ENTITY MATCHING IN VECTOR SPATIAL DATA

FU Zhongliang[a,b*] , WU Jianhua[a]

[a]School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, Hubei, 430079, China - wjhgis@126.com
[b]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,Wuhan University,129 Luoyu Road, Wuhan, Hubei, 430079,China - fuzhl@263.net

Commission IV, ThS–9

KEY WORDS: GIS, Vector Data, Entity Matching, Spatial Relation, Spatial Query, Similarity Measure, Update

ABSTRACT:

Entity matching is a crucial and hard technology in application of vector spatial data integrating, data updating and map differential analyses. According to the disadvantage of matching algorithms nowadays, from the candidate searching algorithm of entity matching, similarity measure and matching strategy, this paper does a deep research in this three aspects. And proposes an area entity searching algorithm based on interior intersection, and another line entity searching algorithm based on buffer division. Both could enhance the seed of searching candidate match set, decreases the number of candidate matching entities. Also in this paper, giving an area entity integrated similarity measure index integrated entity area, overlap area, barycentre distance etc, and a line entity integrated similarity measure index based on buffer overlap area, azimuth code, length etc. . Both of them improve the recognizing ability of identical entity. Through the match strategy of bidirectional matching and clustering combination, effectively achieves the entity matching of one-to-many and many-to-many.

1. INTRODUCTION

Entity matching means that, through a series similarity measures, to distinguish identical entities from the spatial data from different sources, and then built the corresponding relations for related spatial entities. Because of multi-source, multi-temporal、multi-scale spatial database, a expression for the same geographic entity could be inconsistent in point position precision, spatial position, geometric construction, topological relation, semantic, attribute structure and types. This causes a relative difficult for spatial identical entity matching. Nowadays, researchers from different countries have done some search on spatial data entity matching. Saalfeld, pioneer of map combination, proposed a point matching method of combining point geometry position and spider code, and another line entity matching method based on L_2 distance (saalfeld, 1993). On matching strategy, Rosen and Saalfeld proposed a strategy: first matching point, then matching line and area; according to first considering strong condition (geometry shape, distance etc.) matching, if can not match, then using weak condition (topological relation). Zhang Qiaoping proposed a matching method: first according to the size of the overlap area of area entities to decide candidate matching set, then using fuzzy topological relations and clustering to affirm matching type. In addition, he also proposed a distance similarity measure based on mid-area between two line entities and a shape similarity measure based on the differences of direction change along each line. Walter and Fritsch proposed a matching method based on probability statistics. Using the method of "buffer growing", to obtain candidate matching set; through region statistics, could decide matching threshold. Finally by using Merit Function in information theory, could confirm matching results. This method is rigorous in theory and good performance on matching results. However, the calculation process is complex and time consuming; at the same time, matching threshold has influences on the results. Cobb etc. proposed a non-spatial property data matching strategy based on knowledge, deciding matching entities through calculating similarity of properties. Do Xiaohua etc. adopted a matching algorithm of multi-indexes fusion based on probability theory. It could avoid selection of precise threshold, and make the use of bi-directional matching . So, it solved the problem of one-to-many's situation, but this method didn't consider the matching instance of two endpoints without matching , viz. it didn't resolve the matching of many-to-many.

Overall, methods above are not ideal solutions for entities matching and the main problems is: (1) the candidate matching search algorithms for entities are so slow, and low efficiency (2) the entities can not handle the situation of many-to-many matching; (3) the spatial similarity is unacceptable, similarity calculation model is either too simple or too complex. In view of the above three defects, this paper mainly does research in three aspects: searching algorithm for candidate matching set, the similarity measure index, and matching strategy. Moreover, proposes an area entity searching algorithm based on interior intersection, and another line entity searching algorithm based on buffer division. Both could enhance the speed of searching candidate matching set, and decrease the number of candidate matching entities; builds an area entity integrated similarity measure index integrated entity area, overlap area, barycentre distance etc; designs two new line entity similarity measure indexes: similarity measure index based on overlap area of buffer and shape similarity measure index based on azimuth code; Through the method of bi-directional matching and clustering combination, the matching problem of one-to-many and many-to-many are resolved effectively.

