Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 10 (2011) 1939 – 1945

Procedia
Environmental Sciences

ELSEVIER

2011 3rd International Conference on Environmental
Science and Information Application Technology (ESIAT 2011)

Method Exploration of Self-adaptive Entity Matching in Map Fusion

WU Jianhua [a, b]*, ZHOU Jingyun[c] and WU Buwei[c]

[a] ***School of Geography and Environment, Jiangxi Normal University, 99 ZiYang Road, Nanchang,Jiangxi,330022, China***

[b] *Key Lab of Poyang Lake Wetland and Watershed Research, Ministry of Education, 99 ZiYang Road, Nanchang,Jiangxi,330022, China*

[c] *School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, Hubei,430079,China*

lihuawu111@sina.cn

Abstract

Entity matching is a crucial and hard technology in map fusion. Current methods still exists some deficiencies, such as matching efficiency is not high, low degree of automation and poor universality, these methods can not meet the matching needs of large data integration, therefore, the urgent need to develop more effective and intelligent methods. This paper analyzed present research situation and existing problems of entity matching, illustrated the necessity of developing self-adaptive entity matching, pointed out urgent research contents and key issues that need to be resolved urgently in self-adaptive entity matching, provided preliminary research scheme of implementing self-adaptive entity matching, finally, introduced characteristics and advantages of self-adaptive entity matching method presented in this paper.

Keywords: Map Fusion, Entity Matching, Entity Similarity, GIS .

1. Present situation and problems

Entity matching means that, through a series similarity indexes,to distinguish identical entities from the spatial data from different sources, and then built the corresponding relationship for related spatial entities. The research results of entity matching are helpful to improve automation degree of map integration and the quality of map data, etc. Entity matching theories and methods firstly came from a experimental map conflation project of U.S. census bureau between 1983 and 1985[1]. During the past twenty years, people have done many researches on entity matching from different application perspective, made certain progress and obtained some research achievements. (1) Classic point entity matching algorithms: matching algorithm based on distance threshold, matching algorithm of mutual

doi:10.1016/j.proenv.2011.09.304

nearest distance, matching algorithm based on combining multi characteristics; (2) Classic line entity matching algorithms: matching algorithm based on distance, matching algorithm based on shape description, matching algorithm based on property, matching algorithm based on spatial correlation, matching algorithm based on probability, matching algorithm of combining multi characteristics; (3) Classic area entity matching algorithms: matching algorithm based on the similarity of area or overlapping area, matching algorithm based on shape description, matching algorithm based on property, matching algorithm based on probability, matching algorithm based on combining multi characteristics[1-8]. Throughout the current research achievements, it's suggested that people have mainly made great progress on the entity similarity and non one-to-one matching strategy. But some limitations of the existing algorithms can still be observed, need further research and innovation. The problems found in the study are analyzed and summarized as follows:(1)*Similarity Index is not mature enough*:Some of the existing similarity indexes are either too simple or too complex when computed, what leads to poor practicality, and need further improvement or innovation. In the event that there are great differences of shape and location between matching data sets, the expression of geography object's geometric structure is inconsistent or data quality is not high (for example: lack of property information or correct topological relation information), the existing similarity indices are difficult to satisfy matching needs, so further research is needed to design a matching index with good recognition ability.(2)*Query algorithm easily leads to omitting match or mismatch and the speed is slow*： It's found in the research that many algorithms use a buffer of reference entity to determine whether the target entity is located within the buffer zone; and if it is, the target entity will be as the candidate matching entity of source entity. Since the buffer size depending on the characteristics of the data set, it's usually difficult to determine the appropriate buffer radius at the beginning. Therefore, this method is likely to cause missing match or mismatch, and the speed of creating a buffer is slow. Some algorithms use the minimum bounding rectangle(MBR) of entity to query the candidate matching sets, but this may lead to choose too much candidate matching entities (such as establishing a buffer for a long stream object to query), thus increasing the matching time. For the line entity matching, when line entity was horizontal or vertical, its minimum bounding rectangle is extremely small, what will cause missing match. And when great differences existed between the matching data sets, it isn't reliable to judge the similarity between entities only in a certain search area, so further studies are expected to find better query algorithms of candidate set.（3）*Matching strategy is not perfect*：Some of the existing matching methods use each matching index alone and calculate the similarity of each index in sequence. This matching strategy needs to determine the threshold of each similarity index, and the order of using indexes is also important for that different order of indexes may obtain different matching result. In order to get better accuracy and recall rates, some algorithms use these indices integrally in recent years, and calculate a total similarity value to determine matching result. Many papers have proposed the weight of each index determined by the expertise. In fact, the index weight is artificially set for specific data in their experiments, so the obtained matching results lack credibility. And it cannot assure the result obtained by integrally using of multiple indexes must be superior to which is obtained by using single index. Therefore, before the implementation of the overall index matching algorithms, there are two issues need to be clearly: first is that there is no need to use multiple indexes; followed by how to choose appropriate indexes for combination. And for matching data sets, using each index in order or integrally is not the best choice. Because in the matching data sets, some data may be sensitive to a index, while the other part of the data are more sensitive to other index. Therefore, application of these two strategies will all affect the matching results. It can be seen that the existing matching strategy is not perfect and better strategies are expected to be explored.（4）*Existing matching algorithm has poor generality*：The existing matching algorithms are all designed for specific data and application purpose, and the selection of index greatly depends on the characteristics of the potential data set, which is artificial selection. The search area (such as buffers) setting of candidate matching set will affect the matching result, and it's now determined rely on a tentative interactive method artificially, what limits the practicality of the method. In addition, the matching strategy of various algorithms is also different. All of these above disadvantages make matching algorithm with poor portability.

