CO-354

MAPPING CYBERSPACE: TRACKING THE SPREAD OF IDEAS ON THE INTERNET

TSOU M.H.
San Diego State University, SAN DIEGO, UNITED STATES

BACKGROUND AND OBJECTIVES

This paper revisited an old research topic in cartography, "mapping cyberspace", which has been frequently discussed in the last decade (Shiode and Dodge 1999; Dodge and Kitchin 2001; Börner et al. 2003). However, different from the previous approaches, the development of new web search engines (Google and Bing) and new social network applications (Facebook and Twitters) provides a great potential to study the dynamic spread of ideas and concepts over the Internet. To visualize space-time dimensions of ideas spreading on the Internet, we need to design an effective method to geo-locate the contents of individual posts and web pages from cyberspace to realspace. This paper introduced an integrated geocoding method by utilizing geographic information systems (GIS), IP addresses, computational linguistics (CL), and computer-based ontology technologies to track and analyze the dynamic changes of ideas on the Internet. By using multiple geocoding methods, including gazetteers, user defined places, URLs, smart phone locating methods, and progressively refined place names, we can generate a visual "information landscape" consisting of the flow of ideas on a global map. When integrated with time-series analyses, this map might allow examination of the paths and speed of information dissemination, as well as the evolving varieties of various ideas and their relationships. To accomplish these goals, we started to develop a Semantic Web Automatic Reasoning and Mapping System (SWARMS) prototype by combing web search engine, GIS tools, and linguistic analysis methods.

In the SWARMS prototype, we first developed a semantic database through identifying the words and phrases that characterize sites related to these events. Then, we collected data on the spread of these words and phrases through web sites over time and space by using web search engines, such as Yahoo, Bing, or Google. By plotting chronological geographic paths, we aim to test the hypothesis that the spread of ideas is not random. That is, there are places, which are more prone to host these websites (and accept and spread an idea) than others over time. We can perform statistical analyses to understand the reasons for particular trajectories along which an idea spreads. In other words, we will be able to identify factors that cause "susceptibility" of and "immunity" from a particular set of ideas. The goal of this research is to develop a theoretical structure and a set of analytical methods on the spread of ideas or events of interest (e.g., impacts of diseases or disasters) in cyberspace.

Hopefully, this research will help us to better understand the 'collective thinking' of human beings and minimize misunderstandings between different groups and people. The following section will introduce the design of SWARM prototype and the major analysis methods.

APPROACH & METHODS

A prototype of Semantic Web Automatic Reasoning and Mapping System (SWARMS) was designed and implemented for displaying the dynamic information landscape (2D and 3D maps) and showed the spread of concepts, ideas, and news. The overall system framework is illustrated in Figure 1. Initial search uses pre-defined keywords in some selected topics (such as nature disasters, continuous threats for human beings, or radical social movements) provided by domain experts to search public accessible websites (using popular search engines, such as Google Search Engine, Microsoft Bing, or Yahoo). The search results were converted into a [Raw Text Database], which include all search results (ranking, titles, partial contents, and URLs). The system used the URLs and geo-locating methods to convert raw texts into [Semantic Web Information Databases], which include both geospatial locations (latitudes and longitudes) and semantic contents (keywords) for each record. By utilizing WHOIS protocol (a converting method from Web Domain Name and IP address to server registration addresses, http://www.whois.net/), we can easily convert Web addresses (URLs) into real places (with latitudes and longitudes). Domain experts can review these texts and use various tools of computational linguistics, GIS, and gazetteers to identify new key words, key phrases, and related spatial place names in the semantic web information databases. The databases (created by MS SQL server) will be converted to [Visualization maps] showing the dynamic information landscape of specific ideas or concepts. GIS, calculation of network connectivity, and space-time analysis will be used to understand the dynamic change of these concepts and events over space and time. Computational linguistics experts will establish frequencies of occurrences of "key terms," separately and in clusters. Multiple [Semantic Knowledge bases] (ontologies) related to ideas, concepts

and special topics will be created and revised based on the visualization maps and space-time analysis. The revised ontology terms and phrases will be used for the next round of Web query process. New web pages and websites will be discovered by advanced keyword clusters and generate new records in the [Semantic Web Information Databases].