[*] Corresponding author: FU Zhongliang, Email: fuzhl@263.net

2. MATCHING STRATEGY OF SPATIAL ENTITIES

Entity matching is a process to identify the same object from different entity sets and establish their relevant relations. The general concept of entities matching is as follows: Let two entity sets to be matched are $A=\{a_1,a_2,\cdots,a_m\}$ and $B=\{b_1,b_2,\cdots,b_n\}$ respectively, if searching the candidate entity $b_j(j=0,1,\cdots n)$ in B according to the entity $a_i(i=0,1,\cdots m)$ in A (b_j may include multi entities), then a_i is called “source entity”, b_j is called “target entity”, and vice versa. If b_j is the optimal matching entity of a_i , then a_i and b_j make the matching pair. At the aims of reducing the computing time of entity matching and checking the entity matching results, this paper divides entity matching into three phases: coarse matching, precise matching and confirming identical entities: ①in the phase of coarse matching, obtains candidate matching set of source entity by calculating spatial relationships and some geometric restriction condition, candidate matching set of a_i is expressed as Ca_i ;②precise matching is a process to identify the optimal matching entity from candidate matching set for source entity by using the similarity measure index which has stronger distinguishing ability; ③confirming identical entities is used to judge whether two entities are identical entities, for the instance of one-to-one matching, if the similarity measure is larger than a certain threshold, then confirming that they are identical entities, for the instances of one-to-many and many-to-many, combining entities which belong to A into a complex entity ,so do the entities which belong to B, and then calculating similarity measure of the two complex entities, if the similarity measure is large than some threshold, then they can be considered as identical entities.

In order to effectively solve the instances of one-to-many and many-to-many matching, this paper adopts a strategy which need a bidirectional matching and clustering combination, bidirectional matching means that, above all, for each source entity in A, finds the optimal matching target entity in B, then takes the entity which didn’t create matching pair in B as source entity, to find its optimal matching target entity in A. finally conduct clustering and combination to all matching pairs and build matching relations among entities.

3. MATCHING ALGORITHM OF AREA ENTITY

3.1 Integrated similarity measure index of area entity

Because of different number of vertices of two area entities which from different sources and position difference between vertices, the endpoint and the starting point of entity boundary are not clear, in the cases of one-to-many, many-to-many matching, it is difficult to calculate the indexes of boundary distance and boundary shape similarity, it results in area entity matching difficultly by vertices in boundary or boundary; but the area entity matching method based on overlap area has best merit that it can easily distinguish the various complex cases of entity matching(Zhang Qiaoping, 2002). Zhang Qiaoping only presents the similarity index based on overlap area in the course of determining the candidate matching set, calculation formula of similarity based on overlap area between the two area entities as follows:

$$Sim(A_i,A_j)=\frac{Area(A_i\cap A_j)}{Area(A_i)}(i,j=1,2,\cdots,n) \tag{1}$$

When $Sim(A_i,A_j)$ or $Sim(A_j,A_i)$ is greater than a certain threshold (such as 0.5), means that there is a possibility of matching between A_1 and A_2 , but the ability of identifying identical entities is weak. literature (TONG Xiaohua etc. , 2007) confirms matching entities according to the probability calculated by similarity measure index differences such as overlap area and barycentre distance, index difference of overlap area is regarded as the difference between overlap area and the minimum area of two area entities, the area similarity of two entities still not be considered, but the area similarity can properly reflect the similarity of two entities, and accords with people’s intuition relatively. According to the deficiencies of the above similarity of area entities, this paper presents an integrated similarity measure index, which fuses a variety of indexes such as entity’s area , overlap area and barycentre distance, the calculation formula of integrated similarity measure index as follows:

$$\rho_{ij}=\alpha\times(\frac{S(a_i\cap b_j)}{S(a_i)})+\beta\times R_{ij}+\gamma\times\left(1-\frac{d}{l}\right) \tag{2}$$

$$R_{ij}=\frac{min[S(a_i),S(b_j)]}{max[S(a_i),S(b_j)]} \tag{3}$$

a_i , b_j are the entities(or entities combination) in A and B respectively, they can match with each other, $S()$ expresses area of some region, α,β,γ are the weights used for measuring the similarity of entities, α+β+γ=1;α, β and γ are the adjustable parameters, they are 0.4,0.4,0.2 respectively here, d is the barycentre distance of a_i and b_j , l is the length of diagonal of minimum bounding rectangle of a_i ,if $d>l$ then $d=l$ (FU Zhongliang and WU Jianhua, 2007).

