing and refinement.

- Simplification.
- Smoothing: replace the sharp and complex objects by the smooth ones.
- Aggregation: replace many types by one type object.
- Amalgamation: replace some objects with representative object (for polygons).
- Merging: replace some objects with representative object (for lines).
- Collapse: replace the region by set of points and lines.
- Refinement: replace the complex objects by the satisfied one.
- Exaggeration: maintain the objects paricularilties.
- Enhancement: bolding, or replace the size or the symbols to enhance the visualization.
- Displacement: moving objects out of the real position to show and distinguish them.

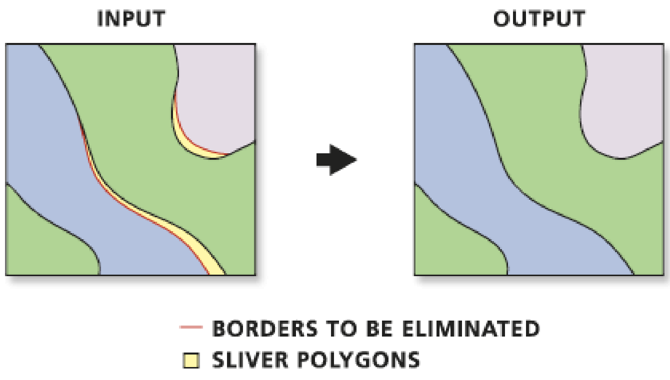

Fig. 3. Eliminate function is used in map error corection process. (taken from ESRI ArcGIS document)

For years, map generalization has been supported in either open and close sources GIS softwares. As in ArcGIS, the desktop software of ESRI, the Generalization functions are implemented in the toolbox, which are:

- Aggregate polygon.
- Collapse Dual Lines To Centerline.
- Dissolve.
- Eliminate.
- Simplify Building.
- Simplify Line.
- Simplify Polygon.
- Smooth Line.

2.3 Geometric meaning of spatial objects

In database, objects are belong to one of two relations: is_a or part_of. In the map, the objects inherits two above relations and have the spatial relations. Spatial relations can be catelogised into these sematic factor groups:

- Geometric sematic factor: distance, topology, bearing are the geometric relations. For examples, if we have 4 points and they may be on the line, then we have the relation bearing.
- Shapeindex sematic factor: for polygons, the ratio of its length and its area will provide us some information about the shape.
- Statistics sematic factor: the objects distribution on the map is the sematic factor. We have three types of GIS data distribution: randomize, cluster and regular. For example, if we search for a clustered points, then we could state that they may not houses map in Vietnam rural areas. Geostatistic could provide more information about the objects on the map such as: the spatial central tendency, the spatial dispersion (mean distance, spatial variation, spatial variant, spatial distance), the spatial orientation, the spatial autocorrelation. More and more, they are attributes classification: quantile classification, histogram classification, optimal classification, and multivariate classification.
- Graph sematic factor: the connection of vertex and edges, the degree of vertex., connected graph. This sematic factor can be the qualitavice factor for the connection. For example, we could find the house near one T-function and next to the alley by firstly finding all T-junction on the map.
- Calculus sematic factor: some calculus distances, such as Hausdorff distance, are a map sematic factor. For line and polygonal layers, Hausdorff distance method is usually used to comparing objects when the scale factor does not care. Indeed, this factor may be the quality of data. For instance, the sets of sample points forming the river border will be not identical but the river could be drawn by any set of these sets.
- Other spatial attribute sematic factors: for instance, the spatial proportion is one of the spatial attribute sematic factor. If the template for searching have a number of polygons which difference areas, then we can sort the objects by area to find the rules.

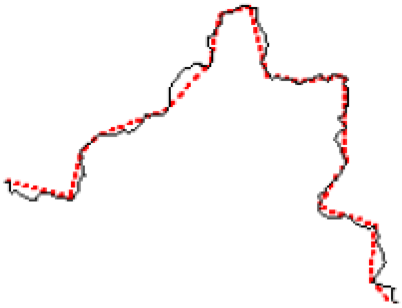

Fig. 4. Map generalization could smooth the water border (taken from map generalization document)

3. THE GENERAL MODEL FOR SPATIAL SIMILARITY SEARCHING

We recommend the overall model for similarity spatial searching based on spatial description. The main idea of this kind of search is how the pattern map is described. By answer the question, maps are made by the sematic obtaining in the pattern map and driven by the searching methods.