The visualization maps and network analysis will constitute data for further quantitative analysis to enrich and refine the search algorithm and to learn more about the nature and specificity of ideas and their characteristic textual architectures. One advantage of this SWARMS framework is its language- and search-engine-independent architecture. This framework can be used to query keywords in multiple languages (Chinese, Arabic, or Japanese) and use multiple Web search engines. The design of SWARMS prototype also focused on how to select the appropriate spatial scales for mapping the locations of web pages and websites. We can analyze the potential errors (uncertainty) in the geo-locationing process and the accuracy of web server registration addresses. For examples, some personal weblogs might be published in a commercial website and the Weblog URL will only be linked to the commercial web server rather than individual blogger's locations. In this case, we might test different scale level maps (from country-scale to city-scale to street-level-scale) and find out the relationship between the spatial scale and the keywords. The spatial scale question is one of the major focuses in the development of SWARMS.

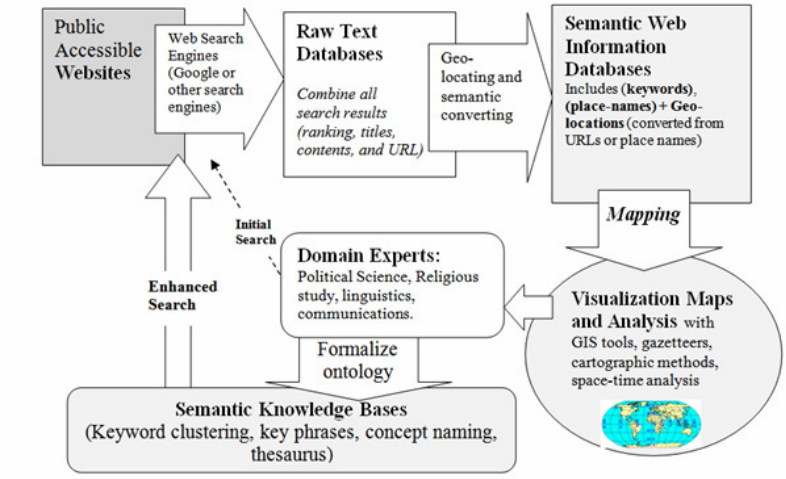

Figure 1. The Semantic Web Automatic Reasoning and Mapping System (SWARMS) framework.

RESULTS

The following examples illustrate four preliminary results of the major procedures and partial demonstration of technology feasibility for SWARMS. The first example is to use Google Search Engine API to query "Wildfires" and "San Diego Wildfire 2007" and retrieve the top 64 ranked records and convert them into maps (Figure 2). Our first preliminary test only used the Google Search Engine API, which returns a maximum of 64 records in a single automatic search. Bigger circles (in Figure 2) indicate higher rankings in the search results. The labels are the ranking of each website. By comparing the two maps, we can analyze the different spatial distribution patterns associated with different semantic terms. The clustered keywords ("San Diego Wildfire 2007") extend the concept of "wildfire" to a spatial semantic ("San Diego") and a temporal semantic ("2007"). Domain experts can then use multiple tools to analyze the differences between the two maps.