3.2 Algorithm of coarse matching of area entity

In order to reduce the computing time of entity matching and improve the efficiency of matching, first of all, coarse entities matching is needed, to calculate entity’s candidate matching set, specific process is described as follows: adds the entities which satisfy the condition of formula(5) into the candidate matching set $Ca_i=(b_1,b_2,b_3,\cdots)(Ca_i\in B)$.

$$\rho_{ij}=\frac{S(a_i\cap b_j)}{min[S(a_i),S(b_j)]}\geq 0.5\ \ (i,j=1,2,\cdots,n) \tag{4}$$

Due to the massive characteristics of spatial data sets which are used for entities matching, in order to improve the efficiency of matching, it should try to avoid searching candidate matching set in large area. Most of the existing entities matching algorithms using the minimum bounding rectangle to search for matching candidates. The literature (Zhang Qiaoping, 2002), firstly, judges whether minimum bounding rectangle of two area entities intersect with each other when search matching candidates , and then gets the polygons intersect with current polygon, if the collection of intersection polygons is not empty, then calculate similarity measure. This method will select too many matching candidates, thereby increases time of matching. Therefore, this article presents a new searching method based on interior intersection for area entity:Let a , b are two entities match with each other, let $I(a)$, $I(b)$ represent internal point set of a , b respectively, if $I(a)\bigcap I(b)\neq\varnothing$, it is possible that b is a candidate matching entity of a .The method has characteristics of less calculation amount and higher searching accuracy.

3.3 Algorithm of precise matching of area entity

After finishing coarse matching, it is necessary to conduct precise matching, to calculate the target entity which best match to the current entity a_i , specific process is as follows: calculates the similarity measure index $I(a_i,b_j)$ of a_i and each entity b_j in Ca_i by using the integrated similarity measure index presented in this paper, the entity b_j which $I(a_i,b_j)$ value is maximum is the optimal matching target entity, then relation of matching pair and index $I(a_i,b_j)$ can be recorded.

After traversing all entities in A, a reverse matching is needed, takes entity b_j which still no matching pair as source entity, and search its optimal matching target entity a_i in A, at the same time, records the relation of matching pair and index value $I(a_i,b_j)$.When finishing bidirectional matching for entities set A and B, one-to-one mapping relation of entities is created, because of the matching instances of one-to-many and many-to-many, further process for the result of entities matching is needed, namely, conducts clustering and combination for all matching pairs created previously. Supposes $Pc_id_i=\{(c_0,d_0),(c_1,d_1),(c_2,d_2),\cdots(c_n,d_n)\}(i=0,1,\cdots,n)$ is a collection of matching pair, $c_i\in\{a_0,a_1,a_2,\cdots a_n\}$, $d_i\in\{b_0,b_1,b_2,\cdots b_n\}$, let (c_i,d_i) is the current matching pair, if there is a matching pair $(c_j,d_j)(i\neq j,j=0,1,\cdots n)$ which has intersection with (c_i,d_i), namely, at least has a common entity, then combine them together, and then combines the matching pair which has intersection with the combined result which obtained in previous step, until no longer be able to combine, lastly, the result of combination is expressed as $Cls_t(t=1,2,......,n)$,Cls_t is divided into two clusters, they are $subClsa_t$ and $subClsb_t$ respectively, $subClsa_t$ is consisted of all entities which belong to A in Cls_t , $subClsb_t$ is consisted of all entities which belong to B in Cls_t , the instances of one-to-many and many-to-many matching are transformed into one-to-one matching by $subClsa_t$ and $subClsb_t$ .For the instance of one-to-one matching, if its value of similarity index is larger than or equals to 0.8 , then considers that they are identical entities; for the instances of one-to-many and many-to-many, it should combine the entities in $subClsa_t$ and $subClsb_t$ into complex area entity respectively, then calculates the similarity measure of the two complex area entities by integrated similarity measure index, if the value is larger than 0.7(if the map scale difference between A and B is larger, the index threshold can be reduced properly), they can be confirmed as identical entities.

