# Tracing Conceptual and Geospatial Diffusion of Knowledge

Chaomei Chen[1,2], Weizhong Zhu[1,2], Brian Tomaszewski[1,3], and Alan MacEachren[1,3]

[1] Northeast Visualization and Analytics Center (NEVAC)
[2] College of Information Science and Technology, Drexel University, Philadelphia, USA
[3] Department of Geography, Penn State University, State College, USA
chaomei.chen@cis.drexel.edu, wz32@drexel.edu, bmt139@psu.edu, maceachren@psu.edu

**Abstract.** Understanding the dynamics of knowledge diffusion has profound theoretical and practical implications across a wide variety of domains, ranging from scientific disciplines to education and understanding emergent social phenomena. On the other hand, it involves many challenging issues due to the inherited complexity of knowledge diffusion. In this article, we describe a unifying framework that is designed to facilitate the study of knowledge diffusion through multiple geospatial and semantic perspectives. In particular, we address the role of intrinsic and extrinsic geospatial properties of underlying phenomena in understanding conceptual and geospatial diffusion of knowledge. We illustrate the use of visualizations of geographic distributions of terrorist incidents, the structural evolution of research networks on terrorism and avian flu, and concept-location relations extracted from news stories.

**Keywords:** knowledge diffusion, geographic mapping, collaboration networks, information visualization.

## 1 Introduction

Understanding the dynamics of knowledge diffusion has become increasingly challenging due to the overwhelming volume of data from multiple sources and the increasing complexity associated with multiple perspectives. The diffusion of knowledge and technical innovations involves individuals, groups, and communities at various stages. Knowledge diffusion in mass opinions is characterized by the emergence of consensus or the expansion of one thematic thesis across a given population. Research in fields such as social network analysis [1], citation mapping and information visualization [2] has addressed various issues concerning the structural complexity challenge. Research in knowledge discovery and data mining is also relevant, notably in the areas of concept drifts [3, 4], topic detection [5, 6], and change detection [7, 8].

The aim of this article is to introduce an integrative approach to tracing conceptual and geospatial aspects of knowledge diffusion so that one can explore knowledge diffusion from different perspectives in a consistent framework. Specifically, our approach is designed to improve our understanding of the structure, the growth, and the spread of knowledge. In this article, we focus on the role of intrinsic and extrinsic

D. Schuler (Ed.): Online Communities and Social Comput., HCII 2007, LNCS 4564, pp. 265–274, 2007.
© Springer-Verlag Berlin Heidelberg 2007

geospatial properties of an underlying knowledge diffusion process. An intrinsic geospatial property identifies the inherent geospatial nature of an event. For example, the location of an avian flu outbreak is intrinsic because it is essential to our knowledge of the event. In contrast, an extrinsic geospatial property is secondary to the understanding of a phenomenon. For example, the interest and expertise of a researcher may or may not have anything to do with the location of his/her institution. In this article, we describe a conceptual framework such that data from multiple sources with various degrees of geospatial relevance can be accessed and contrasted in a unifying analytic environment.

## 2 Related Work

The study of information diffusion is concerned with how and why people accept or reject a new idea. The simplest model of information diffusion is the “Magic Bullet” model, which is also known as the *hypodermic needle model*. In this model, an intended message is sent directly to a receiver and unconditionally accepted by the receiver. A more sophisticated model is the *two-step flow model*. It suggests that the spread of information from mass media to the society takes two steps. First, the information is filtered through opinion leaders, who then influence others. The classic example is that undecided voters before an election tend to vote the way their friends and colleagues voted later in the election [9]. Opinion leaders play a vital role in this model.

Rogers [10] explains how a new innovation or idea would spread through society in terms of adopters of five types. Innovators would be the first to accept the new idea, then early adopters, early majority, late majority, and finally laggards. Early adopters are usually social leaders, popular, and educated, whereas laggards tend to have neighbors and friends as their main sources of information and fear of debt. The adopter-based model has been modified to account for the spread of high tech products [11].

Knowledge diffusion within scientific communities has largely attributed to the role of invisible colleges [12]. The role of social networks in diffusion of technical innovations was identified in [13]. Studies of scientific collaboration networks found that co-authorship is the best predictor of subsequent citations [14]. There is a growing interest recently in how information spreads over web logs, or blogs [15].

