AUTOMATICALLY AND ACCURATELY MATCHING OBJECTS IN GEOSPATIAL DATASETS

L. Li [a, *], M. F. Goodchild [a]

[a] Dept. of Geography, University of California, Santa Barbara, CA, 93106 US - (linna, good)@geog.ucsb.edu

KEY WORDS: Object Matching, Linear Programming, Assignment Problem, Optimization, Greedy

ABSTRACT:

Identification of the same object represented in diverse geospatial datasets is a fundamental problem in spatial data handling and a variety of its applications. This need is becoming increasingly important as extraordinary amounts of geospatial data are collected and shared every day. Numerous difficulties exist in gathering information about objects of interest from diverse datasets, including different reference systems, distinct generalizations, and different levels of detail. Many research efforts have been made to select proper measures for matching objects according to the characteristics of involved datasets, though there appear to have been few if any previous attempts to improve the matching strategy given a certain criterion. This paper presents a new strategy to automatically and simultaneously match geographical objects in diverse datasets using linear programming, rather than identifying corresponding objects one after another. Based on a modified assignment problem model, we formulate an objective function that can be solved by an optimization model that takes into account all potentially matched pairs simultaneously by minimizing the total distance of all pairs in a similarity space. This strategy and widely used sequential approaches using the same matching criteria are applied to a series of hypothetical point datasets and real street network datasets. As a result, our strategy consistently improves global matching accuracy in all experiments.

1. INTRODUCTION

1.1 Motivation

High-quality data are always the prerequisite for meaningful analyses. Since no single geographical dataset is a complete and accurate representation of the real world, we usually require data from diverse sources in scientific research and problem solving. In a particular geographical application, we need to obtain data from multiple sources that represent different properties of objects of interest. Unlike old days when research was impeded due to lack of data, rapid development of technologies for data collection and dissemination creates abundant opportunities for manipulating and analyzing geographical information. However, it is not always straightforward to take advantage of large volumes of geospatial data because data created by different agencies are usually based on different generalization schemes, using different scales, and for different purposes.

As it is impossible to directly collect all data by ourselves, we often need to utilize secondary data sources. Thus it is usually inevitable to combine multi-source data in science, decision-making, and everyday life. For example, in an emergency such as Jesusita Fire in Santa Barbara, effective evacuation requires integrative geospatial information about the affected area, probably including DEM, land use, residence and facility locations. Another typical application is the creation of an integrated database from two input datasets. It is possible that one dataset has all necessary features and attributes, but the other one bears a higher accuracy of positions. For instance, we have an old street network stored as vector data, and a recent remote sensing image that covers the same area. After extracting streets in the image, we want to identify the same streets in the outdated vector database in order to improve its positional accuracy. In all these cases, accurate identification of objects that represent the same entity in reality is an essential prerequisite to further analyses.

1.2 Objective

Object matching can be divided into two steps: the first step is to define a proper similarity measurement between objects, and the second step is to search for matched pairs based on this measurement. This paper focuses on the second step of this procedure by providing a new strategy for matching objects in multiple sources given a certain criterion. Rather than adopting a sequential matching procedure that is widely used in existing literature, we propose a matching algorithm according to an optimization model by regarding object matching as an assignment problem. In the remainder of this paper, Section 2 discusses two types of methods for object matching: widely used greedy method and proposed optimization method. In section 3, we describe two sets of data used in our experiments for a comparison between two methods. In section 4, we present the percentage of correctly matched pairs using different methods, followed by some conclusions in section 5.

1.3 Related Work

Object matching in geospatial datasets has been a fundamental research problem for decades. Most efforts have focused on the definition of similarity between objects. If two objects in different datasets are similar in terms of positions, shapes, structures, and topologies and so on, it is probable that they represent the same entity in the real world. The similarity metric varies from one application to another due to the inherent characteristics of input data and the availability of data properties.