In conclusion, the current matching algorithms also have many deficiencies, and they are difficult to meet the needs of multi-source, multi-scale, multi-temporal map data integration. In order to improve the processing efficiency of entity matching and break the "high pertinence, poor universality" situation of traditional algorithms, developing more rapid and intelligent method is urgently expected. Therefore the author makes a preliminary study and exploration of the self-adaptive spatial entity matching method.

2. Concept and research content

The so-called "adaptive" generally refers to that system adjust its own according to changes in the environment so as to make its behavior achieve the best or at least allowance in the new or have changed circumstances. In this paper, the definition of self-adaptive spatial entities matching is given as follows: In the process of entity matching, matching flow or algorithm has universality for different series scale map data, it could dynamically determine the search range of candidate set, matching similarity index and its index weight and matching strategy according to its own characteristics and environment features, and then it could realize fast accurate matching. Adaptive spatial entity matching should be launched research around the following content:(1)*Candidate set rapid inquiry algorithm*:The optimized query algorithm of spatial relationship and the method to determine the search range dynamically should be studied to improve matching processing efficiency.(2)*The similarity calculation model suitable for complex data environment:*Further researching the point, line, area entities similarity, developing favorable matching basis, focusing on shape similarity and environment (entity surrounding scene) similarity adapted to the complex data environment (large difference between the position of namesake entity, different offset distance and direction between different matching pair), and establishing the comprehensive similarity index with strong discerning, what is helpful to improve the precision of matching.(3)*Research on map data characteristics automatic analysis and similarity index suitable evaluation method* :The existing matching algorithms are all designed for specific data and application purpose, and the selection of index greatly depends on the characteristics of data set, which is artificial selection, all of above result in poor universality of matching algorithm. This is largely due to lack a set of effective method for map data characteristics automatic analysis, as well as the similarity index suitable evaluation method for specific matching data currently. Therefore, research on map data characteristics automatic analysis and similarity index suitability evaluation method will be listed as one of the key research contents, so that in the process of matching, index could be screened out dynamically according to the characteristics of matching data even entity surrounding environment and its weight could be decided.(4)*Research on adaptive spatial entities matching method*:Doing research on the work mode and flow of self-adaptive spatial entity matching method, and establishing a set of harmonious, flexible and efficient operation mechanism; researching the factors which affect the search area of matching candidate set, similarity index selection, similarity index weight and matching strategy in matching process as well as their self-adaptive control method; researching the structure of similarity index model library, management methods and call method of similarity index.