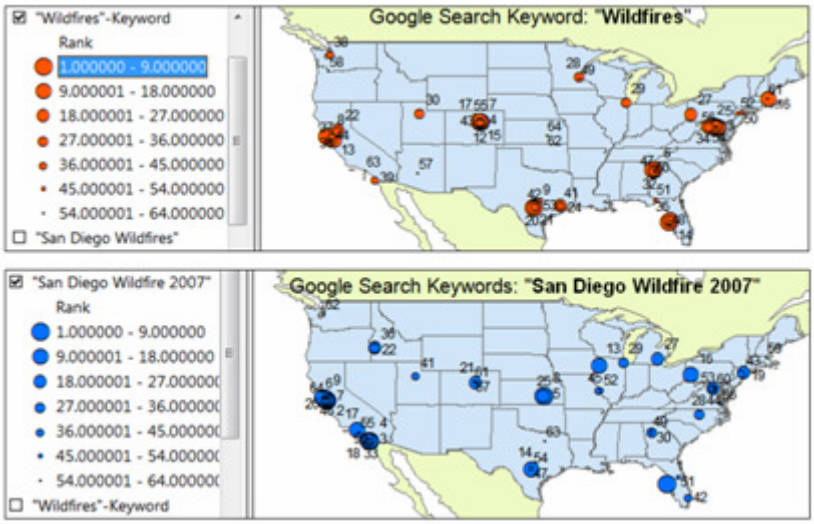

Figure 2. The information landscapes of "Wildfires" (a single keyword) and "San Diego Wildfire 2007" (clustered keywords). The labels are the ranks for each website.

	Rank	Search Engine	Keyword	Search Date	URL	Title	Latitude	Longitude	ZipCode	IP
2	1	Google	San Diego Wildfire 2007	1/11/2010	http://en.wikipec	October 2007 California wildfires -	27.7788	-82.6823		208.80.152.2
3	2	Google	San Diego Wildfire 2007	1/11/2010	http://sosdfirebl	Wildfires 2007	37.4192	-122.057	94043	74.125.95.191
4	3	Google	San Diego Wildfire 2007	1/11/2010	http://map.sdsu.	Internet Mapping Services for San	32.7751	-117.076	92182	130.191.118.202
5	4	Google	San Diego Wildfire 2007	1/11/2010	http://www.sign	San Diego Fires 2007 - Informatior	32.7773	-117.101	92108	69.43.137.203
6	5	Google	San Diego Wildfire 2007	1/11/2010	http://www.msn	S. California fires destroy hundred	38	-97		65.55.53.233
7	6	Google	San Diego Wildfire 2007	1/11/2010	http://helpsandi	Help in San Diego: Wildfires 2007	37.4192	-122.057	94043	74.125.95.191

Figure 3. A preliminary example of semantic web information databases (partial content).

Figure 3 illustrates a screen shot of the preliminary [semantic web information database] created by using WHOIS service, Java Scripts, and MS SQL server. Each database can be converted into an Excel spreadsheets and then generate a visualization map. The semantic web information databases include ranking, search keywords, search dates, URL, web page title, geo-locations, and related items.

The second example is the spread of news about an epidemic, "swine flu." In late April, 2009, incidences of swine flu in the United States raised the specter of a global epidemic and in mid-June the World Health Organization announced that a global pandemic of swine flu was underway. In the intervening period, a variety of news sources participated in spreading the concepts of "swine flu" and related terms, such as "H1N1 virus." Figure 4 shows the spatial diffusion of the "H1N1" keyword from April-29, 2009 to May-06, 2009. The maps show the top 100 Google search results for "H1N1" (using manually methods, so no 64-records limits). Comparing the April-29 map to the map of May-06, we find three new dots in Canada, Singapore, and Sweden. Given the more refined concept identification possible with clustered keyword search, the spatiotemporal development of much more subtle ideas and sentiments can be tracked. The approach extends to various issues related to natural disasters, continuous human threats, and radical social movements, enabling more in-depth analysis of the dynamics of these ideas.