The most popular similarity measurement is the proximity between objects. One typical criterion is the absolute proximity

* Corresponding author.

measured as a distance, such as the Euclidean distance between points or Hausdorff distance between polylines (Yuan and Tao, 1999), and some other distances in particular applications, like the discrete Frechet distance (Devogele, 2002) and radial distance (Bel Hadj Ali, 1997). The Hausdorff distance has been proved proper in calculating the proximity between linear features (Abbas, 1994). It is defined as the maximum distance of the shortest distances between each point on one linear object and a set of points constituting another polyline. When the distance between two objects is smaller than a threshold, they may be regarded as a corresponding pair. In addition to a distance threshold, another measurement based on a relative proximity is usually called the nearest neighbour pairing. This criterion intends to find the nearest neighbour of a particular object in the other dataset regardless of its absolute distance. If an object A in the first dataset is the closest object for object A' in the second dataset, and meanwhile object A' is the closest one for object A in the second dataset, objects A and A' are defined as a matched pair (Saalfeld, 1988; Beeri *et al*., 2004).

Besides proximity, other geometric information is also used in object matching. For example, matching between street segments may be reduced to node matching since nodes, especially intersections, are usually taken as control points (*e.g*. Cobb *et al*., 1998; Filin and Doytsher, 2000). The number and directions of connecting segments for a node are usually used to refine the matched candidates as a result of proximity criterion (Saalfeld, 1988). The angles between two street centrelines or between GPS tracks and street networks are also widely used in polyline and map matching (Walter and Fritsch, 1999; Quddus *et al*., 2003).

Another category of information for object matching is semantic similarity, including two important considerations: similarity between geographic types and similarity between individual geographic objects. In any dataset that involves geographical classes, it is critical to establish a mapping between different classification systems because any classification entails loss of information and usually subjective judgment. On the other hand, similarity between geographic objects may be defined according to attribute values, either numeric or string-similarity (Cohen *et al.*, 2003). Hastings (2008) used both types of semantic similarity - geotaxonomic and geonomial metrics - in conflation of digital gazetteers.

Furthermore, contextual information is also helpful for refining matching results based on the relationship between investigated objects and its surrounding environment. For instance, Filin and Doytsher (2000) developed an approach called "round-trip walk" to take into account contextual information. The counterpart nodes at two ends of the arc are called connected nodes. Two nodes are identified as matched only under the condition that they are similar enough and at the same time their connected nodes are also similar enough. When no explicit contextual information is available, Samal *et al.* (2004) proposed proximity graphs as an aid to incorporate context when landmarks are not connected with other features by constructing topology among them.

While these different methods all focus on the definition of similarity measurement in various datasets, few efforts, if there's any, have been made to improve the search process and consequent matching results given a selected similarity criterion. Rather than comparing different similarity metrics, we propose a new search strategy that minimizes the global mismatch errors after a certain similarity measure is selected.

2. METHODS

Automatic object matching requires an objective function or a series of functions, the solutions to which lead to matched pairs. This function provides a rule to determine whether two objects should be matched and a search path to find all matched pairs. The variables in this function could be any similarity metrics, such as Euclidean distance or Hausdorff distance, or a combination of a set of measurements. In this section, we will discuss two search strategies in object matching after a similarity metric is selected: the first one is the popular greedy method that aims to always find the possible minimum dissimilarity between paired objects in each step, and the second one is our proposed optimization strategy that intends to minimize the total dissimilarity between all matched objects.

2.1 Matching Objects Using Greedy

Greedy is a simple way to achieve local optimum at each stage. Its essence is to make the optimal choice at each step even in a problem that requires multiple steps to solve. It has been studied in many fields such as operations research and computer science (Wu *et al*., 1990) and widely implemented in many applications. One obvious problem with the greedy algorithm is that an addition of a new item to the solution set may render the solution not optimal and it does not provide a mechanism to remove items already in the solution. For example, if we match two objects incorrectly in a previous step, there is no way to correct that mistake in later stages. Therefore, in a greedy-based algorithm, a mismatch error in any step will result in at least two mistakes because it will make it impossible for the omitted object to be matched to the correct one in a later stage.