4. MATCHING ALGORITHM OF LINE ENTITY

4.1 Integrated similarity measure index of line entity

In the aspect of line entities matching, people also put forward some entity similarity indexes, geometric similarity measure indexes which are usually used including distance similarity measure index, shape similarity measure index, direction similarity measure index. Over all, the geometric similarity measure calculation model either too simple or calculating complexly. Aim at deficiencies of the existing similarity measure index, this paper proposes the integrated similarity measure index based on weight for line entity, this index fuses three similarity indexes: First, the line entity length similarity measure index, second, the similarity measure based on overlap area of buffer, third, similarity measure based on azimuth code along each line in polyline. Next, this paper will respectively describe the various similarity measures and give the formula of line entity complex similarity measure based on weight.

4.1.1 Length similarity measure: This index is used to judge whether entities are similar or not by comparing length of a_i and b_j , the calculation is simple. The formula as follows:

$$\rho L_{a_i,b_j}=\frac{min[L(a_i),L(b_j)]}{max[L(a_i),L(b_j)]} \qquad (5)$$

Where $L()$ denotes the length of line entity, $\rho L_{a_i,b_j}$ is the length similarity measure of a_i and b_j .

4.1.2 Similarity measure based on overlap area of buffer: Though the research of the line entity's geometric characters, finds some good characters of the overlap area of two line entities' buffer, the buffer size can reflect the vicinity degree of two line entities, viz. it can reflect the transverse distance and lengthways distance of two line entities. As Figure 1 shows, according to the overlap area size of buffer, it is easy to see that b_1 is more similar to a_1 , b_4 is more similar to a_2 .

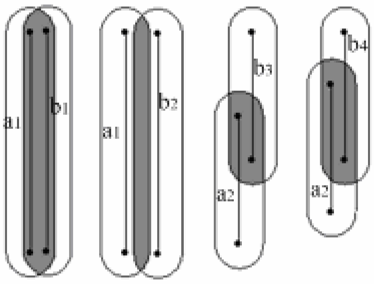


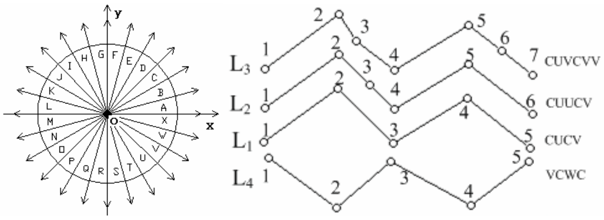
Figure 1. Overlap area of two line entities' buffer

So, this paper proposes similarity measure index based on overlap area of buffer, this index's function is similar to that of distance index. If a_i is the source entity, b_j is the target entity, then the calculation formula of similarity index can be denoted as follows:

$$\rho S_{a_i,b_j} = \frac{S(\,Buffer(\,a_i\,) \cap Buffer(\,b_j\,))}{S(\,Buffer(\,a_i\,))} \tag{6}$$

S() denotes the area of certain region, *Buffer()* denotes the buffer of line entity, $\rho S_{a_i,b_j}$ is the similarity measure index of a_i and b_j based on overlap area of buffer. This method is more intuitionistic than the distance index based on middle area between two line entities proposed by Zhang QiaoPing, furthermore, doesn't need to calculate the dual points of one line to another, amount of calculation is fewer, reduces complexity and difficulty of calculation.

4.1.3 Similarity measure based on azimuth code of polyline: The similarity measure index is used to investigate shape similarity measure between two line entities. The meaning of azimuth code is: First of all, the circumference around the origin of rectangular coordinate system is divided into some angle sub extents according to certain angle interval, and encoding for each angle sub extent; Figure 2 shows that the interval angle is 15 degrees, the entire circumference is divided into 24 sub extents, which can be encoded by A~X respectively. For each line segment in polyline, needs to find which sub extent the azimuth of line segment belongs to, if it locates a certain sub extent, then the azimuth code is equals to the sub extent's code . Code calculation formula as shown in formula (8). When the azimuth happens on the circle division line(except the positive direction of X coordinate axis), it is necessary to degrade the code calculated by formula (8) to it's next level, for instance, D is converted into C.



(a) (b)

Figure 2. Azimuth code of polyline. (a) angle division of circumference; (b) azimuth code of polyline.

$$code = Chr(65 + Int\left[azimuth(\overrightarrow{P_iP_{i+1}})/15\right])\ (i=1,2,\cdots,n) \tag{7}$$

Where *Chr()* is the function which can convert the number into code character(A~X), P_i , P_{i+1} are two sequential neighbour nodes, $azimuth(\overrightarrow{P_iP_{i+1}})$ denotes the azimuth of vector P_iP_{i+1} , $Int[]$ is the function used for returning integer.