Purely geographically driven diffusion paths are relatively straightforward to derive. In this article, we address the challenge of integrating geographic and semantic perspectives within a unifying framework such that users are able to explore salient patterns back and forth between a geographic space and a social-semantic space of knowledge. In particular, we focus on collaboration networks of individual researchers by co-authorship and geographic distributions of events in the physical world. Our goal is to combine the two potentially interrelated and complementary but so far inadequately integrated perspectives, namely geographic-centric and semantic-centric perspectives. Such integration is significant because of the societal nature of knowledge creation, information seeking and sense making, and the potential of advances in human-computer interaction to enable these processes.

# 3 Methods

Our approach consists of four major components: information extraction, geographic coding, constructing associative networks, and constructing thematic layers for geographic visualization (see Figure 1). We explain each component in terms of the example datasets used in this study.

## 3.1 Information Extraction

In this article, we consider two types of data in terms of the role of geospatial properties: Type-A, geographic properties are secondary, and Type-B, geographic properties are primary. Collaboration networks are an example of Type-A data because geographic properties of such networks are secondary in nature. Vertices in such networks are individual researchers in a specific knowledge domain, such as avian flu or terrorism. Edges in these networks are collaborative ties between researchers. The strength of a collaboration tie between two researchers is measured by how many times they published joint papers with each other. On the other hand, terrorist incidents, such as suicide bombing, kidnapping, and shooting, are an example of Type-B data. Geographic locations are essential in this case.

Several Type-A datasets were retrieved from the Web of Science, including research on terrorism (1990-2006), avian flu (2001-2006), and astrophysics (a subset known as Sloan Digital Sky Survey – SDSS) (1996-2006). We also collected 1,427 terrorist incidents as a Type-B dataset from a website maintained by the Israeli International Policy Institute for Counter-Terrorism (ICT[1]). These incidents took place between May 1980 and December 2002 world wide. Each incident is recorded with date, location, responsible terrorist organization, type of incident, causality, and a short description of the event (See Table 1).

**Table 1.** An example record from the ICT dataset

| | |
|---:|---|
| Date | *Sept 9, 2001* |
| Location | *Nhariya, Israel* |
| Attack Type | *Suicide Bomb* |
| Target | *Civilian* |
| Casualties | *3 Killed; 90 Injured* |
| Organization | *Hamas; Number of Terrorists: 1* |
| Description | *Three people were killed and some 90 injured, most lightly, in a suicide bombing near the Nahariya train station in Northern Israel. The terrorist, killed in the blast, waited nearby until the train arrived from Tel-Aviv and people were exiting the station, and then exploded the bomb he was carrying. Hamas claimed responsibility for the attack.* |

## 3.2 Geographic Coding

Geographic coding identifies the altitude and longitude of an event or the institution of a researcher so that we can precisely locate the event or the institution on a

[1] http://www.ict.org.il/

geographic map through a thematic overlay. Locations in the USA are resolved by using zip codes, whereas locations outside the USA are resolved based on postal codes and direct name searches. We use the web services provided at www.geonames.org.

The successful rate of geocoding is high for Type-A datasets: 92.28% for the avian dataset (2001-2006), 95.98% for the terrorism dataset (1998-2003), and 96.53% for the SDSS dataset (2001-2006). However, the successful rate of 60.05% is relatively low for the Type-B dataset, i.e. terrorist incidents. Among the total of 846 locations in the terrorist event dataset, 62 are unspecified locations. Here are some examples of other types of problematic locations: 'enroute to London from Jedda," "outskirts of Jerusalem," "Atarot Industrial Zone," "Arc Royal aircraft carrier," and "Egyptian Border."

### 3.3 Collaboration Networks

An integration of conceptual networks with geospatial perspectives is illustrated with examples of collaboration networks of individual researchers. Such networks are derived from bibliographic records retrieved from the Web of Science. Each bibliographic record contains its authors, title, abstract, and a list of addresses of the authors. These authors are called co-authors. Co-authorship indicates the existence of a collaboration tie between co-authors. For example, a record shows that a 2006 article, entitled '*Avian influenza H5N1 in viverrids: implications for wildlife health and conservation*,' has coauthors from the following institutions:

Univ Hong Kong, Dept Microbiol, Hong Kong, Peoples R China.
Univ Hong Kong, Dept Pathol, Hong Kong, Peoples R China.
Univ E Anglia, Ctr Ecol Evolut & Conservat, Norwich NR4 7TJ, Norfolk, England.
World Hlth Org, Natl Inst Vet Res, Dept Virol, Hanoi, Vietnam.
World Hlth Org, Communicable Dis Surveillance & Response Unit, Hanoi, Vietnam.
Owstons Civet Conservat Program, Ninh Binh, Vietnam.
Endangered Primate Rescue Ctr, Ninh Binh, Vietnam.