Two greedy methods were implemented in MATLAB in our study. Greedy1 adopts a sequential identification and removal procedure: it identifies the closest pair of objects as corresponding counterparts and removes both from the candidate set; then it identifies the closest pair in the remaining objects and removes them, until all objects are matched. Greedy2 is a modified version of greedy1 by adding a random component to the procedure in order to jump out of local optima. It starts with a random object in one dataset and identifies the closest object in the other, followed by the elimination of matched pairs; then it selects another random object and identifies its matched correspondence until the process is finished. This procedure could be repeated as many times as necessary (*e.g.*, 100) and the best result would be the final result.

2.2 Matching Objects Using Optimization

In order to rectify mistakes introduced in previous stages in a greedy algorithm, we propose another strategy to rely on a global measurement of similarity by regarding object matching as an assignment problem that takes into account all corresponding pairs of objects simultaneously. The search for corresponding objects is based on minimization of dissimilarity between matched objects and can be formulated as the following objective function:

$$Minimize \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} x_{ij} \quad (1)$$

where i = index for the objects in the first dataset
j = index for the objects in the second dataset
n = the number of objects in each dataset
c_{ij} = the dissimilarity between object i in one dataset and object j in the other. c_{ij} could be any form of similarity measures or any combination of multiple metrics that jointly decide the similarity between two objects
x_{ij} = a Boolean indicator: when object i in the first dataset and object j in the second dataset are matched, it is assigned to 1, and assigned to 0 otherwise

The constraints for this objective function are as follows:

$$\sum_{j=1}^{n} x_{ij} = 1, \quad \forall i \tag{2}$$

$$\sum_{i=1}^{n} x_{ij} = 1, \quad \forall j \tag{3}$$

These two constraints ensure that every object in each dataset is matched to exactly one object in the other dataset.

This form of objective function is well known as the assignment problem in the operations research. It is generalized from the problem of assigning a set of tasks to a group of agents with the objective to minimize the total cost of performing all tasks, under the constraints that each task can only be assigned to one agent, and each agent can only accept one task (Hillier and Lieberman, 2001). Our task in object matching is to assign each object in one dataset to its corresponding counterpart in the other one, satisfying the objective function that minimizes the total dissimilarity between matched pairs.

In real applications, two datasets that represent the same area rarely have the same number of objects, so we relaxed the constraints:

$$\sum_{i}^{m} x_{i,j} <= 1, \forall j \tag{4}$$

$$\sum_{j}^{n} x_{i,j} = 1, \forall i \tag{5}$$

where m = the number of objects in dataset 1
n = the number of objects in dataset 2
$m<=n$

Therefore, each object in the smaller dataset is matched to one object in the other, and some objects in the larger dataset will be identified as having no corresponding pair. This assignment problem was implemented using the GNU MathProg modeling language in the GLPK (GNU Linear Programming Kit) package that provides a platform for solving linear programming problems. The similarity criterion is Euclidean distance in the point datasets, and Hausdorff distance in the polyline datasets.

3. DATA

Two sets of data were used to test the differences between greedy and optimization methods in object matching: hypothetical point datasets and real street network datasets.

3.1 Hypothetical Data

Hypothetical data were generated by a random process. The first set of point data were created by a bivariate point process and the second set of point data were created by the following formula: $x_2 = 0.1+x_1$, $y_2 = 1.1*y_1$, where x_1, y_1 are the coordinates of points in the first set of datasets, and x_2, y_2 are the coordinates of points in the second set of datasets. Within a square area, the number of points varies from 10 to 100 with an interval of 5. Some examples of these datasets are demonstrated in Figure 1.