After calculating azimuth code of polyline, code string is obtained; afterward, still needs to deal with code string as follows: ①to merge codes: If code string has two or more of consecutive same code, only one allowed to be reserved, in Figure 2 (b), the code of L_2 is merged into CUCV, become the same azimuth code as L_1 ; this step mostly aim to unify the number of nodes of line entities, facilitates comparing comparability; ②to eliminate exceptional azimuth codes: after merging codes, if length of the line segment which corresponds to the current code is less than 10 percent of length of line entity, then deletes the code. The step is designed to ignore the line entity's minor dithering in entity matching.

After dealing with code string, the shape similarity can be calculated by the azimuth code string of two line entities. Suppose $\rho F_{a_i,b_j}$ is the shape similarity measure of entity a_i and b_j , *sCode()* is the final azimuth code string of some line entity, *sNum()* is the character number of azimuth code string, if $min\,Num = min[\,sNum(\,sCode(a_i\,)),sNum(\,sCode(b_j\,))]$, $maxNum = max[\,sNum(\,sCode(a_i\,)),sNum(\,sCode(b_j\,))]$, and the character number of $sCode(a_i\,)$ less than or equals to the character number of $sCode(b_j\,)$, uses each code in $sCode(a_i\,)$ subtracts the corresponding position code in $sCode(b_j\,)$ from $K(K=1,2,\cdots,maxNum\text{-}minNum+1)$ to $K + min\,Num - 1$, if there is $m_k \in \{0,1,\cdots,min\,Num\}$ $(k = 1,2,\cdots,max\,Num\text{-}minNum\text{-}1)$ sequential same code, then calculation formula of similarity measure index based on azimuth code as follows:

$$\rho_{a_i,b_j}(\,F\,) = \frac{max\{m_k\}}{max\,Num} \tag{8}$$

Obviously, when two code string is equal, $\rho F_{a_i,b_j} = 1$, while the similarity measure of CDE and CDEFG is 0.6.

The shape similarity measure index based on azimuth code of polyline segment has some merits such as immutability of translation, rotation and proportion, as well as allows tiny dithering. Compares to Zhang QiaoPing's shape similarity measure, which based on the differences of direction change along each line, this method has the characteristics of intuitionistic and simple calculation.

According to the above three measurement indexes, the calculation model of line entity's integrated similarity measure index which considering weight as follows:

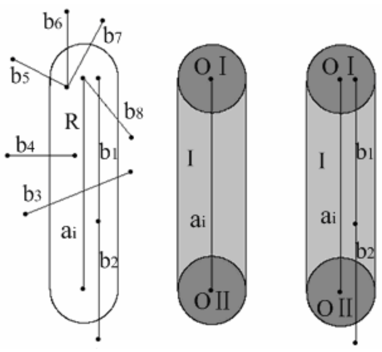
$$\rho_{a_i,b_j} = \alpha \times \rho L_{a_i,b_j} + \beta \times\ \rho S_{a_i,b_j}\ + \gamma \times \rho F_{a_i,b_j} \tag{9}$$

ρ_{a_i,b_j} is the integrated similarity measure index of entities a_i and b_j ;α,β,γ are weights of some similarity measure index of entity, α+β+γ=1;α,β,γare adjustable parameters, they can be set by the situation of entity matching.

4.2 Algorithm of coarse matching of line entity

Line entity's coarse matching is a process of searching target entities to be matched in B by using line entity's spatial relationships and putting the entities which meet demand into the candidate matching set $Ca_i = (\,b_1,b_2,b_3,\cdots)(\,Ca_i \in B)$ of

a_i .Due to expression of the same line entity from multi-source data exists differences at aspects of position of start point and endpoint, the number of nodes, the number of segments, components, the line entity matching is more complex. Using A and B to stand for two line entity sets, using a_i and b_j to stand for the $i(i=0,1,\cdots n)$ entity in A, and the $j(j=0,1,\cdots m)$ entity in B. a_i and b_j might have the spatial relationships of “disjoint”, “equal”, “contain”, “partial overlap” or “one dimensional intersection”. In order to obtain the entities to be matched for the current entity, may make full use of line entity’s buffer to search candidate matching set, the radius of the buffer determined by the actual situation of data, such as map scale, data accuracy, data intensive degree, and so on, determine the optimal value through several tests. When search candidate matching set by using spatial topological relationships of buffer such as “intersect” and “contain”, the situation is often complex, as it shown in Figure 3 (a).