Altitudes and longitudes of cities such as Hong Kong, Norwich, Hanoi, and Ninh Binh would be resolved. These cities would be connected by lines in the 2006 model of the collaboration network. In this way, we expect to reveal collaboration patterns with reference to geospatial distributions.

### 3.4 Multiple Layers of Thematic Overlays

Once collaboration networks are constructed and geospatial details are retrieved, multiple layers of thematic overlays are produced so that they can be viewed with Google Earth. A useful way to explore the diffusion of knowledge over geographic boundaries is to compare and contrast thematic layers of adjacent time intervals. For example, one may identify the expansion of a collaboration network over time by tracking the locations of major hubs year by year. In addition, one can hide and show multiple layers by topic and look for cross-layer patterns.

The four layers in Figure 1 illustrate the benefit and flexibility for users to study the diffusion of conceptual patterns as well as geospatial patterns. These overlays show the expansion of the avian flu research network between 2003 and 2006 and spread among countries in Southeast Asia.

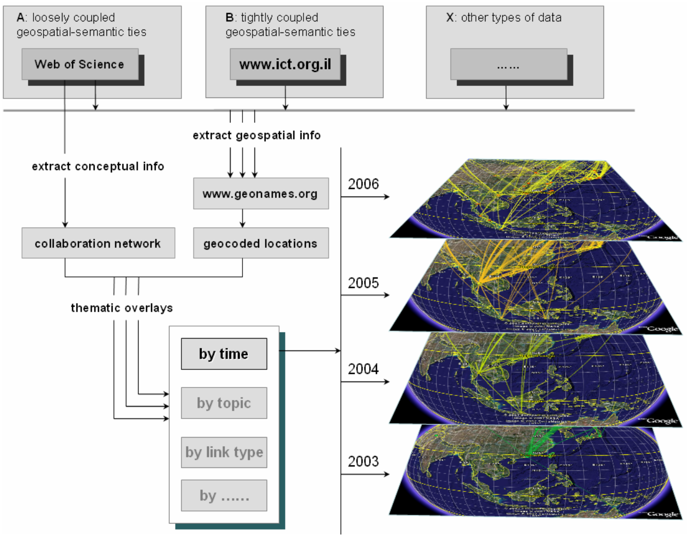


**Fig. 1.** A conceptual overview of our approach shows key components, processes, and results. The four thematic layers depict the growth of the avian flu research network in Southeast Asia.

Geospatial patterns play distinct roles in different conceptual spaces. The structure of a geographic space of researchers may not match the structure of a semantic space of knowledge because researchers do not always choose their collaborators based on geographic proximity. Furthermore, a geographic space of events, such as terrorist incidents or avian flu outbreaks, may or may not have anything to do with a geographic distribution of relevant expertise. For example, a group of researchers in Hong Kong publishing on avian flu may not imply a nearby avian flu outbreak. However, if researchers in several Southeast Asian countries were found to collaborate frequently with researchers in Memphis in the USA, such collaboration ties in a knowledge space could be particularly valuable because the absence of geographic proximity means that their connections seem to be unlikely in the geographic space. Facilitating users to explore patterns in two spaces can help users recognize and better understand patterns that may not be obvious if users only search in one space. The possibility of bridging geographic gaps via semantic proximity in a collaboration network is also encouraging for responders to terrorist incidents or avian flu outbreaks. For example, they would need to find not only experts on a specific subject but also experts who have knowledge of specific geographic areas. From a knowledge diffusion point of view, being able to see collaboration networks over a geographic map is the first step towards a deeper understanding of the interplay between knowledge and the context of its application.

The following examples demonstrate the use of different thematic layers to highlight potentially interesting patterns. These examples all have geospatial references. However, geospatial references in some of the datasets are essential to the underlying phenomena, but they are secondary in other datasets.

## 4 Examples

We first depict geospatial distributions of terrorist incidents such as suicide bombings, shooting, and kidnappings. Then we impose thematic overlays of research networks of terrorism. The second example illustrates the expansion of collaboration networks on avian influenza across Southeast Asian countries.