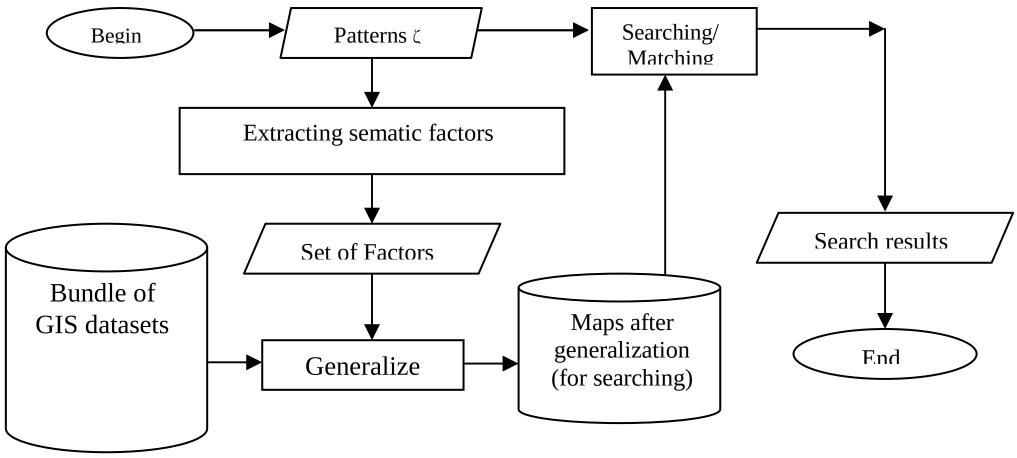

Fig. 5. The overall model for similarity spatial searching based on spatial description

- Step 1st: Input the searching pattern set ζ.
- Step 2nd: Extract the "geometry-meaning" information and decide the transform sets.
- Step 3rd: Establish the set of map for searching by generalization the bundle dataset.
- Step 4nd: Do searching the similarites from the maps by the patterns.
- Step 5nd: List the results with some threshold parameters.

Based on searching model, we must generate many maps by map generalization. Then, we will have the set of maps in general:

Table 1.Map generalization with geometry "meaning"

Generalization	***1st meaning***	***2nd meaning***	**…**	***Nth meaning***
Data layer 1	Map $_{11}$	Map $_{12}$	…	Map $_{1n}$
Data layer 2	Map $_{21}$	Map $_{22}$	…	Map $_{2n}$

…	…	…	…	…
Data layer m	Map $_{m1}$	Map $_{m2}$	…	Map $_{mn}$

After generalization, we will get a set of map. Then, we can employ an search technique such as some measure method (Hausdorff,…) or shape transformation into alphabetic string as the KMS project implemented by National Mapping Agency of Denmark.

In this step, we can established the mapping between geometry and some method for representative the spatial object, such as the method to convert the shape to the Left-Right string. Thus, the computational demands mainly on each layer with the feature type and the feature quantity parameters. So, we can build a data structure, such as the tree index on data to help quickly searching.

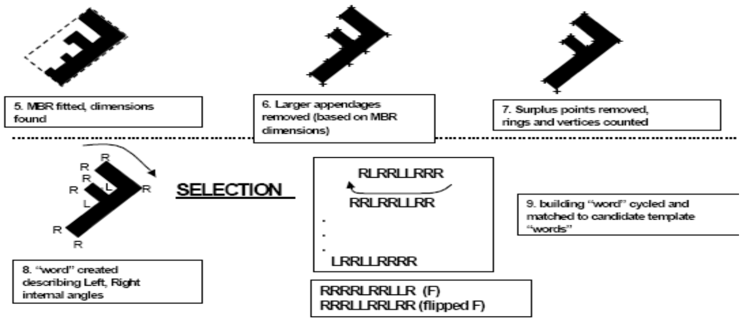

Fig. 6. One of the methods to string the shape (taken from KMS project).

4. PARALLELIZATION SPATIAL SIMILARITY SEARCHING MODEL

From the above analysis, we have an strategy to parallize after the extraction geometric information:

- *Strategy 1^{st}: Parallezation the map generalization per each geometric meaning. That is, each computation node will handle a set of layer doing the same generalization rule.*

- *Strategy 2^{nd}: Parallezation the map generalization per each data layer. That is, each computation node will handle some layer and do all generalization .*

Each strategies has its advantages and disadvantages. The former method may causing unbalancing when the set of transformation calculations from geometric meanings not identical. In constrast, the later will lead to the speed not uniform by the differences on layers such as the number of spatial objects on each layer, the complexity of each spatial objects…

When doing map generalization, each geometric meaning is the set of the transformations:

$$\text{Geometric meaning } k = \lambda_1 \text{Transform } k_1 + \lambda_2 \text{Transform } k_2 + \ldots + \lambda_p \text{Transform } k_p$$

λ_I is the difficulity and complexity of the i^{th} transformation. The λ_I is coefficient of computing time.

Each layer has its own properties: the quantity of elements, geometric type of objects... And the data of one layer can be divided into blocks θ_1, θ_2,…, θ_r . Assuming that the procesing speed of processors is the same, then we can establish the optimal programming to find the solutions.

The parallel searching strategy will be chosen from 2 two posibilities: spatial extent or data layer. Spatial extent division may face with the limitation of searching extensive region. And strategy will be searching with many map partitions including non-divide case. For splitting cases, we should take care of the overlapping regions.