3. The key issues to be solved urgently

(1) *The problem of spatial entity fast and accurately matching*:This problem mainly associates with the query algorithm of candidate set and the similarity calculation model. Therefore, the design and

composite application of methods to determine the search range of candidate set, the spatial query algorithm and the index with strong operation and recognition ability is the key to this project.

(2) *Map data characteristics automatic analysis and similarity index suitability evaluation method*: to realize the automation of matching algorithm and make it have the universality, the first important problem to be resolved is that formally describing and quantitatively calculating the characteristics of matching data and data differences. The solution to this problem depends on the integration of statistics, spatial analysis, computational geometry and other aspects of knowledge. Secondly, whether the similarity index is suitable for quantitative evaluation of the current matching data is also a key issue.

(3) *In the process of adaptive entity matching, how to accomplish self-adaptive control of search area of candidate set, similarity index selection, index weight and matching strategy is a key problem of adaptive spatial entity matching research*.

4. Implementation scheme of the research

(1)Research scheme of self-adaptive searching algorithm for candidate matching set

For point spatial data, the key to achieve self-adaptive searching area of the candidate matching set is to explore a method which can automatically determines the searching area of the candidate matching set. A preliminary research idea is to set a reasonable searching range based on convex hull area of the union set of candidate matching sets and the number of points in convex hull area. For line spatial data, ①if we query by using minimum bounding rectangle of line entities, firstly, it is noteworthy that when the line entity is horizontal or vertical direction (or close to that direction), the minimum bounding rectangle is too small, which would lead to leaky matching. This problem can be solved by creating buffer or expanding the area of the minimum rectangle according to maximum distance (or its multiples) of conjugate points in early analysis of data characteristic,secondly, if there are too many target entities queried by minimum rectangle (more than the set threshold), it will obviously increase the matching time, then the program should automatically choose to use the buffer to query (if the queried entities are more than the former, then we choose the former method), in order to achieve self-adaptive query range. ②if we use buffer to query, then we would achieve self-adaptive settings of buffer radius according to aggregation index (an index of aggregation level between current entity and entities around, in terms of dataset). For area spatial data, according to analytical result of data position difference, query is conducted by using the following three options: ①query the candidate matching set through the spatial relationships—interior intersection of planar entity (early experiments has proved that this method is faster) .②if there are too many target entities queried by minimum rectangle of planar entity (exceed the threshold), it will obviously increase the matching time,then the program should automatically choose to use the buffer to query so as to achieve self-adaptive query range. ③when the data offset is too big and have no regularity, then we need a wider area to query, the maximum area can be set to a reasonable value based on the analytical result of data position difference.

(2)Research scheme of similarity calculation model adapted to complex data environments

Because there are differences in precision, coordinate systems and other differences in geographic data from different sources of the same area, sometimes even they are after processing, it is still difficult to put them together completely. Apparently, in this complex data environment, the matching result derived from position -based matching method would be very poor. Aimed at this case, we focus on researching the shape similarity and environment similarity of an entity, based on a comprehensive study of the existing similarity calculation models. For shape similarity, using some existing methods for reference, we can use the moment method or multiple characteristics (For example: size, spindle orientation, bump area, compactness, solid degree, eccentricity, etc.) from the classical geometric theory to achieve a description of geometry. As for environment similarity (a describes the surrounding environment similarity of two entities to be matched (or entity set)), we can design by topological characteristics of entity, spatial direction relations, the distribution of the surrounding entities, etc.

(3)Research scheme of automatic analysis of map data characteristics and suitability evaluation method of similarity index

Preliminary idea of research is: the first step, to carry out various characteristic analysises for select map data by using statistical methods. The main tasks include evaluation of the integrity of attribute information, extraction of information-rich properties, topological quality assessment, distribution mode of geography objects and computing of length, direction, area and perimeter of geography objects, etc. The second step, make a comparison and association between the characteristics of reference data and target data. For example: If there are the same information-rich properties, then record the result of property comparison as "excellent" and establish relationship in properties. If both of their topological quality are not high, then the result of topology comparison record as "poor ", etc. The third step is to analyze the location difference between reference data and target data: On experimental software platform, establish some matching relationship between the identical points by artificial interpretation. Based on these conjugate points, calculate their spatial position difference (distance and offset direction), and compare the difference between the cognominal points, analyze whether the offset direction and offset distance of these identical points is same (or similar), and record the maximum of offset distance. These results will affect the later matching index and strategies to use. Suitability evaluation method of similarity index: suitability of similarity index is the degree that the similarity index suitable for current matching data. According to data characterstics,marking these similarity indexes in index database by using designed index suitability calculation model.

