Address Matching: An Expert System and Decision Support Application for GIS

Peter Y. Wu
wu@rmu.edu
Dept of Computer & Information Systems, Robert Morris University
Pittsburgh, Pennsylvania 15108, USA

Eugene Rathswohl
gener@sandiego.edu
School of Business Administration, University of San Diego
San Deigo, California 92110, USA

Abstract

Address matching is the function to convert textual address description into a map location. It is often incorporated into the geographic information system (GIS) because of the need to do address matching in GIS projects for business today. It is a commonly needed task in community mapping for analysis of market situation and identification of opportunities as well. While GIS courses are becoming more common in the IS curriculum, the topic of address matching is rarely covered. We discuss how the topic may be covered, and suggest its inclusion into the GIS course syllabus. The paper explains the basics of address matching: find the street center line segment and perform ratio of division to extrapolate the portended location on the street segment. However, correctly interpreting written textual address can be a daunting task, requiring human intelligence to handle cultural habits, ambiguities, and mistakes. The paper describes a framework of using an expert system to batch process a large collection of addresses, and how it also turns into a decision support system to manually discover mistakes to correct unmatched addresses. Use of GIS is multi-disciplinary; a generic model of how address matching works should be helpful for its teaching and learning. Hands-on experience of doing the work will be imperative for the student.

Keywords: geographic information system, GIS, address matching, geocoding, expert system, decision support.

1. INTRODUCTION

The geographic information system (GIS) has emerged as a powerful tool for the intelligent use of information, and a viable tool in business decision making (Gewin, 2004; Douglas, 2008). GIS courses also begin to become popular in the IS curriculum. (Wu & Kohun, 2005; Sinton & Lund, 2007). Granted that GIS skills are very much multi-disciplinary, different GIS courses offered in different schools do have widely different emphases (Glover, 2005; Reames 2005; Wu 2007). It is quite necessary to continue the discussion about what to cover in the GIS course. This paper focuses on the special topic of address matching, and discusses the issues about address matching and its inclusion into the GIS course syllabus.

Section 2 will discuss address matching in the business use of GIS. The function is needed

so often that the topic must be in the GIS course syllabus. Section 3 then explains the address matching basics: how it is done with the digital map. Sections 4 and 5 then go on to discuss the problem of processing a large collection of addresses. Section 4 presents the framework in which an expert system is used. Section 5 deals with interactively re-matching the unmatched addresses, using the expert system for decision support. Section 6 discusses the related issues of having address matching in the GIS course design, and the prerequisite knowledge students may need. To close, section 7 presents the summary of the paper.

2. ADDRESS MATCHING IN BUSINESS

Address matching is the function to convert textual address description into a location on the map. GIS applications in business very often need to use the function. Take for example the contact addresses of attendees in a conference (such as ISECON/CONISAR). They may represent the target customers for a certain market, and we may like to be able to visualize where they are gathered or spread. The address matching function will convert the information into a format ready for presentation in a map. Consider another example. Suppose we want to select an appropriate site for our gourmet coffee shop to start a new business. Since all other coffee shops in the vicinity will be our competitors, we would like to gather their addresses and plot them on a map to help our site selection.

ORG_NAME	ADDRESS	CITY	ST	ZIP	SICDESC
Steven T Hesky Ph.D.	1 Allegheny Ctr St 218	Pittsburg	PA	15212	INDIVIDUAL AND FAMILY
Doorway Inc.	10 California Ave	Pittsburg	PA	15202	INDIVIDUAL AND FAMILY
St Clair Hosp Lfeline Program	1000 Bower Hill Rd	Pittsburg	PA	15243	OUTREACH PROGRAM
Mercy Hospital	1004 Arch St	Pittsburg	PA	15212	INDIVIDUAL AND FAMILY
Northco Cmnty Employment Ctr	104 Beta Dr	Pittsburg	PA	15238	JOB TRAINING SERVICES
La Societe De Femme	107 Cresthaven Ln	Pittsburg	PA	15237	SOCIAL SERVICES, NEC
Paula Teacher & Associates	10700 Frnkstwn Rd 310	Pittsburg	PA	15235	JOB TRAINING AND RELA
Lullaby Day Care Center	10985 Frankstown Rd	Pittsburg	PA	15235	CHILD DAY CARE SERVIC
Southwinds Inc. M H M R	1121 Boyce Rd Ste 2000	Pittsburg	PA	15241	RETARDED HOME
Ymca Pittsburgh-OHara	115 Cabin Ln	Pittsburg	PA	15238	CHILD DAY CARE SERVIC
North Avenue Daycare	116 East North Ave 118	Pittsburg	PA	15212	INDIVIDUAL AND FAMILY
Don Mollenauer Memorial Fund	1204 Lindendale Dr	Pittsburg	PA	15228	FUND RAISING ORGANIZ
Just For Kids Learning Center	1220 Powers Run Rd	Pittsburg	PA	15238	CHILD DAY CARE SERVIC
Temple Emanuel Nursery School	1250 Bower Hill Rd	Pittsburg	PA	15243	PRESCHOOL CENTER
Mc Donough Lisa M Msw	128 North Craig St	Pittsburg	PA	15213	INDIVIDUAL AND FAMILY
Elaine S Portner	128 North Craig St	Pittsburg	PA	15213	INDIVIDUAL AND FAMILY
Velletri Andrea S Ph.D.	128 North Craig St Ste 21	Pittsburg	PA	15213	GENERAL COUNSELING
Love & Learn Childrens Center	1300 Brinton Rd	Pittsburg	PA	15221	GROUP DAY CARE CENT