If the data contain many kind layer such as points, lines and polygons, then the set of spatial meaning will be assessed whether to do search on every single layer. Because the map generalization will affect to the layers. But the "meaning" may be the set of geometries with the existence of three categories (point, line and polygon).

From all above analysis, here we recommend the overall parallel processing for the spatial similarity searching:

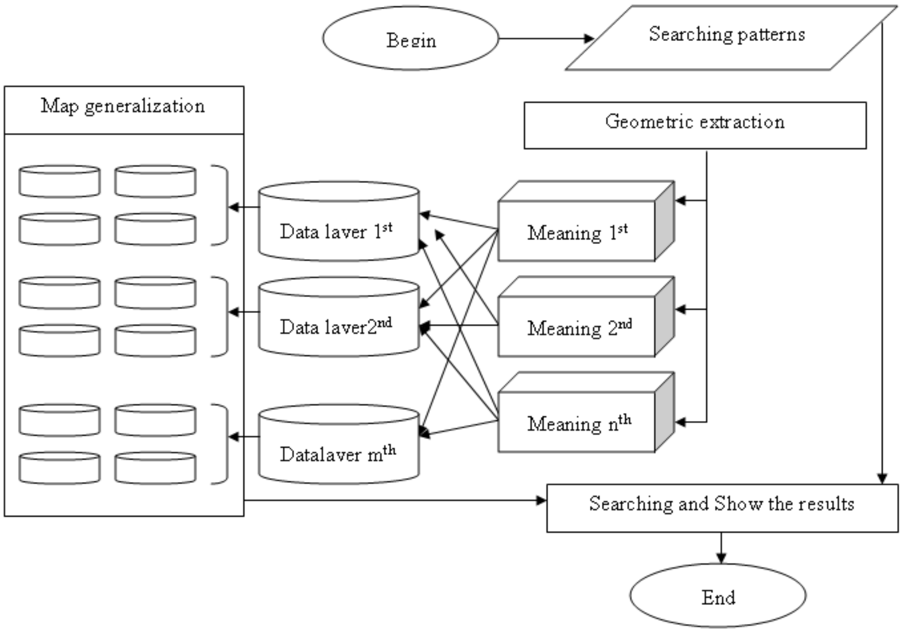

Fig. 7 . Overall model for parallel spatial similarity searching processing

In the overall model, the parallelization work focuses on the meaning finder, the map generalization process and the searching on each machine. The optimal system will schedule the parallelization to speeding the computations.

5. CONCLUSION AND DIRECTION

There are a lot of works to do for the similarity search based on spatial description problem. And the key problem is the sematic extraction for the searching pattern. This paper presents some factors and they take a role of search orientation.

Finally, beside refinement and optimization the code for searching, the next step of this research may be study on many kinds of data, such as for 3D data or timeseries data or data of nonlinear prediction spatial patterns.

6. REFERENCES

DƯƠNG Minh Đức, 2006. Textbook: *Lý Thuyết Độ Đo (Theory of measure).*

NGUYỄN Đức Nghĩa, 2008. Textbook: *Bài giảng môn Tính toán song song, Đại học Bách Khoa Hà Nội* (*Graduated lectures on Parallel Computing, Hanoi University of Technology).*

LÊ Thị Ngọc Liên, 2001. Textbook: *Giáo trình Biên tập bản đồ.*

TRẦN Văn Hoài, 2010. Slides: *Graduated lectures on Parallel Computing.*

Richard Healey, Steve Dowers, Bruce Gittings and Mike Mineter, 1998. Textbook: *Parallel Processing Algorithms for GIS*, publisher Taylor&Francis.

J.P. Marques de Sá, 2001. Textbook: *Pattern Recognition – Concepts, methods and applications*, Springer.

Kulik, L., Duckham, M., and Egenhofer, M., 2005. *Ontology-driven map generalization*.

Barbara P.Buttenfield, 1991. *Map generalization: Making rules for knowledge representation, A rule for describing line feature geometry*, Longman Scientific&Technical, 150-171.

Eric Reiher, Fady Said, Ying Li and Ching Suen, 1996. *Map Symbol Recognition Using Hausdorff Distance and a Neural Network Classifier*, Canada.

Jeff Danciger, Satyan L. Devadoss, John Mugno, Don Sheehy, and Rachel Ward, 2007: *Shape deformation in continuous map generalization*.

Alec Holt, 2003. *Spatial Similarity*.

Alec Holt, 1999. *Spatial similarity and GIS: the grouping of spatial kinds*. Spatial Information Research Center – University of Otago, Dunedin, New Zealand.

Maylor Leung, Yu Xiaozhou, 2004. *Hausdorff Distance for Shape Matching*.

Desmond Rainsford và William Mackaness, 2001. *Template Matching in Support of Generalisation of Rural Buildings*.

Mindshade project, http://www.cs.uu.nl/groups/AA/multimedia/publications/pdf/mindshade.pdf