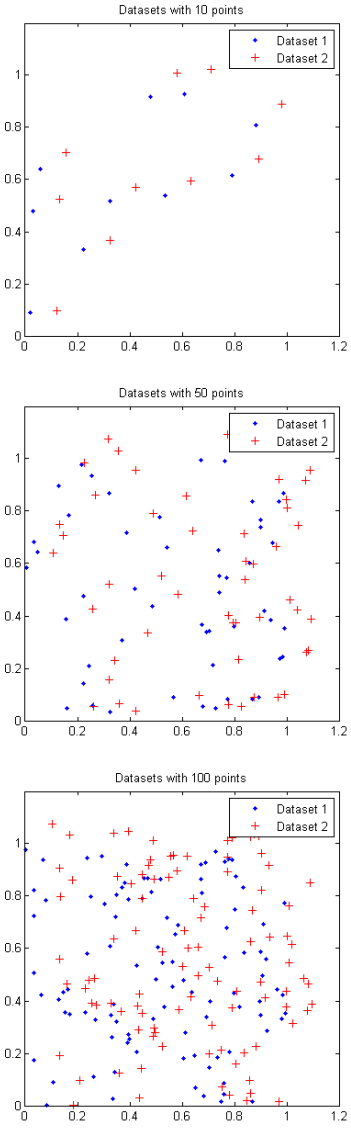

Figure 1. Hypothetical datasets with different numbers of point objects

3.2 Real street data

Real street data are more complex than the hypothetical point data, since they are composed of multiple points and the offsets

between objects are not uniform. In our experiment, street network data in Goleta CA were created under different standards by two agencies. These data represent approximately the same streets in a neighbourhood of Goleta (Figure 2). These two datasets have 236 and 223 objects, respectively. As shown in the figure, there are some discrepancies between these two datasets, and some streets are missing in one version of the data. These data were prepared in a way that they are under the same coordinate system and internally consistent.

Pre-processing was performed in the datasets to maximize 1:1 correspondences, since our optimized object matching strategy is designed for 1:1 matching. Due to the difference in generalization of real streets, the same street may be represented as different numbers of segments. For example, the street Hollister could be described as 5 segments (objects) in one dataset, and as 7 segments (objects) in the other. Therefore, it is helpful to make as many pairs of 1:1 correspondences as possible. In our experiment, we merged street segments based on the name attribute and the topology of polylines. In each dataset, if multiple street segments have the same name and they are connected, they are merged to form one object after pre-processing.

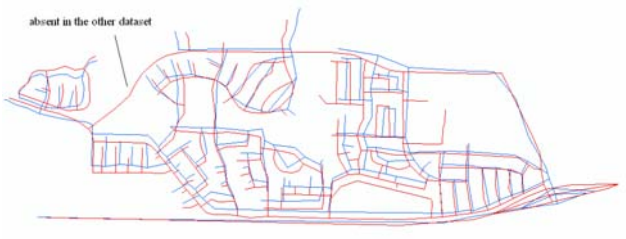

Figure 2. Street networks in a neighborhood of Goleta, CA.

4. RESULTS AND DISCUSSION

Both greedy and optimization methods were tested in these datasets. The sum of distances between matched objects using each of the three methods is displayed in Figure 3. When the number of points is small, the total distances calculated from different methods are similar. As the density of points becomes larger, the difference of total distance becomes more obvious between greedy and optimization methods, but the results are relatively close between the two greedy methods. In any dataset, the total distance of matched pairs is consistently smaller using the optimization method. In Figure 4, the relationship between the percentage of correctly matched pairs and the number of points is displayed. The trend shows that there is a drastic drop in the percentage of correct matches using the two greedy methods as the number of points becomes larger. However, the percentage of correct matches using the optimization method is stable and robust in all tested datasets. While the percentage of correct matches decreases from 100% or 80% to less than 20% using the greedy methods, the percentage of the optimization method maintains at a level close to 100% even in dense datasets. Therefore, when the density of points gets larger, the probability of mismatch becomes larger, and consequently the superiority of the optimization method becomes more obvious.

