July 2008

GEOGRAPHICAL AND TEMPORAL VISUALISATION OF SOCIAL RELATIONSHIPS

Majigsuren Enkhsaikhan
The University of Western Australia, majigaa@csse.uwa.edu.au

Wei Liu
The University of Western Australia, wei@csse.uwa.edu.au

Mark Reynolds
The University of Western Australia, mark@csse.uwa.edu.au

GEOGRAPHICAL AND TEMPORAL VISUALISATION OF SOCIAL RELATIONSHIPS

Majigsuren Enkhsaikhan, University of Western Australia, Crawley, Perth, WA 6009, Australia, majigaa@csse.uwa.edu.au

Wei Liu, University of Western Australia, Crawley, Perth, WA 6009, Australia, wei@csse.uwa.edu.au

Mark Reynolds, University of Western Australia, Crawley, Perth, WA 6009, Australia, mark@csse.uwa.edu.au

Abstract

Knowledge discovery from large data collection is increasingly important for knowledge-intensive information systems. In addition, effective visualisation are vital for understanding the knowledge embedded in the data. This paper aims to identify specific named entities from structured text content and visualise them in terms of social relations and geographical locations. The system presented here retrieves author information from publication data, disambiguates people names and creates a graph that visualises the co-author connections in order to build co-author networks for a specified topic in a given time period. Each co-author connection at a different time is shown using a coloured line to effectively visualise the co-author relations over time. The system also retrieves addresses of organizations for authors and displays them on a geographical map. Thus, the places that concentrate on a specific research topic for a given time period are geographically identfied. The paper presents an effective way to discover and visualise social networks of authors and geographical locations of researchers on a particular topic.

Keywords: Social network analysis, Social network visualisation, Data sharing.

1 INTRODUCTION

Knowledge retrieval from large data collection is extremely important. In addition, effective representation of the retrieved information is essential for understanding the knowledge. “Graphical excellence is what gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space” (Tufte 1983). Effective representation makes retrieving and using information easier when the form of the information sought cognitively fit with the form of the information presented (Petre et al, 1997). Discovering the evolution of social relationships through time and across different geographical regions is essential to many information systems that support decision makings. For example, analysing data related to the spreading of disease and the tracking of criminal activities relies heavily upon geographical and temporal information. Visualising these kinds of data using social networks and geospatial frontend is critical to help tracking the development process and the relations among people.

This paper uses publication data as an example to show how the visualisation of geographical and temporal information can improve the understanding of social relationship between co-authors. Findings in this paper can be generalised to other similar applications. For instance, in the case of the spreading of pandemic influenza (e.g., Bird Flu), visualising the victims and their social relationship as well as the geographical location can provide vital information. In our example, we retrieve and process authors' information, create social networks and visualize them with respect to time and geographical aspects, which makes hidden information apparent and visible. This allows users to find authors and co-authors of the scientific publications sought and visualise geographical locations of the places of publication using digital library data. The implementation includes tasks that retrieve and disambiguate authors' information from the *Citeseer*[1] metadata, build co-author networks and visualise the networks and geographical locations using *Google Maps API*[2]. CiteSeer is a scientific literature digital library that covers 767,558 documents primarily in the area of computer and information science. It provides metadata for all of its articles and is used primarily for searching publications. In this paper, we use it to analyse social networks of authors. Author networks help to identify key researchers for a specified topic given certain time period. Google Maps provide free web mapping services that include street maps, a route planner, and an urban business locator for numerous countries around the world. Google Maps Javascript API allows developers to embed Google Maps into their own web applications.

The paper is organized into five sections. Section 2 discusses related work. Section 3 explains the system architecture and the purpose of each system component. Section 4 includes experiments and discussions on the results of our system at work. The paper concludes in Section 5 with a discussion on further work.

2 RELATEDWORK

2.1 Social Network Analysis

Social network analysis (SNA) provides an analytical paradigm for defining, mapping and measuring important concepts and structured social relationships. The nodes in the network are the people, groups or objects, while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of the relationships between nodes.