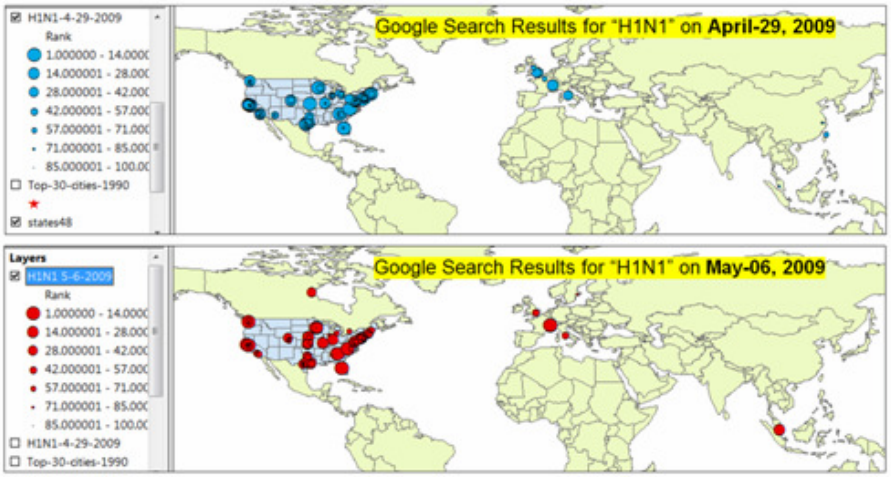

Figure 4. The spatial diffusion of the "H1N1" keyword from April, 29, 2009 to May, 6, 2009.

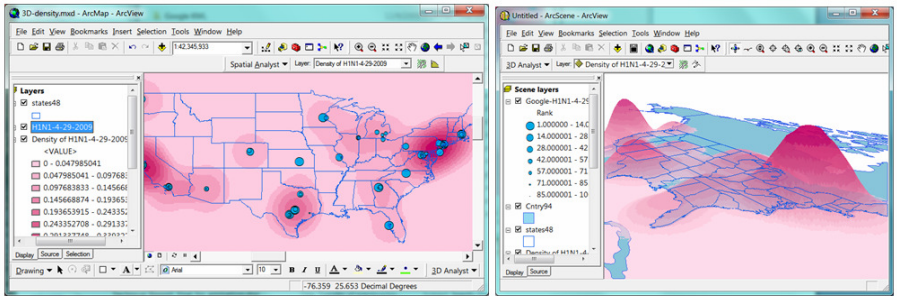

Figure 5. The density analysis and 3D information landscape (using H1N1 keyword search).

The third example (Figure 5) illustrates several possible spatial and GIS analysis methods for the semantic web information databases. We used GIS density analysis function (dot density) to create a density map (left) for our keyword clusters and then converted the density to a 3D landscape map (right). Various cartographic representation methods and GIS techniques can be adopted in this research for the visualization of information landscapes and semantic webs.

The final example (Figure 6) illustrated the scale dependent problem in creating visualization maps and the information landscape. In this example, we used Yahoo Search Engine to query the "Brotherhood of Klans" keyword and retrieved top one thousand search hits with associated ranks. Then, our SWARMS prototype utilize a commercial GIS software (ArcGIS) to create dot density maps by defining the threshold (radius) of kernel density algorithm. The scale dependent characteristic can be observed in the two maps using different map unit radius (Figure 6). The top map in Figure 6 showed the spatial pattern with 8 map unit radius (one map unit is approximate 50 miles in the real world). The bottom map showed the different spatial pattern with 3 map unit radius. The red dot indicated the locations of website associated with the keyword (Brotherhood of Klans). The density calculation also considered the ranking number for each website as its population parameter. The color scheme in Figure 6 is adopting a regular elevation color scheme to illustrate both high density areas and low density areas.