### 4.1 Terrorist Incidents and Research Networks on Terrorism

Figure 2 shows an example in which thematic layers are selected, namely the terrorist attack layer and the collaboration network layer of terrorism research. The occurrence of each incident is marked as a red translucent disc at where the incident took place. Thus multiple incidents in an area will accumulate higher density than geographically isolated incidents. As shown in Figure 2, Israel is marked by a dense cluster of incidents. The figure also shows a higher concentration of collaboration links in Europe, especially in Britain, than other areas. A few long lines across the globe indicate joint publications between Israeli researchers on terrorism and remote collaborators.

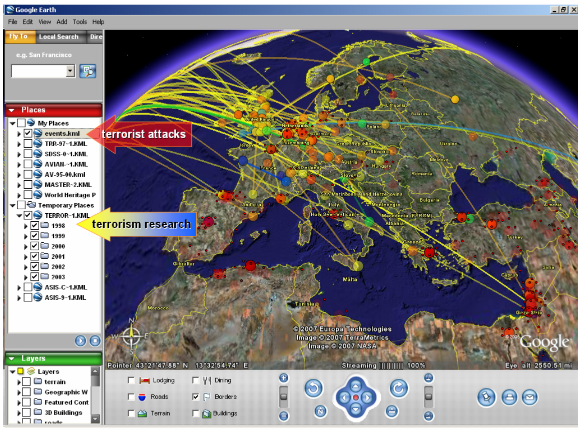
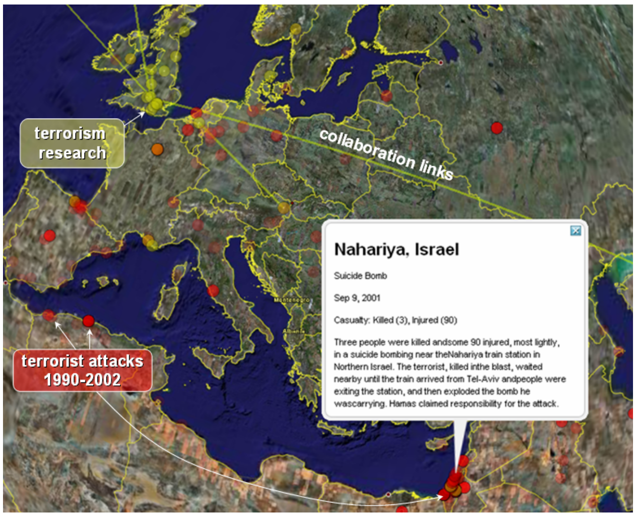


**Fig. 2.** Geographic distributions of terrorist incidents and research networks on terrorism

Figure 3 shows the details of a terrorist incident in Nahariya, Israel and a few collaboration links in the same timeframe, 2001. Terrorist incidents are shown as red markers, whereas research sites are depicted as greenish yellow markers.



Nahariya, Israel

Suicide Bomb

Sep 9, 2001

Casualty: Killed (3), Injured (90)

Three people were killed andsome 90 injured, most lightly, in a suicide bombing near theNahariya train station in Northern Israel. The terrorist, killed inthe blast, waited nearby until the train arrived from Tel-Aviv andpeople were exiting the station, and then exploded the bomb he wascarrying. Hamas claimed responsibility for the attack.

**Fig. 3.** Details of terrorist incidents are available by clicking on corresponding markers

Figure 4 illustrates the use of multiple themes simultaneously on the same map, including collaboration networks on three different subjects and terrorist incidents. This would make it easy for users to identify various interrelationships.

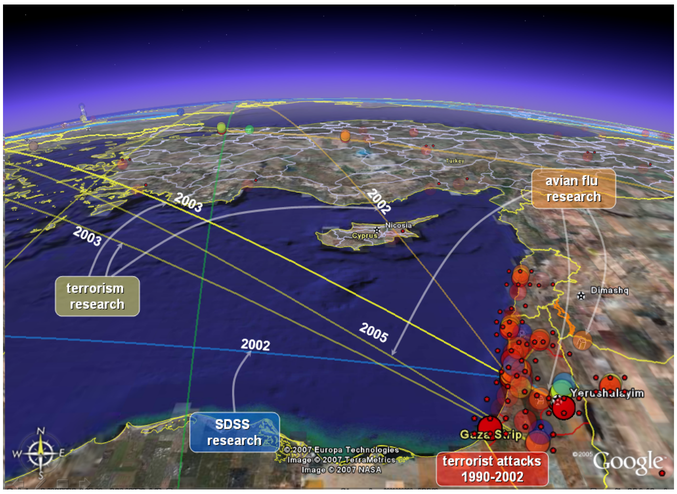


**Fig. 4.** Israel, as seen here, is the site of a large number of terrorist attacks. Multiple layers of thematic overlay also reveal other research activities in the areas of astronomy, avian flu, and terrorism in terms of collaboration links.