Main concepts in SNA include actor, relational tie, dyad, triad, subgroup, group, relation and network (Wasserman and Faust, 1994). *Actors* are individuals, organizations or a group of individuals.

[1] http://citeseer.ist.psu.edu/
[2] http://www.google.com/apis/maps/documentation/

Relational ties are links between actors: for instance, evaluation of an actor by another, transfer of resources, association of affiliation, behavioral interaction, physical movement, physical connection, formal or biological relation. A *dyad* consists of two actors and a tie between them. *Dyadic analysis* studies the properties of the relationships for a single tie. A *triad* consists of three actors and connections among them. A *subgroup* consists of actors who are all tied among them as any subset of common actors. A *group* is a collection of all actors on which ties can be measured. Modeling finite group concerns network boundaries, sampling, and the definition of a group. Relation is the collection of ties of specific actors. A social network consists of a finite set of actors and the relations defined on them.

There are three popular individual centrality measures: *Degree centrality, Betweenness centrality*, and *Closeness centrality*. *Degree centrality* refers to the number of direct connections for a single node. *Betweenness centrality* refers to the node who has one of the best locations in the network. This person plays a *broker* role in the network, but can become a single point of failure. Without this person, some people would be cut off from information and knowledge from this broker person's side. A node with high betweenness has great influence over network flows and holds power over the outcomes in the network, even if the person has not got a high degree centrality. *Closeness centrality* refers to the direct and indirect ties that are able to access all the nodes in the network quicker than anyone else. They are the nodes with shortest paths to everyone. They can monitor the information flow in the network and have the best visibility over the network.

Centralities of all nodes can reveal the overall network structure. A centralized network is dominated by one or a few central nodes. If they are removed, the network falls into unconnected sub-networks. A less centralized network allows the nodes to reach each other over other network paths, when some connections fail.

There are works that analyse the social networks in order to build online social networks for communities of people to share or explore interests and activities from the bibliographical data. Social networks in bibliographic data have been investigated in several applications that extract co-authors, referenced authors, and similar authors from literature library. Chan et al. (2006) visualised author information by clustering the author groups into social networks by investigating the authors' research areas and co-author groups. Klink et al. (2006) presented the user interface that helps searching authors and publications and analyses social networks from the bibliographical data. They implemented a social network analyser to find co-authors, co-authors of co-authors, similar area authors, authors within same and similar conference, workshop or journals, and similar authors.

Web-based social network tools provide various ways for user interaction including text or multimedia chat, messaging, email, file sharing, blogging and online discussion. For example, Flink (http://flink.semanticweb.org/) for the Semantic Web community and Babble for an online conversation system (Erickson and Laff, 2001).

Lauw et al. (2005) proposed a model for constructing a social network from events and mining these events from the given data. They tried to track people to find out their spatio-temporal co-occurrences in cyber locations. These kind of systems can be useful for law enforcement organizations to investigate collaborations among criminals, for businesses to exploit relationships to sell products and services, for healthcare professionals to manage institutional knowledge, disseminate peer to peer knowledge and to highlight individual physicians and institutions or for individuals who wants to network with others.

In this paper, we focus on co-author relationships on a specified topic over time.

2.2 Data Mining and Text Mining

Data Mining explores and analyses large quantities of observational data in order to discover meaningful patterns and models. Data mining mostly processes highly structured numeric data, while *Text Mining* processes mostly unstructured text written in natural language format. While data mining prefers data in spreadsheet or matrix format, text mining sees data in document format (Weiss et al, 2005). Text mining methods usually apply data mining techniques by organizing statistical evidences of the text into spreadsheet based numerical data.