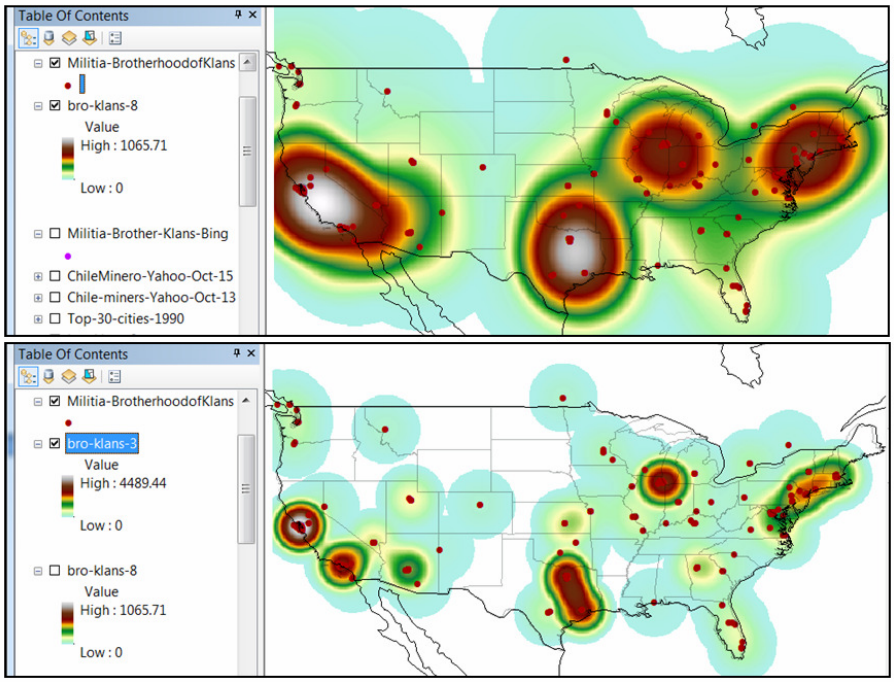

Figure 6. The scale dependent characteristic of dot density maps for keyword search results of "Brotherhood of Klans" in Yahoo Search Engine. (Top map: kernel density method with 8.0 map unit radius; bottom map: kernel density method with 3.0 map unit radius)

One interesting finding in Figure 6 is that the spatial pattern generated with radius 8.0 units clearly illustrates a "regional" perspective in United States (east, west, north, south). The map with radius 3.0 units shows a "state" level perspective in U.S. (California, Texas, Arizona, etc).

CONCLUSION AND FUTURE PLANS

In summary, this project seeks to map both the geography and the chronology of ideas over cyberspace, as the ripples of information usage radiate outward from a given event epicenter. By mapping and analyzing

such ripples, new insights will be provided into the role of new media in biasing, accelerating, or otherwise influencing social and political uses of such information.

We also realize that our mapping methods are still very premature and need to be improved. For example, we need to exclude several search hits from Wikipedia websites and public news websites. Also, the geolocation method in the SWARMS can only convert 85% of URLs into latitude and longitude coordinates. The rest of 15% of websites were not used in the visualization maps yet. Moreover, our current SWARMS prototype can only use Yahoo and Bing search engines because the Google Search Engine limits the API search up to 64 records only. Hopefully, these limits and restrictions will be addressed in our future research.

During our testing and prototyping, we also find out that the following three research topics are essential to the future development of cyberspace mapping.

1. How to analyze the spatial relationships among points (websites and individual web pages), lines (hyperlinks within web pages), and polygons (community groups or social networks) on cyberspace?

2. How to develop effective cartographic representation methods and map symbols to illustrate the dynamic flows of ideas and concepts on the Internet?

3. Which spatial scale is the best scale to represent specific ideas or concepts on maps?

These research questions along with the new SWARMS research framework may help us explore the dynamic interactions of various ideas in cyberspace. Cartographers might be able to use the new research framework to develop more advanced visualization principles and methods.

ACKNOWLEDGEMENT

The SWARMS framework is developed by the funding support from NSF project “CDI-Type II: Mapping Cyberspace to Realspace: Visualizing and Understanding the Spatiotemporal Dynamics of Global Diffusion of Ideas and the Semantic Web”. (NSF Award# 1028177). This project is a multidisciplinary collaboration among Ming-Hsiang Tsou (Geography), Dipak Gupta (Political Science), Jean Marc Gawron (Linguistic), Brian Spitzberg (Communication) and Li An (Geography) at San Diego State University.

REFERENCE

Dodge, M. and Kitchin R. (2001). Mapping Cyberspace. London: Routledge.

Shiode, N. and Dodge, M. (1999) Visualising the Spatial Pattern of Internet Address Space in the United Kingdom. In Gittings, B. (ed.) Innovations in GIS 6: Integrating Information Infrastructure with GI Technology, 105-118. London: Taylor and Francis.

Börner, K., Chen, C., and Boyack, K. W. (2003). Visualizing Knowledge Domains. Annual Review of Information Science & Technology, 37, 179-255.