Figure 1. Service Agencies and Addresses

Figure 1 shows of the database table of social service agencies for senior citizens along with the addresses in Allegheny County of Pennsylvania. We may want to analyze how well the agencies may be serving our elderly population in the region. Address matching can generate the point locations of the agencies in a map layer.

Figure 2 illustrates the map plotted along with senior population percentage by census tracts published by US Census Bureau for year 2000. The choropleth map has each of the tracts in shades of gray to symbolize the percentage of population 65 or above, at 0-12%, 12-18%, 18-24%, and above 24% within the tract, respectively. We can then visualize the situation on the map, and we are ready for further analysis using the GIS.

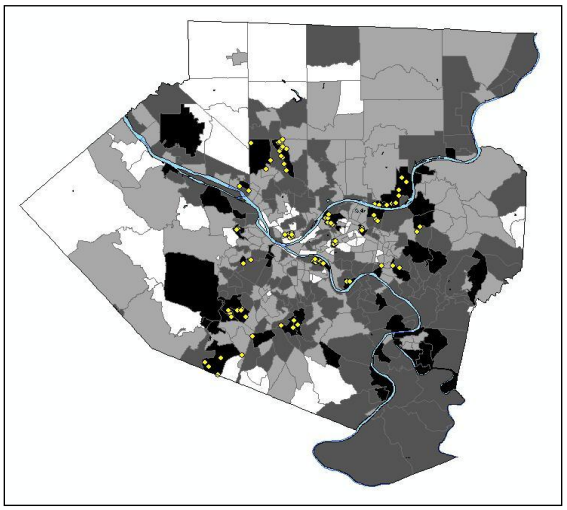

Figure 2. Senior Population by Census Tracts along with Social Service Agencies

Address matching is very often a necessary first step in most GIS projects for business purpose in mapping the community. While it constitutes a special topic in its own right, we propose to include the topic in the GIS course for the IS curriculum.

3. ADDRESS GEOCODING BASICS

Geocoding is the processing of textual coded information to determine the location on a map. The textual description may be global latitude and longitude values, coordinates in the state plane, or more commonly used in business, postal addresses. Geocoding postal address is also called address matching.

Address matching requires a reference map with information to look up for an address. For US postal addresses, the Census Bureau publishes these reference maps in the TIGER map format (Marx, 1986). We can download digital maps of street center lines from the

Census Bureau website (www.census.gov) for anywhere in the United States.

Each line segment of a street center line map carries with it the attribute information for address look up. The following lists the fields for relevant information.

- Street Name
- Street Type (such as St, Ave, Blvd...)
- Prefix Direction (such as East, West…)
- Suffix Direction (such as North, Ext…)
- From-Number on Left
- From-Number on Right
- To-Number on Left
- To-Number on Right
- Zip Code on Left
- Zip Code on Right

A simple select query to the relational table of street center line segments can determine the line segment of the map location can then indicate the specific location within the street line segment by the ratio of division from both ends. Figure 3 below illustrates the case of address 333 Sample Street. The segment of Sample Street at the 300 block has From-Number 301 and To-Number 399 on the Left. The location for 333 is therefore one third of the way on the left side. The Zip Codes should verify, if provided.