Disambiguation of named entities is one of the text mining problems. In this paper, we discuss about author name disambiguation. Person name disambiguation is required to identify a specific person. Matching name variations is used to determine a person whose name is written in different forms. For example, John K. Smith can be written as J. Smith, Smith, J., J.K. Smith, John Smith or Smith, J.K.. Also different people may share same names. For instance, let us assume that John K. Smith could be a lecturer in one of universities in Australia while another John Smith could be a research assistant in one of research institute in Canada. Therefore, here we used not only person names but also affiliations and addresses to identify an author. Because we give keywords for a particular research area as an input for the system, measuring document similarities is not included here, although this is an effective way of differentiating authors with same names. Similarity metrics including Soft-TFIDF and Jaro-Winkler are used for our purpose. Cohen et al. (2003) compared and evaluated string distance metrics on the task of matching entity names and covered different types of metrics including edit-distance metrics, fast heuristic string comparators, token-based distance metrics and hybrid methods. They concluded that Soft-TFIDF is the best performing method and also Monge-Elkan method and Jaro-Winkler method as surprisingly good edit-distance based metrics.

Visual Data Mining presents the data in a visualisation, allowing human to get more insights of the data, analyse and explore the data interactively (Keim 2002).

A relatively small amount of research to seek communality between textual sources and spatial representation has also been reported. If the geometry data of an object is available, such as the 3D map of a brain, a mechanical device or the World, *Spatial Data Mining* can be performed. “The parcellation of the human brain by combining text mining and spatial data mining within a neuroinformatics database" in (Nielsen 2006) was one of the text-spatial data mining research that uses 3D human brain visualisation and related database to represent the changes in brain activities. “The fault analysis of complex technical devices" project also explored the semiautomatic analysis of database which contained products’ failure-related text documents in order to interactively explore and reveal the problem sources of the complex technical devices from problematic spatial configurations (Gtzelmann et al, 2007). They applied text mining and information retrieval techniques and 3D models of geometric approximations of the real objects.

When the graphic model represents a geographic information system (GIS), the task is called *Geospatial/Geographical Data Mining*. The project for identifying specific health conditions and analysing correlations between text data and geospatial locations is reported in (Petrou 2005). They applied statistical text mining, clustering and data visualising methods to achieve the proposed goal. In this paper we discuss analysing the text content to retrieve geographical data and visualising it with the geographical model through Google Maps API.

3 SYSTEM ARCHITECTURE

Figure 1 shows our system architecture consisting of four main modules: retrieving publication and author information, processing metadata while disambiguating people names and converting data into XML (eXtensible Markup Language) format, defining co-author relationships and visualising geographical and temporal aspects of author information. When a search phrase and a time period are given, the system finds and retrieves metadata of the related documents from a Citeseer metadata file. The document parts for titles, subjects and abstracts, are checked against the key phrase. Then authors, publication date, location and title information are retrieved from structured data and named entities are disambiguated. In short, author information of relevant publications is collected, disambiguation is executed and necessary data is written into an XML file. The file will only include the *document name, author names, institute names, institute addresses* and *date of the publication*.

We visualise the information in two ways: networks of- for co-authorship and a geographical map of publication locations. Firstly, authors' information for each publication is used to define co-author relationships and co-author networks are created and visualised in graphics. The multiple connection lines can be drawn depending on the number of shared works, but the colours of lines will differentiate multiple relationships between the same people. The lines with colours except black, represents the co-author connections created during the given time. Black lines refer to the co-author connections for other times, not in our selected time period. In short, this graph can give an idea of the main authors and their networks for a topic of interest and for the specific time period. An author with more connections of many similar colours, is one with many co-author connections. An author with lines of many colour shows that the author has many publications. Secondly, the publication locations are visualised on the geographical map using the XML data. Google Maps are used as a tool for implementation, which can show us the main physical locations of related publications. By clicking on the marker, the publication information in the XML file, can become available. The user can then see the main research institutes or organizations on the map.

These visualisations can reveal co-author networks developed over time on a selected topic. Therefore, they help to identify main researchers at research area of interest and locations of strong research institutes.