(4)Research scheme of self-adaptive entity matching method and experiment

①Use the workflow of the existing entity matching for reference, and considered the characteristics of self-adaptive entity matching (for example we need analysis of the data characteristics, index screening and suitability assessment, establishment of index database, etc.), establish workflow of self-adaptive spatial entity matching. Preliminary designed process as shown in figure 1:

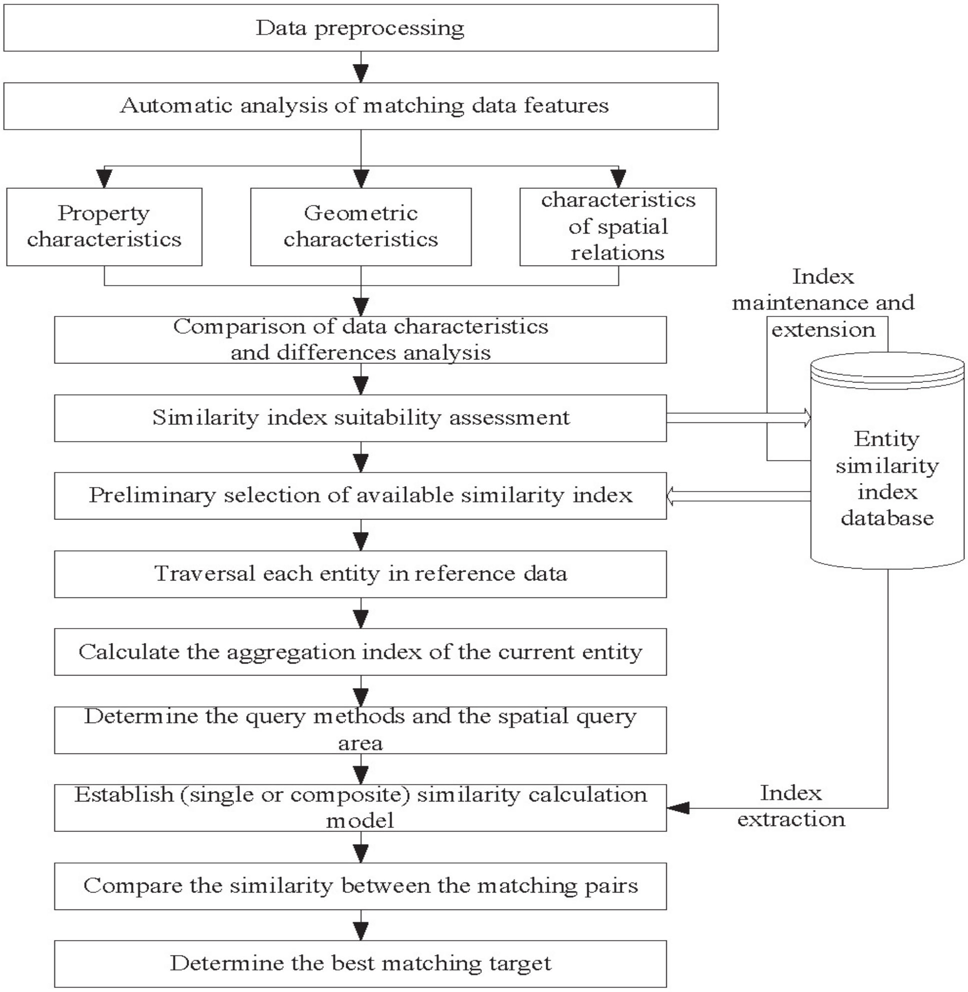
Data preprocessing
Automatic analysis of matching data features
Property characteristics
Geometric characteristics
characteristics of spatial relations
Comparison of data characteristics and differences analysis
Similarity index suitability assessment
Preliminary selection of available similarity index
Traversal each entity in reference data
Calculate the aggregation index of the current entity
Determine the query methods and the spatial query area
Establish (single or composite) similarity calculation model
Compare the similarity between the matching pairs
Determine the best matching target
Index maintenance and extension
Entity similarity index database
Index extraction