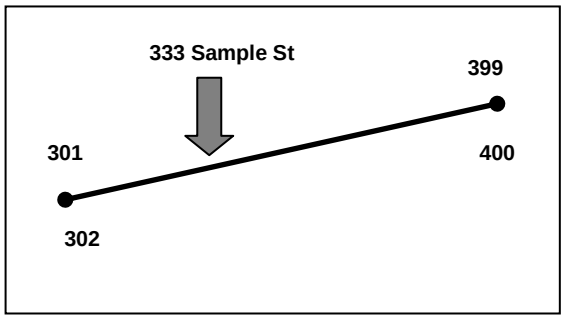

Figure 3. Ratio of Division applied to the Street Line Segment for 333 Sample St.

However, address matching is complicated by the different formats an address may be recorded, including cultural habits and human mistakes. For example, Fifth Avenue may be written as Fifth Av, or 5th Ave. The address of 1 Montgomery Place is also often written as One Montgomery Pl. The post office can sometimes handle incorrect zip code when a letter is delivered to the correct city, but the GIS will need an expert system to deal with a large collection of addresses contaminated with mishaps and mistakes.

The generic solution to the problem is rooted in the mathematics and computer science of pattern recognition (Freeman, 1996), and the theory of unification search (Wolff, 2003; Wang, 2009). In terms of bringing the science into practicable solution, the seminal work of ChoiceMaker Technologies, Inc. (www.ChoiceMaker.com) is notable (Buechi, et al, 2003; Goldberg and Borthwick, 2004). Specific to the GIS setting, there were also studies on searching for matches in digital maps. (McElroy, et al, 2003; Pavia, 2005) In this paper, we will focus on using available solutions for GIS projects and the framework for teaching.

4. BATCH GEOCODING EXPERT SYSTEM

GIS today is usually bundled with the expert system tool to do address matching. While different vendor systems may be different, we attempt to sketch the generic framework based on how ArcGIS, the GIS software product from Environmental Science Research Institute (www.esri.com).

Since ArcGIS is prevailing GIS product in use in many different countries, the address matching expert system is designed to deal with many different postal address formats. In order to manage and re-use the detailed setup of parameters for the address matching expert system, ArcGIS gathers together the setup data into an Address Locator file. To set up for address matching, we create a new Address Locator file. In the file, we specify the map layer of street center line segments used as the reference map, and we need to specify the actual field names for the fields of information relevant to address matching as explained in the previous section. Since we also want to tolerate human mistakes such as spelling and others, the parameters include the tolerance level of spelling mistakes and the acceptable level of matching scores. For output data, that is, the location computed by ratio of division in the street line segment, we may specify the desired limit of accuracy. A good system should be designed so that these parameters will default to appropriate values for users not yet familiar with the tool.

Once we set up an Address Locator file, we may apply it to do address matching on any table of addresses brought into the GIS. (Hence we may re-use the Address Locator.) The process will generate a map layer of point locations, each of which corresponds to

an address in the table, if an acceptable match is found. Figure 4 shows the report pop-up window from ArcGIS on the address matching performance: how many matches found, etc.

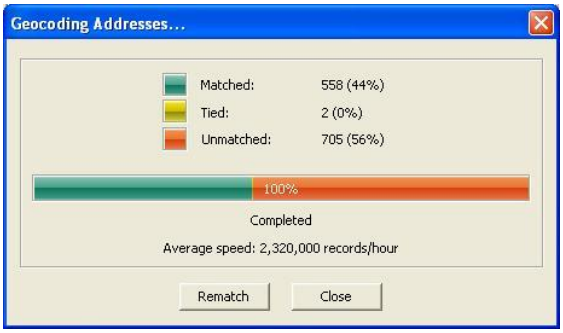

Figure 4. Batch Process Performance Report of Address Matching in ArcGIS

The process leaves us with the problem of unmatched addresses. The next section will discuss how we may perhaps resolve some of the unmatched addresses, manually, using the expert system for decision support.

5. RE-MATCHING INTERACTIVELY

Unfortunately, only human intelligence will resolve the arbitrary nature of human mistakes. We will have to handle the unmatched addresses one at a time. Fortunately, the expert system may provide us with helpful assistance. The address matching tool in ArcGIS provides an Address Re-match panel to interactively handle re-matching of unmatched addresses. Figure 5 in the appendix shows the panel, which incidentally we can bring up by clicking on the re-match button as shown in figure 4 above.