(a) (b) (c)

Figure 3. Searching algorithm based on buffer division

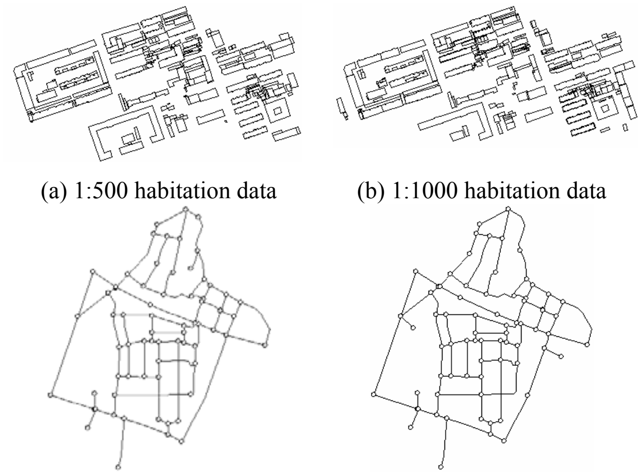
In order to quickly get the candidate matching set, and to improve efficiency of matching, this paper proposes the target entity’s searching algorithm based on buffer division: for each entity a_i in A, first get its buffer R according to buffer radius, in order to eliminate the entities which are obviously not identical entities, divides R into OI , OII and I three regions, as shown in Figure 3 (b), OI and OII respectively are the buffers of two endpoints of a_i (two dark grey regions in Figure 3 (b)), and named the union of OI and OII as O region, $I=R-O$ (Figure 3 (b), the light grey region). When searching, firstly, the entities which are fully included in R and belong to B will be seen as the candidate matching entities of a_i ; Secondly, if some entities in B have intersection with I and O of a_i at the same time, then the target entities which accord the following two situations can be added to the candidate matching set of a_i : First, the entity which has intersection of OI , OII , I three regions, Second, the entity which intersects with I and O regions, and an endpoint in I region. As shown in Figure 3 (c), the candidate matching entities of a_i are b_1 and b_2 .

4.3 Algorithm of precise matching of line entity

At the stage of precise matching of line entity, traverses each entity a_i in A, and searches its candidate matching set in B by using the searching method based on buffer division. The target entities in candidate matching set can be divided into several situations: ① two endpoints match with the two endpoints of entity a_i (the endpoint matching means that if an endpoint of other line entity in the buffer of an endpoint of a_i , then the two endpoints match); ②contained by the buffer of a_i and an endpoint match with the endpoint of a_i ; ③ contained by the buffer of a_i and no endpoint match with the endpoint of a_i ; ④ intersect with O and I region of a_i at the same time, and no endpoint match with endpoints of a_i . computes the integrated similarity index of each target entity and a_i , then selects the target entity which similarity measure index is maximum as optimal matching entity of a_i ; For the first two cases, the value of α,β,γ are set to 0.4,0.2,0.4 respectively, while the third case, the value of α,β,γ is set to 0.5,0.5,0, and for the forth case, the value of α,β,γ is set to 0,1,0. for the above four case, it is necessary to record relation of matching pair and it’s similarity index $I(a_i,b_j)$. After traversing each entity in A, takes entity b_j which still no matching pair as source entity, and search its optimal matching target entity a_i in A, at the same time, records the relation of matching pair and similarity measure index $I(a_i,b_j)$. After finishing bidirectional matching, conducts clustering and combination for all matching pairs created previously, and calculates the similarity measure index of complex line entities which are consist of the entities from one-to-many or many-to-many instances, while only adopts length similarity measure index, if the index is larger than or equals to 0.7, then the two complex line entities can be considered as identical entities. Process of confirming identical entities in the instance of one-to-one is same as area entities.

5. EXPERIMENTS AND RESULTS

Aiming at the above matching algorithms and strategies, this paper selects two kinds of habitation data of the same area but map scale is 1:500 and 1:1000 respectively, and two kinds of road data of the same area but from different sources, (figure 4), and makes some experiments on them. When doing the experiments, takes the habitation data of map scale 1:500 as source data, takes the habitation data of map scale 1:1000 as target data to be matched, regards the road data from first source as source data, regards the road data from second source as target data to be matched, Table 1 shows entities matching results of the above two types of data, Figure 5 shows the statistical results of different types of matching.