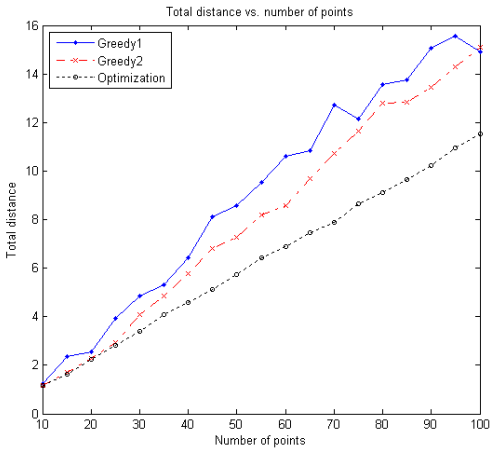

Figure 3. Total distance of matched pairs.

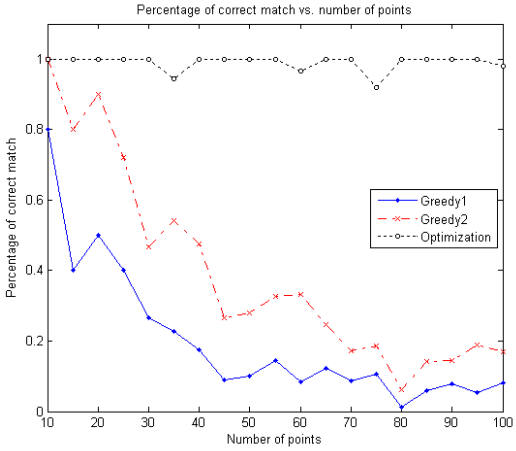

Figure 4. Percentage of correct matches.

The results of object matching in real street data using the three methods are demonstrated in Table 1. The total distance between matched pairs is smaller using the optimization method than using greedy methods. As a result, the percentage of correct matches using the optimization method is about 10% higher than that using the optimization method.

Table 1. Results of object matching for street datasets

	Total distance	Percentage of correct match
Greedy1	13104	88.14%
Greedy2	13078	88.56%
Optimization	12369	97.03%

In all experiments, either with hypothetical point data or real polyline data, object matching using the optimization method consistently achieves better results. When the density of a dataset increases, the probability of mismatch becomes larger, and consequently, the advantage of the optimization method becomes more obvious. While a denser dataset makes object matching more susceptible to mismatches, the spatial arrangement of objects within the study area is also another important factor that affects the matching result. These experiments indicate that the optimization method for object matching is more robust than greedy methods. In some datasets, object matching using a greedy method may also result in a good percentage of correct matches, but in other cases, the

percentage could be not acceptable. Since it requires a lot of time and labour to identify and correct even a small number of mismatches, it is important to maximize the percentage of correct matches in the automatic stage of object matching.

In terms of the choice of similarity measurement in our experiments, when the total distance of matched pairs is small, the percentage of correctly matched objects is high. Therefore, Euclidean distance and Hausdorff distance are proper indicators of point and linear object similarity in these datasets, respectively. However, when more attributes are available, not only relying on the geometric distance in a geographical space, we can also construct a similarity space according to a weighted combination of these properties, and use that metric as a similarity measurement in our objective functions. Furthermore, additional attributes may also be used to reduce search space in particular applications. Although the emphasis of this paper is not the selection of similarity measurement, a proper similarity metric is a necessity for effective and efficient object matching. A measurement that is an adequate indicator of the likeness between two objects should be included in the objective function.

5. CONCLUSIONS

Object matching is a fundamental problem in spatial data handling and many related applications. How to identify objects in different data sources that represent the same entity in reality is a prerequisite for data manipulation and analyses in later stages, such as accuracy improvement, change detection, and geospatial analysis using multi-source data. There are two major components in the object matching process: selection of an appropriate similarity measurement and identification of matched objects according to this measurement. Most existing literature has focused on the definition of a proper similarity metric in particular applications. They usually adopt a sequential procedure to find object pairs one after another based on the chosen metric. In our paper, we focus on the other aspect of the problem: how to effectively search for corresponding objects once a similarity measurement is chosen. Rather than using a greedy strategy that consecutively adds more matched pairs into the solution set, and never removes any mismatched pairs from the solution, our optimized object matching takes into account all possible matched objects simultaneously with the aim to minimize the total dissimilarity between all corresponding objects.