In ArcGIS, the address re-match panel brings up all the addresses in a table, which one can manipulate similar to that of a spread sheet. We can then find the unmatched addresses and examine each one of them. Once an unmatched address is selected, the system then finds many potential matches and lists them in the lower pane there. We can visually detect what is wrong conveniently. The panel also provides for us to edit any data field when we may identify any possible human error in the data. Thus, the address matching expert system provides decision support in the inter-active re-matching of unmatched addresses.

6. DISCUSSION

Address matching certainly high-lights the multi-disciplinary nature of GIS use. For its importance in the business application of GIS, we recommend its inclusion in the GIS course syllabus.

Although the theory behind address matching requires some understanding of coordinate geometry, a full course treatment as pre-requisite should not be necessary. A good illustration as explained in Section 3 would be sufficient.

It is then more important for the students to understand the complications due to different address formats and style, including human errors in recording addresses. The science of record matching is beyond the scope of the GIS course, but the use of an expert system tool for the purpose certainly makes it very interesting. The student does need to understand the relevant parameters necessary to set up and control the. Pre-requisite knowledge of expert systems may be helpful. Yet we believe it is more constructive for the student to build a conceptual model of the system to think critically about the design and how the tool works. In terms of pedagogy, it will be interesting further work to develop the lesson plan based on Vygotsky's theory of instruction in cultural context – specifically, the scaffolding zone of proximal development (Chaiklin, 2003).

In the case of using ArcGIS, the critical part is for the IS/IT student to understand the need for the Address Locator file, and how the file may be re-used. The interactive re-matching of addresses resembles that of working with a spread sheet. All the more, the student will need hands-on experience of actually using the system to do the work.

It is therefore our recommendation for the GIS course in the IS curriculum to include address matching as a special topic. We should discuss the theory for a system model to facilitate critical thinking, but hands-on experience will be imperative for the student.

7. SUMMARY

Address matching is an important function in the business application of GIS. It would be essential to include it as a special topic in the GIS course for the IS curriculum. The basics of address matching involve the division of line segment to interpret the map location of address, but the student does not have to be

well versed in coordinate geometry. All we need may be a simple illustration. Rather, it is more important to understand the problem of dealing with different address formats and human mistakes. Using an expert system tool can become a challenge for the beginner. A framework sketch on how to set up the parameters to control the expert system in batch processing will be necessary. The student's prerequisite knowledge of expert system can be helpful, but actual hands-on experience of doing the work is imperative in the course. The batch process of address matching using an expert system tool may leave us with some unmatched addresses. We will need to re-match them one at a time. The address matching expert system can also provide appropriate help for the user to do this interactively. Since this may be quite different for different vendor systems, the hands-on experience for the students all the more must be required. This certainly highlights the multi-disciplinary nature of GIS applications.

8. REFERENCES

Buechi, M., Borthwick, A., Winkel, A, Goldberg, A. (2003) "ClueMaker: A Language for Approximate Record Matching," Proceedings of the 8th International Conference on Information Quality, Cambridge, MA. Retrieved Sept 15, 2010 from http://wotan.liu.edu/docis/dbl/iqiqiq/2003__207_CALFAR.htm

Chaiklin, Seth. (2003) "The Zone of Proximal Development in Vygotsky's Analysis of Learning and Instruction." Vygotsky's Educational Theory in Cultural Context, Alex Kozulin (eds, et al), Cambridge University Press, pp.39-64.

Douglas, Bruce (2008) Achieving Business Success with GIS, John Wiley & Sons.

Freeman, Herbert, eds. (1996) Studies in Pattern Recognition, World Scientific Publishing Co. Pte. Ltd.

Gewin, V. (2004) "Mapping Opportunities." Nature, Vol.427, pp.376-377.

Glover, B. (2005), "Curriculum Update for GIS Programs in Texas Community Colleges," 5th Annual ESRI Education User Conference, July 2005, San Diego, CA.

Goldberg, A. and Borthwick, A. (2004) "The ChoiceMaker 2 Record Matching System," ChoiceMaker Technologies, Inc., Nov 2004. Retrieved Sept 16, 2010 from http://www.cs.nyu.edu/cs/faculty/artg/publications/goldberg_borthwick_The_ChoiceMaker_2_Record_Matching_System_2007.pdf

Marx, Robert W. (1986) "The TIGER system: automating the geographical structure of the United States Census," Government Publications Review, Vol.13, Issue 2, March-April 1986, pp.181-201.