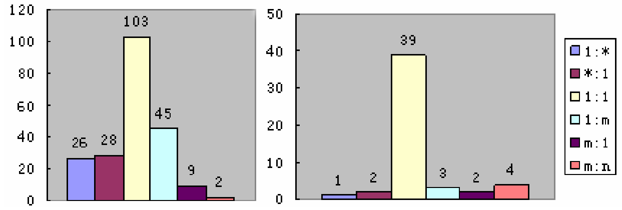
(a) 1:500 habitation data (b) 1:1000 habitation data

(c) road from first source (d) road from second source

Figure 4. Experimental data for entity matching

Data name	Habitation	Road
amount of entities in source data	198	55
amount of entities in target data	244	59
amount of 1:1,1:*,*:1 matching pair	283	66
amount of matching pair after clustering and combination	213	51
amount of identical entities matching pair	145	47
amount of error matching pair	0	1
ratio of error matching pair and successful matching pair	0	1.82%
amount of entities which matching successful in source data	158	53
matching success ratio of source data	79.79%	96.36%

Table 1. Results of entity matching



(a) (b)

Figure 5. Statistical results of different types of matching (*: expresses empty): (a) statistical result of habitation data; (b) statistical result of road data.

6. CONCLUSION

This paper studies on searching algorithms of candidate matching set, similarity measure index, matching algorithms and strategies of area entities and line entities. Compares to the existing methods of entity matching, these matching methods of area entity presented here have some merits: such as quick searching for candidate matching set, calculation of similarity measure index is simple and considers the area similarity of two entities; and matching method of line entity has following advantages: ①compares to the method of obtaining candidate matching set by using method of "buffer growing" presented in literature(WALTER and FRITSCH, 1999) , does not need strict topological adjacency of neighbor entities, and does not need union the buffers of entities, decreases the time of calculation, just depends on the current entity, the method is relatively flexible; ②this method of searching target entity based on dividing buffer into several regions, which can effectively eliminate the entities which are not appropriate to matching, improves efficiency of matching ; ③resolves line entities matching in the case of "two points of polyline all can not match" proposed in literature(TONG Xiaohua etc., 2007). Through the strategies of bidirectional matching and clustering-combination, effectively implements one-to-many and many-to-many matching for area entities and line entities.

Experiments show that matching methods presented by this paper have advantages such as calculating easily, high speed and robust, these methods can effectively resolve many-to-many matching in the condition of complex data. But some aspects are worth of further improving, for example, in one-to-many or many-to-many matching, if an entity matches with an entity with a number of overlapping entities, when using the strategy of bidirectional matching and clustering combination, often leads to mistaken matches, so, needs to explore a good way to remove the wrong entities in spatial data, in addition, this paper only makes use of geometric similarity measure indexes, but not adopts attribute similarity measure index and topology similarity measure index, which also a field worthy of further study in the future.

REFERENCE

Cobb M., Chung M., Foley H., 1998. A rule-based approach for the conflation of attributed vector data. *GeoInformatica* ,2(1): 7-35

FU Zhongliang,WU Jianhua. Update Technologies for Multi-scale Spatial Database. *Geomatics and Information Science of Wuhan University* ,2007,Vol. 32 No. 12:1115-1118

SAALFELD A. Automated Map Conflation. Washington D C: University of Maryland, 1993.

TONG Xiaohua, DENG Susu, SHI Wenzhong. A Probabil istic Theory-based Matching Method. *.ACTA GEODAETICA et CARTOGRAPHICA SINICA*, 2007,Vol.36, No.2: 210-217

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

ZHANG Qiaoping, LI Deren, GONG Jianya. Areal Feature Matching among Urban Geographic Databases. *Journal of Remote Sensing*, 2004, Vol.8, No.2:107-112

ZHANG Qiaoping. Research on Feature Matching and Conflation of Geographic Databases. Wuhan:Wuhan University, 2002

ZHANG Qiaoping, LI Deren, GONG Jianya. Shape Similarity Measures of Linear Entities. *Geo-spatial Information Science*. 2002, Vol 5, No.2, 62-67