Therefore, object matching is formulated as an assignment problem that intends to assign each object in one dataset to an object in the other dataset, with the objective to minimize the sum of dissimilarity between object pairs. Unlike the widely used greedy procedure for finding matched pairs, this strategy makes it possible to rectify mismatch errors made in early steps. Although only point and polyline data were tested in this paper, this method can also be applied to other types of data as long as the selected metric is adequately representative of the resemblance between objects. Our experiments demonstrate that optimized object matching method is robust and always achieves a higher percentage of correctly matched pairs in both hypothetical and real datasets.

Although our research points out a new research direction in object matching, there are some limitations. First, formulation of object matching as an assignment problem entails the constraints that one object can only be assigned to one or none object in the other dataset. Therefore, this strategy is appropriate for 1:1 correspondence. In real applications, there are cases when an object in one dataset is represented as several parts in the other dataset (1:*n* correspondence), or several objects are corresponding to a different number of objects (*m*:*n* correspondence). Therefore, one of our future research questions is to find a way to maximize the 1:1 correspondence in different datasets before the execution of the optimized object matching strategy. Another problem we are going to investigate is to directly tackle the 1:*n* and *m*:*n* relationships by examining partial similarity between objects. Finally, as the input datasets become larger, the matching procedure may degrade rapidly, and makes it difficult to finish matching within a reasonable time frame. Therefore, we will study the improvement of the algorithm using heuristics to reduce the search space, such as divide-and-conquer technique (Preparata and Shamos, 1985).

References

Abbas, I., 1994. Base de données vectorielles et erreur cartographique: problèmes posés par le contrôle ponctuel; une méthode alternative fondée sur la distance de Hausdorf. *Computer Science*. Paris, Université de Paris VII.

Bel Hadj Ali, A., 1997. Appariement geometrique des objets géographiques et étude des indicateurs de qualité. Saint-Mandé (Paris), Laboratoire COGIT.

Cobb, M. A., Chung, M. J., Foley III, H., Petry, F.E. and Shaw, K.B., 1998. A rule-based approach for the conflation of attributed vector data. *Geoinformatica*, **2**(1), pp. 7-35.

Cohen, W., Ravikumar, P. and Fienberg, S. E., 2003. A comparison of string distance metrics for name-matching tasks. *IJCAI-2003*.

Devogele, T., 2002. A new merging process for data integration based on the discrete Frechet distance. In: *Advances in Spatial Data Handling*. D. Richardson and P. van Oosterom. New York, Springer Verlag: pp. 167-181.

Filin, S. and Doytsher, Y., 2000. The detection of corresponding objects in a linear-based map conflation. *Surveying and Land Information Systems*, **60**(2), pp. 117-128.

Hastrings, J. T., 2008. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, **22**(10), pp. 1109-1127.

Hillier, F. S. and Lieberman, G. J., 2004. *Introduction to Operations Research* (McGraw-Hill).

Preparata, F. P. and Shamos, M. I., 1985. *Computational Geometry: An Introduction* (New York, NY: Springer-Verlag New York, Inc.).

Quddus, M., Ochieng, W., Zhao, L. and Noland, R., 2003. A general map matching algorithm for transport telematics applications. *GPS Solutions*, **7**(3), pp. 157-167.

Saalfeld, A., 1988. Conflation automated map compilation. *International Journal of Geographical Information Systems*, **2**(3), pp. 217-228.

Samal, A., Seth, S. and Cueto, K., 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, **18**(5), pp. 459-489.

Walter, V. and Fritsch, D, 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, **13,** pp. 445-473.

Wu, S., Manber, U., Myers, G. and Miller, W., 1990. An *O*(*NP*) Sequence comparison algorithm. *Information Processing Letters*, **35**, pp. 317-323.

Yuan, S. and Tao, C., 1999. Development of conflation components. *The Proceedings of Geoinformatics'99 Conference* (Ann Arbor).