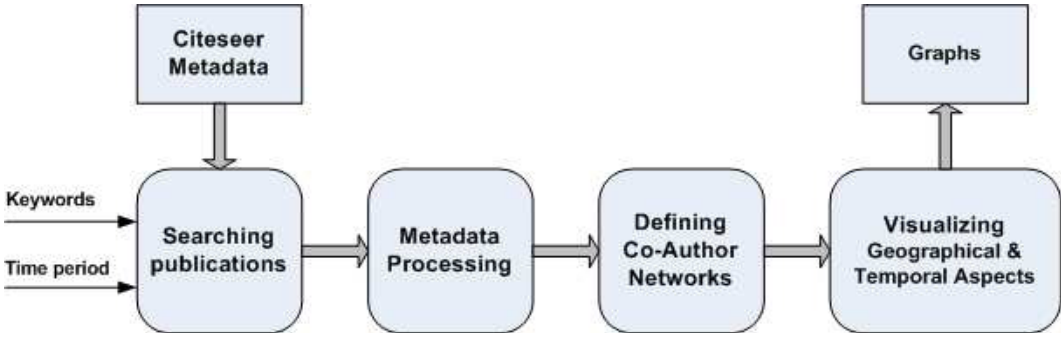

Figure 1. Architecture of the Spatial Text Mining System

3.1 Publication Search

This part of the system searches related documents by finding keywords in the document abstracts, subjects and titles using string pattern matching. The input file contains metadata about scientific publications registered in Citeseer. Topic keywords are used to find the scientific publication metadata

from this file and retrieve the useful information such as document names, published dates, author names, institute names and addresses.

3.2 Metadata Processing

This system component is responsible for disambiguating author names and converting the information that we retrieved from the Citeseer metadata into XML format. In order to determine name variations of the same person, we applied string similarity metrics. We used named entities such as author names and institute names, as well addresses from the publication information. To identify similarity of affiliations and addresses, we applied Jaro-Winkler similarity, which is one of the successful string similarity metrics. Then hybrid similarity metric Soft-TFIDF is used for finding similarity of person names. Figure 2 represents the approach to solve name ambiguity.

```
FOR each name x of List
    FOR each name y of List
        IF similarity(x, y) > φ
            IF similarity(x_affiliation, y_affiliation) > ψ
                IF similarity(x_address, y_address) > μ
                    x and y are name variants;
                END IF
            END IF
        END IF
    END FOR
END FOR
```

Figure 2. Algorithm for Name Disambiguation

XML files are well supported by Google APIs, so they make the geographical visualisation part of the system easier to implement. Creating our own tags to support the application coding was important in our system and using XML files facilitate the data sharing across the different information systems.

3.3 Defining Co-Author Networks

Authors' information is retrieved and co-author networks are generated in this part of the system. Metadata of a publication, which is written by several authors in a specific time period, is used to set co-author relationships between the authors. Information of every publication includes all author names with their affiliations and addresses, so co-authorship relations are created using literature information.

3.4 Visualisation

A graphic image of co-author networks is created to visualise the co-author relationships. Different colour of connection lines are used to identify relationships over time. Connection lines in the selected time period, is marked with different colours except black, but the relations in other time periods are shown with black lines.

We embedded Google Maps services into our web page by using Google Maps Javascript API technology. Data in XML format, of found publications is used for visualizing in the Google Maps. We applied Google's GClientGeocoder to find geographical locations (latitude and longitude) from textual data which is primarily physical addresses of author institutes. Locations of publications are marked on the map. The markers are able to show notes about a publication which was written at that place.

4 EXPERIMENTS AND RESULTS

Our experiments use the bibliographical data, which is taken from Citeseer metadata and covers 6110 scientific publications. The main user interface shown in Figure 3 allows us to search publications by specifying a time period and a search key phrase. When search is done with *Find articles* button, the system finds publications from Citeseer metadata file and present them in the user interface. The interface displays result of the search for a given key phrase and a time period. Button *Show co-authors* creates a graph to show co-author networks. Button *Show places* presents a geographical map and marks geographical locations of the scientific publications using Google Maps API. Because of incomplete or empty address fields on the metadata for some publications, the map may miss some marks for locations of those publications. In order to visualise on the geographical map, the data in XML format is created from the metadata and used in visualisation with the map.