## 4.2 Avian Influenza

Figure 5 shows knowledge diffusion paths of avian flu research in Southeast Asian countries. The regional hubs of collaboration moved from Tokyo to Hong Kong, then reached Bankok and later on Jakarta. A strong collaboration triangle among Hong Kong, Bangkok, and Hà Noi emerged since 2004 and subsequently strengthened in 2005 and 2006.

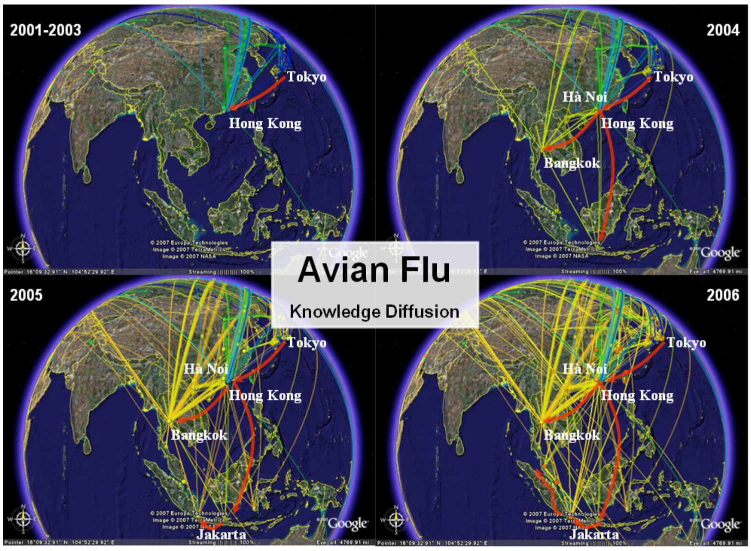
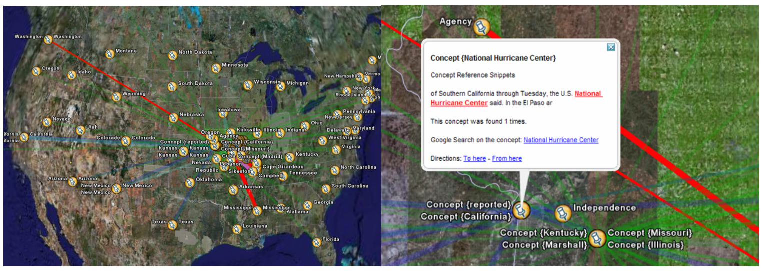


**Fig. 5.** Knowledge diffusion paths of avian flu research in Southeast Asia



Concept {National Hurricane Center}

Concept Reference Snippets

of Southern California through Tuesday, the U.S. National Hurricane Center said. In the El Paso ar

This concept was found 1 times.

Google Search on the concept: National Hurricane Center

Directions: To here - From here

**Fig. 6.** Left: Geographic locations computationally extracted from news articles about flooding in the mid-Western United States. Right: The detail of an abstract concept (National Hurricane Center) found in a news story and a related geographical location.

### 4.3 News Stories

News stories implicitly contain numerous geographic references such as towns, cities, and counties. Such information can be used to geographically contextualize situations such as disaster recoveries and humanitarian relief missions [16]. Figure 6 shows geographic locations computationally extracted from news articles about flooding in the mid-Western United States. Lines indicate connections between a news articles geographical origin and other geographical locations found in the article. Colors represent different individual news articles, thickness of line indicated the frequency of mention in the news article, transparency of a line indicates how old a story is (the older the story, the more transparent the line). Locations are plotted based on user-selected geographic scales allowing the user to view data at varying local, national, or international scales.

## 5 Conclusions

The work contributes to human-computer interaction in several ways: 1) It provides an intuitive and consistent framework to combine the visualization of concrete and abstract diffusion processes; 2) It introduces a potentially effective way to explore conceptual and geospatial diffusion of knowledge over time; and 3) This is a generic approach that can be applicable to a wide range of domains.

This is our ongoing effort to facilitate the understanding of dynamic and complex information processes and sense making activities involving large volumes of information, which is applicable to understanding online communities and social phenomena. Research on geocoding of locations derived from implicit geographic data sources will focus o establishing a document's geographic origin [17], examining multiple geo-spatial contexts within a document [18], improved disambiguation of geographic names [19, 20],and improved matching of abstract concepts with geographical locations [21, 22].