Fig.1 Self-adaptive entity matching process

②Self-adaptive control of spatial query area of candidate matching set, selection of similarity index, index weight and matching strategy. The spatial query area of candidate matching set is dynamically determined by geometric type, analytical result of data position difference, etc. Similarity index screening is preliminarily determined by analytical result of data features before matching, and it is dynamically determined by "aggregation index" (a characterization of aggregation level between current entity and entities around) of the entity during the matching process. According to suitability evaluation result of the selected similarity index, dynamically determine index weight by weighted average method. For matching strategy, we need to dynamically establish similarity calculation model which combining multi characteristics, according to suitability evaluation result of similarity index,and considering the "aggregation index" of current entity, forming a good information filtering mechanism, to improve the matching accuracy.③Research on methods of construction and management of similarity index database, and method of call index. Construction of similarity index database: establish similarity index information table in accordance with geometric type classification, which is based on a comprehensive analysis of similarity index of existing point, line and area entities,the table should contain the following information: index name, index algorithm description, calculation formula of index (it should be convenient for analysis and program call). In the aspect of index database management, we can accomplish management by developing special program, which would be convenient for maintenance and extension of the index. In the aspect of index call: to achieve the index call by developing special program for parsing and conversion . ④Using many groups of same (/similar) scale map data from the same region (include map data of city interesting points, urban drainage, electricity and other facilities, capital cities of province, towns, roads, river system, residential areas, etc.) to conduct experiments and algorithm optimization.

Summary

Innovations of the self-adaptive spatial entity matching methods presented in this paper includes three aspects as below:(1) in the aspect of matching efficiency and accuracy: it may improve processing efficiency of entity matching by optimizing query algorithms of candidate set from the following aspects: setting fit query area of matching candidates, spatial relations, spatial index and so on, By means of establishing similarity measure calculation model based on matching data characteristics and the "aggregation index ", the calculation model dynamically combining multi characteristics, and characteristic's weight is self-adaptive. (2)the existing matching algorithms select similarity indexes and set their weight commonly for specific data, so the matching results usually lack of credibility. Through the establishment of an extensible similarity index database, and presenting automatic analysis method of map data 's characteristics and suitability evaluation method of similarity index , that provide theory basis for filtering similarity indexes and determining their weights. (3) proposed theoretical framework of self-adaptive spatial entity matching and related algorithms, attempt to improve algorithm's universality and break the situation of poor universality.

References

[1] SAALFELD A. Automated Map Conflation[D] .Washington DC: University of Maryland, 1993.

[2] COBB M ,CHUNG M , FOL EY H. A Rule-based Approach for the Conflation of Attributed Vector Data [J] . Geo Informatica , 1998 , 2 (1): 7-35.

[3] VOLKER WALTER and DIETER FRITSCH. Matching spatial data sets a statistical approach [J].Geographical information science, 1999, Vol.13, No. 5: 445- 473.

[4] LI De-ren，GONG Miao-Yao，ZHANG Qiao-ping.On the conflation of geographic databases[J].Science of Surveying and Mapping,2004, 29(1):1-4.

[5] TONG Xiao-hua,DENG Su-su, SHI Wen-zhong. A Probabilistic Theory-based Matching Method [J]. ACTA GEODAETICA et CARTOGRAPHICA SINICA,2007, 36(2): 210-217.

[6] FU Zhongliang, WU Jianhua.Entity Matching in Vector Data[C]. 21th SPRS2008, 2008, 7: 1467-1472.

[7] WANG Yu-hong , CHEN Jun.An Instance-Based Approach for Schema Matching Between GIS Databases [J].Geomatics and Information Science of Wuhan University,2008, 33(1):46-50.

[8] FU Zhong-liang, SHAO Shi-wei, TONG Chun-ya.Multi-scale Area Entity Shape Matching Based on Tangent Space[J]. Computer Engineering,2010,36(17):216-220.