Figure 3. Interface with search results

Some co-author networks are shown in Figure 4. The first triad between three authors is shown in blue colour and there are only single lines between each two authors and all in the same blue colour. It means that this publication written by these people, was published during our selected time period. Also we can say that there is only one publication between them in the part of Citeseer metadata that we selected, since the blue lines are single for every two authors. The next triad is represented in black lines, so that these authors' publications were published in a different time period than our given time. We can see the authors for our topic of interest from this graph. The last network includes eight people and lines of seven different colours. So the network has eight authors and seven or more publications

between them, because black may refer to multiple documents which are published outside of our selected time period. As we can see in blue connections, an author named Paolo Bresciani has cooperated with four authors for a paper: Anna Perini, Paolo Georgini, Fausto Guinchiglia and John Mylopuolis. Author Paolo Georgini is located in the central location and participated in all publications, which means he has a high degree of centrality with many direct lines. He is also a node of high *betweenness centrality* and *closeness centrality*, which indicates a powerful role in terms of social network analysis. Without him, two actors Paolo Bresciani and Anna Perini would be cut off from this network. We can say also, he connects several sub groups. We refer to subgroups as a small number of people and their connection and they work together often. Paolo Georgini, John Mylopuolis and Manuel Kolp have three papers together, so they can be one subgroup. Also Paolo Georgini, Enrico Blanzieri, Fausto Guinchiglia and Claudio Zanoni published three publications together. They can be another subgroup. The main connector of them here is Paolo Georgini, who has high *degree centrality*. These kinds of graphs can give clear view about the network of people and their connections. Therefore, main researchers for the interested research area can be determined from the graph. For instance, according to our graph, Paolo Georgini was an important and strong researcher on *multi-agent* topic during June 2003 and January 2005, but let us remind you that the bibliographic information that we used for our experiment is a small part of Citeseer metadata and covers only 6110 publications.