Future research directions include exploratory visual analytics methods that support comprehensive analysis of three-way spatial, temporal, and semantic relationships that are embedded (and potentially) hidden in these rich data sources. One strategy we will adapt is the *group, select, and filter* methodology in which analysts can select arbitrary subsets of entities in multivariate space to explore the cross-connections [23]. We have demonstrated the integration of two types of thematic perspectives in this article. Integrating spatial data with non-spatial data is an even more challenging area of research for future research.

**Acknowledgments.** The work is in part supported by the National Visualization and Analytics Center (NVAC) through the Northeast Visualization and Analytics Center (NEVAC) and the National Science Foundation under grant IIS-0612129.

## References

1. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
2. Chen, C.: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. 57, 359–377 (2006)
3. Klinkenberg, R., Renz, I.: Adaptive information filtering: learning in the presence of concept drifts. In: Klinkenberg, R., Renz, I. (eds.) Learning for Text Categorization, pp. 33–40. AAAI Press, Menlo Park, CA (1998)
4. Tsymbal, A., Pechenizkiy, M., Cunningham, P., Puuronen, S.: Dynamic integration of classifiers for tracking concept drift in antibiotic resistance data. Technical Report TCD-CS2005-26. Department of Computer Science, Trinity College, Dublin, Ireland (2005)

5. Morinaga, S., Yamanishi, K.: Tracking dynamics of topic trends using a finite mixture model. In: KDD'04, pp. 811–816. ACM, Seattle, Washington (2004)
6. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: KDD'04, pp. 306–315. ACM, sEATLEWashington (2004)
7. Kleinberg, J.: Bursty and hierarchical structure in streams. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 91–101. ACM Press, Edmonton, Alberta, Canada (2002)
8. Kumar, R., Novak, J., Raghavan, P., Tomkins, A.: On the Bursty Evolution of Blogspace. WWW2003, Budapest, Hungary, p. 477 (2003)
9. Lazarsfeld, P.F., Berelson, B., Gaudet, H.: The people's choice: How the voter makes up his mind in a presidential campaign. Columbia University Press, New York (1944)
10. Rogers, E.: Diffusion of Innovations. The Free Press, New York (1962)
11. Moore, G.: Crossing the Chasm. Harper Business, New York (1991)
12. Crane, D.: Invisible Colleges: Diffusion of Knowledge in Scientific Communities. University of Chicago Press, Chicago, Illinois (1972)
13. Singh, J.: Social networks as determinants of knowledge diffusion patterns. Vol. 2004 (2004)
14. Newman, M.: The structure of scientic collaboration networks. Natl. Acad. Sci, vol. 98, pp. 404–409, USA (2001b)
15. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proceedings of the 13th international conference on World Wide Web, New York, NY, pp. 491–501 (2004)
16. Mubareka, S., Khudhairy, D.A., Bonn, F., Aoun, S.: Standardising and mapping open-source information for crisis regions: the case of post-conflict Iraq. Disasters 29, 237–254 (2005)
17. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-Where: Geotagging Web Content. In: SIGIR'04. ACM, Sheffield, South Yorkshire, UK, pp. 273–280 (2004)
18. Graupmann, J., Schenkel, R.: GeoSphereSearch: Context-Aware Geographic Web Search. In: SIGIR '06. Workshop on Geographic Information Retrieval, ACM, Seattle, WA, USA (2006)
19. Wang, X.: Robust utilization of context in word sense disambiguation. In: Dey, A.K., Kokinov, B., Leake, D.B., Turner, R. (eds.) CONTEXT 2005. LNCS (LNAI), vol. 3554, pp. 529–541. Springer, Heidelberg (2005)
20. Rauch, E., Bukatin, M., Baker, K.: A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References (2003)
21. Liu, X., Pezanowski, S., MacEachren, A.M.: Bridging forms of knowledge for crisis management: Concept maps to Geographic maps. In: Kraak, M.-J. (ed.): ICA Commission on Visualization and Virtual Environments (2006)
22. Mitra, P., Pan, C.: Extracting Semantic Networks among Named Entities from Websites. International Conference of the Association of Computational Linguistics (submitted)
23. Weaver, C., Fyfe, D., Robinson, A., Holdsworth, D., Peuquet, D., MacEachren, A.M.: Visual analysis of historic hotel visitation patterns. Information Visualization (2007)