McElroy, J.A., Remington, P.L., Trentham-Dietz, A.T., Robert, S.A., Newcomb, P.A. (2003) "Geocoding Addresses from a Large Population-Based Study: Lessons Learned." Epidermology, 14(4), pp.299-407. Retrieved July 4, 2010 from http://www.jstor.org/stable/3703788

Pavia, Isidor. (2005) Interactive Querying of Topolographic Maps, Report K268, Dublin Institute of Technology, 2005.

Reames, S. (2005) "Business Geographic Information Systems - A Course in Business Geomapping," *Proceedings of ISECON 2005 v.22* (Columbus OH): § 2334.

Sinton, D.S. and J.L. Lund. (2007) Understanding Place: GIS and Mapping Across the Curriculum, ESRI Press, Redlands, CA.

Wang, ShuXi. (2009) "A New Algorithm for Pattern Matching and Unification," Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China, pp.405-409.

Wolff, J.G. (2003) "Information Compression by Multiple Alignment, Unification and Search as a Unifying Principle in Computing and Cognition," Artificial Intelligence Review 19(3), pp.193-230.

Wu, Peter Y. & Kohun, Frederick G. (2005) "Designing Geographic Information System Courses in the IS Curriculum," *Proceedings of ISECON 2005 v.22* (Columbus OH): §2564.

Wu, Peter Y. (2007) "Introducing Geographic Information Systems into the IS Curriculum: GIS Tutorial and Preparation Workshop," *Proceedings of ISECON 2007 v.24* (Pittsburgh PA): §3732.

9. APPENDIX

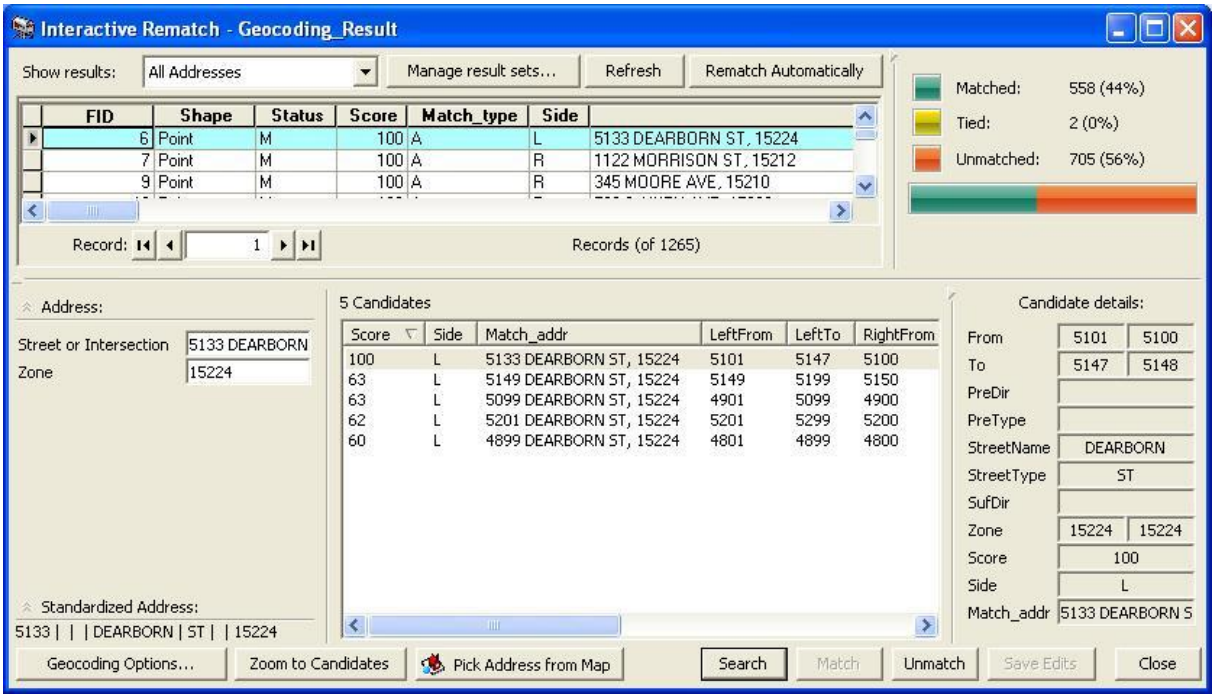

Figure 5. Address Re-Match Panel in ArcGIS