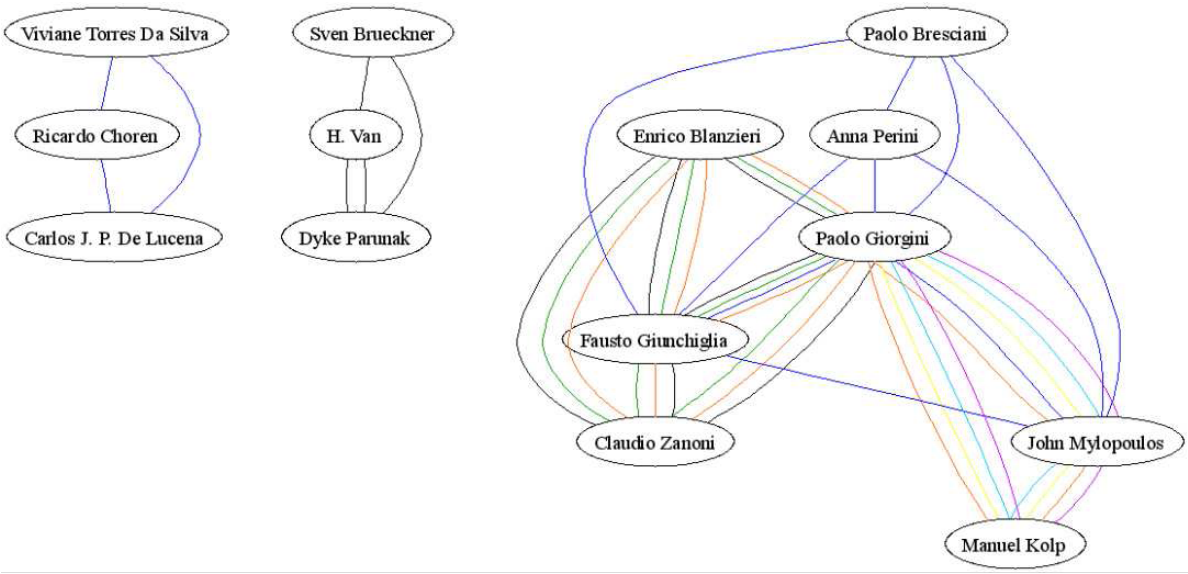

Figure 4. Co-author networks

Figure 5 represents places of the publications. The topic we chose was *Multi-agent* and publications are all related to this topic. In other words, these publications titles, subjects and/or abstracts include the key phrase. The map here can show the places where these publications are written. As we see, USA, Canada, Europe, Australia and New Zealand published some scientific publications on *Multi-agent* topic between June 2003 and January 2005. However, some places may not be marked on the map, because of the inaccurate address data stored in Citeseer database.

Figure 5. Places of found publications on the Google Maps

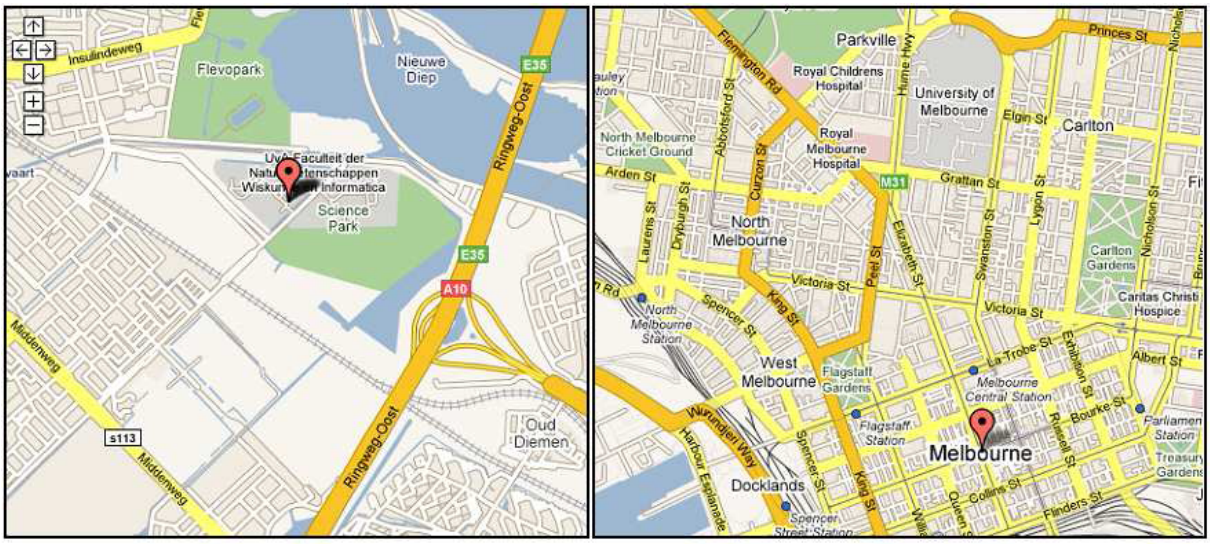

Figure 6. Place marking: (a) Detailed address is given (Kruislaan 403, 1098 SJ Amsterdam, The Netherlands); (b) Address is not detailed (Melbourne , Vic, Australia);

Our system uses data in the address field of Citeseer metadata as geographical locations for the publications. This geographical visualisation part is therefore highly dependent on the accuracy of the address information in Citeseer. Unfortunately, we find that many authors do not insert their addresses properly into the Citeseer database. Because of empty address fields, incorrect addresses, grammatical errors or incomplete address, the geographical visualisation may appear incomplete. As such, incorrect location taggings can happen. Also, depending on the given address, some places are visible to the detailed location, but some just directs to the general place like city or country. For instance, Figure 6 shows two geographical locations of two publications. Figure 6(a) shows very detailed location,

because the address was entered in detail with street number, city name and country: Kruislaan 403, 1098 SJ Amsterdam, The Netherlands. Figure 6(b) points to the city centre, because the address did not include street address, only city, state and country names, for example, Melbourne, Vic, Australia. Therefore, inserting the correct and full addresses is important in this application.

5 CONCLUSION AND FUTURE WORK

Tracking people's activities, disease spreading or object movement is important in many applications. Using social network analysis and geographical visualisation can uncover information behind large quantities of text content in an effective and informative way. This paper discusses visualisation for the scientific publications in two ways: networks of co-authors and physical locations of selected publications. Firstly, the co-authors information is used for creating the co-author networks and visualised in graphics. The connection lines with different colours help to show the collaborative work among authors. This kind of graphs brings clear idea of main authors and their networks for an interested research area during a specific time period. An author with more connections of many similar colours, is one with many co-author connections. An author with connection lines of many colours refers to the author with many publications. Secondly, the locations of found publications are visualized on the Google Maps using the XML data. This can show us the main places such as research institutes or organizations of the certain research have taken place. Using digital literature library for not only searching publications, but also defining important researchers in a given field during a given time period and also showing networks of researchers and geographical locations of them can give important information to people who are doing literature review or searching people on a specific research area.

This system employs the already prepared and rich database Citeseer, but is not restricted to it. A web crawler for collecting publications is under development in order to extend this work. In order to provide rich text data that is populated automatically, we need a web crawler which checks and finds the scientific publications on-line from the World Wide Web. This process needs to check the publication whether it is scientific or not, apply different text mining techniques for retrieving the useful information such as author , location, co-authors and referencing authors. Moreover, most publications do not include its publication date in its text, thus finding the dates is a difficult task and needs to be considered.

This prototype development will lay the foundations of much further development, such that a visualisation tool for geographical text mining can be deployed for general purpose social network evolution analysis.

References

Chan, S., Pon, R. and Cardenas, A. (2006) Visualization and clustering of author social networks. In Distributed Multimedia Systems Conference, pp. 174-180.

Cohen, W., Ravikumar, P. and Fienberg, S. (2003) A Comparison of String Distance Metrics for Name-Matching Tasks. In Proceedings of the IJCAI-2003.

Erickson, T. and Laff, M. (2001) The Design of the 'Babble' Timeline: A Social Proxy for Visualizing Group Activity over Time. In Human Factors in Computing Systems: The Proceedings of CHI 2001, ACM Press.

Gtzelmann, T., Hartmann, K., Nrnberger, A. and Strothotte, T. (2007) 3D Spatial Data Mining on Document Sets for the Discovery of Failure Causes in Complex Technical Devices. In 2nd Int. Conf. on Computer Graphics Theory and Applications.

Keim, D. A. (2002) Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1), pp. 1-8.
Klink, S., Reuther, P., Weber, A., Walter, B. and Ley, M. (2006) Analysing social networks within bibliographical data. In DEXA 2006, LNCS 4080, pages 234{243. Springer-Verlag.
Lauw, H., Lim, E., Tan, T. and Pang, H. (2005) Mining Social Network from Spatio-Temporal Events. In Proceedings of Midwest SAS User Group.
Nielsen, F.A. (2006) Text and spatial data mining.
Petre, M., Blackwell, A. and Green, T. (1997) Cognitive questions in software visualisation. In Stasko, J., Domingue, J., Brown, M. H. and Price, B. A. editors, Software Visualization: Programming as a Multimedia Experience, chapter 30, pages 453-480. MIT Press.
Petrou, C. (2005) Using SAS for Spatial Analysis.
Tufte, E. R. (1983) The Visual Display of Quantitative Information. Graphics Press.
Wasserman, S. and Faust, K. (1994) Social Network Analysis: Methods and Applications. Cambridge University Press.
Weiss, S. M., Indurkhya, N., Zhang, T. and Damerau, F. (2005) Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer.