International Journal of Digital Earth, 2014
Vol. 7, No. 4, 316–335, http://dx.doi.org/10.1080/17538947.2013.781240

Taylor & Francis
Taylor & Francis Group

Mapping ideas from cyberspace to realspace: visualizing the spatial context of keywords from web page search results

Ming-Hsiang Tsou[a]*, Ick-Hoi Kim[a], Sarah Wandersee[a], Daniel Lusher[a], Li An[a], Brian Spitzberg[b], Dipak Gupta[c], Jean Mark Gawron[d], Jennifer Smith[e], Jiue-An Yang[a] and Su Yeon Han[a]

[a]*Department of Geography, San Diego State University, San Diego, CA, USA;* [b]*School of Communication, San Diego State University, San Diego, CA, USA;* [c]*Department of Political Science, San Diego State University, San Diego, CA, USA;* [d]*Department of Linguistics and Oriental Languages, San Diego State University, San Diego, CA, USA;* [e]*Department of Geography, Pennsylvania State University, University Park, PA, USA*

(*Received 9 May 2012; final version received 22 February 2013*)

We introduce a new method for visualizing and analyzing information landscapes of ideas and events posted on public web pages through customized web-search engines and keywords. This research integrates GIScience and web-search engines to track and analyze public web pages and their web contents with associated spatial relationships. Web pages searched by clusters of keywords were mapped with real-world coordinates (by geolocating their Internet Protocol addresses). The resulting maps represent web information landscapes consisting of hundreds of populated web pages searched by selected keywords. By creating a Spatial Web Automatic Reasoning and Mapping System prototype, researchers can visualize the spread of web pages associated with specific keywords, concepts, ideas, or news over time and space. These maps may reveal important spatial relationships and spatial context associated with selected keywords. This approach may provide a new research direction for geographers to study the diffusion of human thought and ideas. A better understanding of the spatial and temporal dynamics of the 'collective thinking of human beings' over the Internet may help us understand various innovation diffusion processes, human behaviors, and social movements around the world.

Keywords: cyberspace; web-search engines; spatial context; geolocation; web information landscape; innovation diffusion

Introducing web information landscapes on digital earth

The world today is constantly awash with a flood of ideas, concepts, and social events, and the diffusion of these information and social communications now leaves measurable traces in digital earth that can be mapped and analyzed in near real time. The digitization of social media and web pages can provide massive human communication records and facilitate the emergence of a data-driven computational social science (Lazer et al. 2009; Miller 2011).

Mapping the diffusion of information and messages in response to various social movements, events, and epidemics is an important research topic of analytical digital

*Corresponding author. Email: mtsou@mail.sdsu.edu

earth. This paper introduces a new research framework for analyzing web keyword search results and web page content analysis, called Spatial Web Automatic Reasoning and Mapping System (SWARMS), to visualize the spread of ideas, events, and trends disseminated in digital earth (via the web and social media). In this article, we define 'the web' as the connected Internet on digital earth and its broader network-based applications, which include the World Wide Web, instant messengers, FTP servers, social media, web services, and so on. On the other hand, 'the World Wide Web' refers to the aggregations of web servers (websites) only, which are built upon the Hypertext Transfer Protocol (HTTP) with HTML documents (Berners-Lee, Hendler, and Lassila 2001).

The SWARMS prototype is designed to track spatial patterns of publically accessible web pages based upon searching predefined clusters of keywords. Web pages and web content associated with the same keywords were converted into visualization maps using GIS functions (e.g. kernel density calculation and raster-based map algebra methods). The resulting maps represent web information landscapes consisting of hundreds of website locations (latitudes and longitudes) ranked by web-search engines, such as Yahoo or Bing. Given the extent to which the human population is 'plugged into' the online world, we hope that the SWARMS prototype may track the social impact of significant events over time and space as they are reflected in digital earth.

The Web, linking millions of networks and billions of people, has become an important base of computer-supported social networks (Wellman et al. 1996). In this paper, we begin by reviewing efforts in mapping ideas and exploring web relationships. We then discuss geolocation methods and then describe the SWARMS approach in more detail. We follow by displaying the application of SWARMS in examining the spatial relationships between web page contents and Internet Protocol (IP) address geolocations and visualizing the spatial context of web page search results from Yahoo Search Engine. Since the Google search engine Application Programming Interface (API) can only provide up to 64 records (compared to 1000 records from Yahoo or Bing APIs), we did not include the Google search results in this paper.

Methods of mapping ideas on digital earth

With the high popularity of web-search engines and social networks, many researchers have focused on the spatial analysis and visualization of web information and keyword searches, such as Google Flu Trends (Ginsberg et al. 2009; Varian and Choi 2009), the Healthmap project (Brownstein et al. 2008), and BioCaster (Collier et al. 2010). While some parallels exist between previous projects and the SWARMS prototype, the SWARMS approach is unique. In contrast to our SWARMS prototype, Google Flu Trends only analyzes user input keywords as the source of web information rather than actual web page contents. The Health Map project created global maps of disease-related websites (Brownstein et al. 2008) based on random submissions by the public and individual researchers only (rather than by systematic web-search engine results). Another research avenue has been the pursuit of data-mining projects investigating word co-occurrence (Ohsawa et al. 2002), centrality (Corman et al. 2002), and sentiment analysis (Chute 2008; Bai 2011). For example, some research indicated that the relational forms of data manifest less in

the message content itself, and linkages within and across communication networks (Monge and Contractor 1998), social networks (Kempe, Kleinberg, and Tardos 2003; Papacharissi 2009; Singh, Gao, and Jain 2010), emails (Matsumura and Sasaki 2007), and websites (Elmer 2006). In particular, projects such as those focused on algorithms and structural topographical configurations and calculations (Shekhar and Oliver 2010) suggest that unique patterns may provide unique geospatial network 'fingerprints' that characterize the evolution of different social dynamic processes (Worboys 2010). Several scholars have suggested that there may be narrative markers of health-based (Little, Jordens, and Sayers 2003) and hate-based or terrorist groups (Leets and Bowers 1999; Hoffman 2005; Brown 2009). Such markers may be discernible through various data-mining techniques. Most web-related projects, however, only emphasize the visualization of information without using advanced GIS tools for further spatial-temporal analysis.

Another related research direction is geographic information retrieval (GIR) (Purves et al. 2007; Jones and Purves 2009). The scope of GIR ranges from the detection of geographic content on the Web (Markowetz, Brinkhoff, and Seeger 2003) to the analysis of IP geolocations (Buyukokkten et al. 1999), to the search engines of geotags (Amitay et al. 2004) and gazetteer reasoning databases (Silva et al. 2006). A seminal work in GIR, Purves et al. (2007) 'The Design and Implementation of SPIRIT: a Spatially-Aware Search Engine for Information Retrieval on the Internet' introduced the design, implementation, and evaluation of a spatially aware search engine. Their prototype identified geographic references from web pages (documents) and automatically created spatial footprints to index the contents. By using web crawlers (Joho and Sanderson 2004), geographical ontology databases, gazetteer lookup services, and geoparsing engines, SPIRIT can index and rank web documents based on their textual and spatial relevance (Purves et al. 2007). However, most of the GIR research projects designed their own search engines (crawlers) and did not utilize powerful commercial web-search engines to collect web contents associated with keywords. Compared to previous related mapping cyberspace research projects, our SWARMS prototype has four unique features:

(1) The SWARMS prototype utilizes powerful commercial web-search engine APIs (such as Yahoo BOSS APIs and Bing APIs) rather than developing small-scale web crawlers or web robots, which might not be powerful enough to index all cyberspace activities of interest on a daily or weekly basis.
(2) The prototype adopts real-world coordinates and real-world distance to geocode web pages as the 'locational proxy' of ideas or concepts on Earth. Different from maps using abstract coordinates (such as multidimensional systems), our visualization maps include real-world latitudes and longitudes. These visualization maps are more compatible with advancing spatial analysis methods. For example, we can compare these maps with census data to explore possible spatial relationships.
(3) GIS software and functions are used in SWARMS to conduct sophisticated spatial cluster visualization and temporal change analysis. Advanced GIS functions, such as kernel density and map algebra, are utilized for web content comparison analysis.
(4) We are mapping published web pages in cyberspace rather than counting user-submitted keyword search frequencies. Web pages are created by

'information providers' which are different from user-input keywords (by information requesters or readers). Our focus is to analyze where the actual information (web pages) were created rather than where people submit their requests (keywords).

Geolocation methods

Recently, the World Wide Web Consortium (W3C) developed a standardized specification for Geolocation APIs (World Wide Web Consortium 2010). The standardized APIs allow various web applications to share and utilize geographic location information gathered from the host devices or users. 'Common sources of location information include global positioning system (GPS) and location inferred from network signals such as IP address, RFID, Wi-Fi, and Bluetooth MAC addresses, and GSM/CDMA cell IDs, as well as user input' (World Wide Web Consortium 2010). So far, however, few web pages have adopted W3C geolocation APIs due to the lack of available geolocation tools. We hope that more web application and web server administrators will adopt the W3C specification in the future.

The SWARMS prototype utilizes automatic geolocation methods to create multiple web information landscapes for different keywords. Table 1 illustrates three popular geolocation methods available for mapping web content and social media: (1) IP geolocation; (2) mobile device tracking (GPS, Wi-Fi, or cellular signals); (3) geographic context analysis (gazetteers, geographic names, and spatial reasoning).

Table 1. Three major geolocation methods for mapping web content and social media.

Methods	Accuracy/spatial resolution	Spatial data availability	Implementation requirement
IP geolocation	US: Zip code level (92119) or city level (10–100 km). International: city level (Taipei, New York, etc.) or country level	All web pages have IP addresses. We can label 90% of web pages with geotags using IP geocoding methods	WHOIS databases (public), commercial or free IP geolocation databases or web services (IPPAGE.org, MaxMind.com)
Mobile device tracking (GPS, Wi-Fi, or cellular signals)	● GPS: 8 meters (average median error) ● Wi-Fi: 74 meters ● Cellular: 600 meters	Public: Twitter has only 1% of tweets with geolocation enabled. Private: mobile phone companies have user location data, but they are confidential	Smart phones or mobile devices with assisted GPS functions or Wi-Fi enabled
Geographic context analysis (gazetteers, geographic names)	Spatial footprint resolutions range from 1 to 1000 km, to 3000 km	10% web pages have US zip code information; 20% have geographic identifiers (Himmelstein 2005)	Challenging. Requires a comprehensive framework with geographic name ontology, gazetteers, spatial reasoning tools

The first method, IP geolocation (Muir and Oorschot 2009), is a popular technique for identifying geographic location of Internet users or web servers. Researchers can convert IP addresses into real-world coordinates (latitudes and longitudes) or geographic regions by using IP geolocation methods. The geolocation analysis of website visitors has become an important component in Web log analysis research (Fleishman 1996; Turner 2004) and has been applied in various domains, including location-based service, target marketing, epidemiology, and criminal investigation (Choi and Tekinay 2003; Lee 2008; Tsou and Kim 2010).

There are two types of IP geolocation techniques: active IP geolocation and passive IP geolocation. Active IP geolocation technique relies on the time delay measure of network routing (such as ping functions) from one IP address to another. However, the active method requires extensive calculations and cannot handle a very large volume of IP geolocations. Passive IP location is a database-driven procedure which relies on relational databases (such as MS SQL or MySQL databases). The IP geolocation databases include the index for mapping different levels of IP address (blocks or prefixes) to countries, cities, zipcodes, and real-world coordinates (Poese et al. 2011). For example, the database can convert the IP address 130.191.118.3 to the US zip code 92182. The database also includes the latitude and longitude coordinates of the central point of zip code polygons.

Currently, there are several commercial or free IP geolocation databases available, such as IPligence, MaxMind, and IP2Location. The spatial resolution of commercial geolocation databases is a probability of 62–73% to place an IP location within 40 km from the 'Points of Presence' (the actual user's device or the registration location of web servers) (Shavitt and Zilberman 2010). Most commercial IP geolocation databases claimed that their spatial resolution can reach to the zip code level in the United States and to the city level for international counties. However, some academic researchers argue that significant uncertainty and accuracy problems exist for IP geolocation. For example, Youn, Mark, and Richards (2009) calculated the median of estimated errors in a statistical geolocation method as 53 km. Poese et al. (2011) argue that the accuracy of IP geolocation in the European region did not reach the city level, but only the country level. Although there are some uncertainty problems for IP geolocation methods, they can create geolocation tags over 90% of web pages. The implementation requirements for IP geolocation methods are relatively easy compared to other methods.

The second geolocation method is mobile device tracking through GPS, Wi-Fi signatures, or cellular signals. This method can track down the coordinates of web devices or users accurately with high spatial resolution. The GPS tracking resolution can reach 8–10 meters, the Wi-Fi signature tracking resolution is around 74 meters, and the cellular signal triangulation methods can have 600-meter resolution (Zandbergen 2009). The major problem for this method is data availability. Most web servers do not have attached GPS devices. In our collected public social media datasets, only 1% of media content (Tweets) contains GPS or Wi-Fi coordinates. For Twitter in particular, research by Takhteyev, Gruzd, and Wellman (2012) also shows evidence as only 6% tweets associated with GPS coordinates in their research. Most smart phone users prefer not to turn on the GPS-geolocation functions in their social media applications (Twitter or Facebook).

The third geolocation method is geographic context analysis using gazetteers, geographic names, and ontology databases (Purves et al. 2007). Although this is a

promising geocoding method for web content and web searches, there are several reasons why we chose to use a different method in our SWARMS prototype. First, the spatial resolution of this method varies ranging from 500 meters (such as Sea World in San Diego) to 800 km (the State of California). Some web pages may not include geographic identifiers in their contents (only 20% of web pages have geo-name identifiers) (Himmelstein 2005). This method cannot efficiently handle large amounts of records due to the manual process.

After comparing three different geolocation methods, we adopted IP geolocation as our major SWARMS prototype geolocation method. In our test, the SWARMS prototype performed geolocation procedures for approximately 90% of Yahoo's top 1000 search results. When a geolocation process fails, the IP address receives the assignment of 0 latitude, 0 longitude (as a point in the middle of ocean south of Ghana and east of Gabon). One issue is that the original website IP addresses could be replaced by proxy servers, and the geolocations of these machines might be incorrect (Svantesson 2005). A proxy server acts as a connector between users and the actual websites. When an IP address of a web server is converted to a geolocation, some Internet machines link to proxy servers in order to protect their geolocations and privacy (Muir and van Oorschot 2009). Due to the limitation of current geolocation technology, we cannot guarantee 100% accuracy for all geolocation procedures (Buyukokkten et al. 1999). Although geolocation may have other accuracy problems, such as geolocation database errors, or address-matching problems, the unsuccessful conversion rate is relatively small (10–12%) in our geocoding tests.

Web-search technologies and methods

Web-search engines, such as Google and Yahoo, have become the *de facto* method for people to find information on the Web. These search engines control how people access websites and what information they can obtain from search requests. The SWARMS prototype relies on the successful development of commercial web-search engine technology to query the related web contents through single or multiple keyword searches. For example, users can submit a keyword (text-based) search to Yahoo.com that returns the 100 top-ranked pages. The higher-ranking sites are usually more relevant to the submitted keywords or more popular among users. In the research reported here, the web-search ranking numbers serve as the 'popularity' index of the web pages. For example, if we search 'SDSU' on Yahoo.com, the first hit (Rank#1) is http://www.sdsu.edu (San Diego State University), which means that this web page is the most popular web page for the keyword 'SDSU'. The second rank is 'http://sdstate.edu' (South Dakota State University), which means that this website is less popular than the #1 website associated with the 'SDSU' keyword.

Our SWARMS prototype adopted both Yahoo and Bing search engines, because they provide up to 1000 web pages from their APIs in a single keyword search. Yahoo's search engine algorithm generates ranking numbers based on the relevancy of web pages to the submitted keywords. Web page titles, header texts, body descriptions, and associated links are analyzed inside the Yahoo search engine algorithms. User click popularity is also one major factor considered by Yahoo search engines. The more users click on a specific website from the list of keyword search results, the higher the ranking of the website will become in the next identical

keyword search. This method allows actual user experience and user feedback to contribute to the calculation of web page ranks. Bing (from Microsoft) is another popular search engine adopted in our SWARM prototype. However, after comparing the top 1000 search web pages between Yahoo and Bing, the Yahoo search became the preferred engine because of Bing's common limitation of web pages to within the United States. Yahoo search engine covers more international websites from different countries among its top 1000 web pages. We also did a comparison between the Yahoo API search results and Bing API search results and found out that the Bing search engine tends to return more commercial, informational (wiki-type), and social media web pages and the Yahoo search engine will provide more blogs, news, and educational web pages.

Visualizing information landscapes

Geographers and cartographers have studied information landscapes and cyberspace mapping for a few decades. However, most of these research activities did not emphasize a strong linkage between real-world coordinates and spatial representations of cyberspace. Many cyberspace maps use alphabetical reference systems, such as Domain Name Systems tree structures or IP addresses rather than real-world latitudes and longitudes. Without the linkage to real places, it is difficult to perform advanced spatial and temporal analysis with census data (collected with real-world locations) or environmental data. The SWARMS prototype aims to bridge this gap by 'spatializing' web-search results and web pages using real-world coordinate systems.

One early example of a web information landscape can be found in Shiode and Dodge (1999), introducing a visualization approach converting thousands of web hosts into real-world coordinates with geolocation methods. A total of 10,183 web servers located in UK with their IP addresses were converted to geographic points according to their registered organizations' locations. These locations were represented with various cartographic methods, including dot density maps, density surface maps, and three-dimensional (3D) density landscapes showing different types of websites (commercial sites, government organizations, and nonprofit organizations). The maps created by Shiode and Dodge focused on the spread of web server infrastructure and physical computing networks rather than on the spread of the ideas or content stored in individual web servers. Our research adopted a similar geolocation method, but focused on the dynamic keyword search results from web-search engines and their spatial relationships rather than the development of generic IT infrastructure and computer networks.

In 2001, Dodge and Kitchin published *Mapping Cyberspace,* an important research contribution to literature in cyberspace visualization. Their project overviews related topics as well as various map examples, including cyberspace spatialization, geographies of cyberspace, spatial cognition, and the cartographies of cyberspace. These maps emphasize the interactions and relationships among diverse people at various scales in cyberspace (Dodge and Kitchin 2001). Other related cyberspace visualization literatures include Börner, Chen, and Boyack (2003) and Schouten and Engelhardt (2006). The concepts of spatialization and information spaces were introduced and formalized by Fabrikant and Buttenfield (2001). Spatialization methods facilitate exploration of massive data archives with spatial

frames. The spatialization of information can create a wide variety of spatial metaphors, such as information landscapes and hotspots, to help people communicate and interact with data. Fabrikant, Montello, and Mark (2010) discussed a few problems associated with the 3D landscape metaphors in information visualization and suggested that landscape metaphor is not as self-evident as designers seem to believe. Therefore, the information landscape created in our research is constrained to 2D representations of web information landscapes rather than using 3D maps. In this article, information landscapes are defined as the visualization of spatial patterns and spatial clusters of web page density in 2D maps.

Designing the SWARMS

We designed and implemented the SWARMS prototype for creating visual maps and web information landscapes. Figure 1 illustrates the overall conceptual framework. Initial searches are conducted by using predefined keywords provided by domain experts to investigate specific topics (e.g. infectious diseases or radical concepts) by searching publically accessible websites (using the Yahoo search engine API). Then we convert the top 1000 search results into a *[Raw Text Database],* which includes all search results (ranking, titles, IP addresses, and URLs). The system uses the IP addresses and geolocation databases to convert raw text files into *[Geocoded Web Information Databases],* including both geospatial locations (latitudes and longitudes) and web information (keywords) for each hit.

By utilizing GIS software (ArcGIS), we convert the geocoded databases (created by a Microsoft SQL server) to *[Visualization Maps]* showing the information

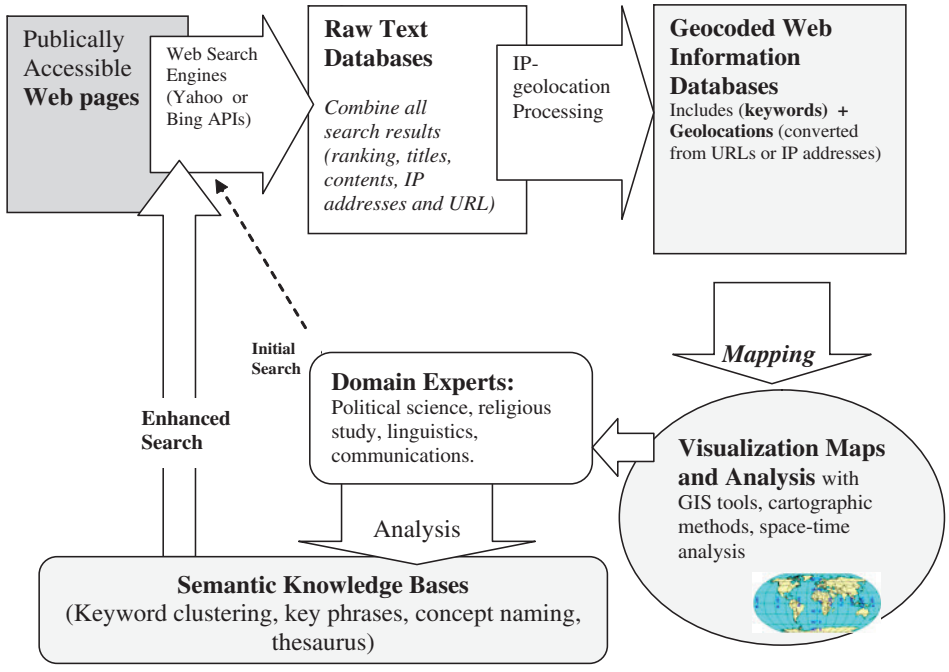

Figure 1. The SWARMS framework.

landscapes of specific ideas or keywords. We then apply advanced GIS analysis and visualization methods to understand the dynamic change of these concepts and events over space and time. Computational linguistics experts can review the resulting maps and then establish frequencies of occurrences of 'key terms,' separately and in clusters. Multiple *[Semantic Knowledge Bases]* related to ideas, concepts, and special topics can be created and revised based on the maps, which may be used in subsequent space–time analysis. The revised keyword clusters and phrases will be used for the next round of Web query process. The visualization maps constitute data for further quantitative and qualitative analysis to enrich and refine the search algorithm and to learn more about the nature and specificity of ideas and their characteristic textual architectures. This iterative process may also identify new web pages by refining keyword clusters and analyzing new information landscapes (Figure 1).

One advantage of this SWARMS framework is its flexible architecture. This framework can be used to query keywords in multiple languages (e.g. Chinese, Arabic, Spanish, or Japanese) and be used in multiple web-search engines. Figure 2 illustrates the screen shots of the keyword query interface of the SWARMS prototype. Researchers can select a search engine (from Google, Bing, or Yahoo) and type in a keyword search. The SWARMS prototype will generate the top 1000 web pages (or up to 1000 web pages) from Yahoo or Bing (or 64 web pages from Google due to the limitation of Google APIs).

Sometimes the system may not return 1000 web pages due to index limitations in search engines or incomplete geolocation databases. For example, when we tested the keyword search of 'Jerry Sanders' on 9 March 2011, the Yahoo search engine only returned 978 web pages rather than 1000 web pages. Each keyword search result table includes both the keyword and the search date. We may be able to use this information visualize and study the dynamic spread of concepts or events among different days, weeks, or months.

To demonstrate our method, we first used the keyword 'Jerry Sanders' to search web pages by Yahoo Search engine on 9 March 2011. The Yahoo API returned 978

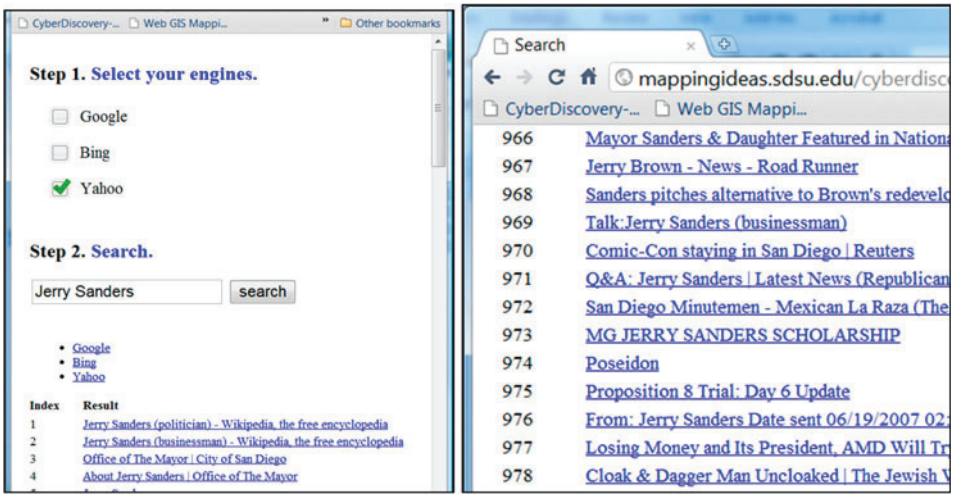

Figure 2. The SWARMS prototype interface for the keyword search of 'Jerry Sanders' and the output of the top 978 web pages from the Yahoo search engine.

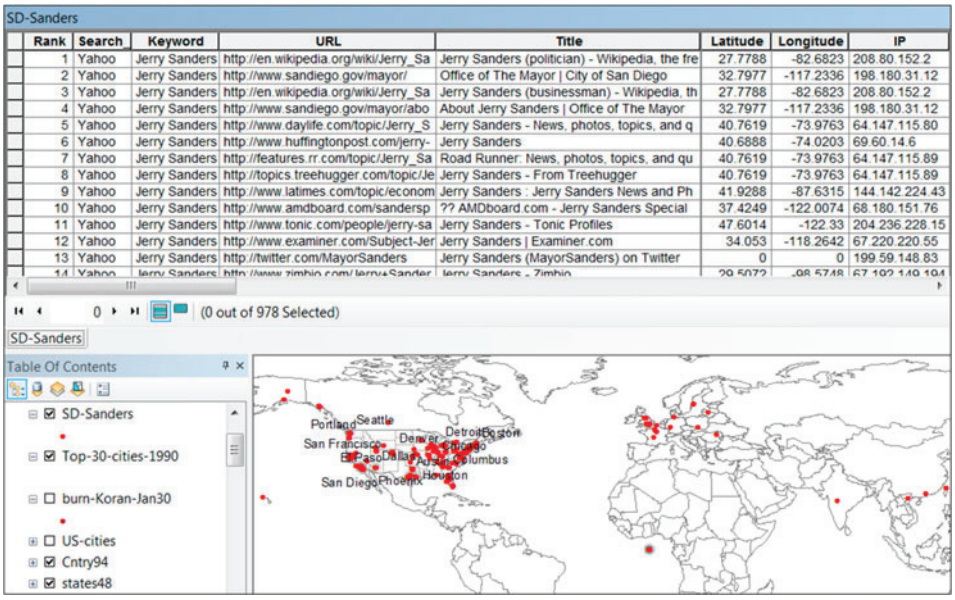

Figure 3. An example of geocoded web information databases (top) and visualization maps (using 'Jerry Sanders' as the keyword search in Yahoo).

web pages related to the keyword with ranks (Figure 3). Since 'Jerry Sanders' is the name of the mayor of the city of San Diego, California, USA, we can use these test results to verify if the spatial pattern associated with the spatial context of keywords (the City of San Diego). We converted 978 search results from the SWARMS prototype into geocoded web information databases and visualization maps using the IP geolocations associated with each web page (Figure 3).The geocoding of these 978 web pages utilized Simple Object Access Protocol and IP Address Lookup Service from the IPPage.com (http://www.ippages.com/lookups/) and the MaxMind database. Within the 978 records, 81 records were not able to generate their geographic coordinates in the IP addresses Lookup Service. The successful geolocation conversion rate of the 'Jerry Sanders' web pages was 91.7% in this test.

It is a challenging task to illustrate the spatial relationships and patterns among the 978 points from web-search results. Many cartographic representation methods could be applied to the creation of information landscapes for web pages, such as kernel density maps, choropleth maps, and graduated circle maps. Some points may be at the same or nearby locations, leading to recognition difficulty relating to scale issues or point overlap. We applied the kernel density method to illustrate the 'hotspots' and 'density' of related web pages. Figure 4 illustrates the web information landscape (web page density) created for the 'Jerry Sanders' keyword search results (with 978 points). The darker shading areas indicate higher density of web pages in the region associated with 'Jerry Sanders'.

There are various spatial analysis methods applicable for mapping web-search results, such as Thiessen (Voronoi) polygons, inverse distance weighting, or simple Kriging. But we selected the kernel density methods based on the following reasons.

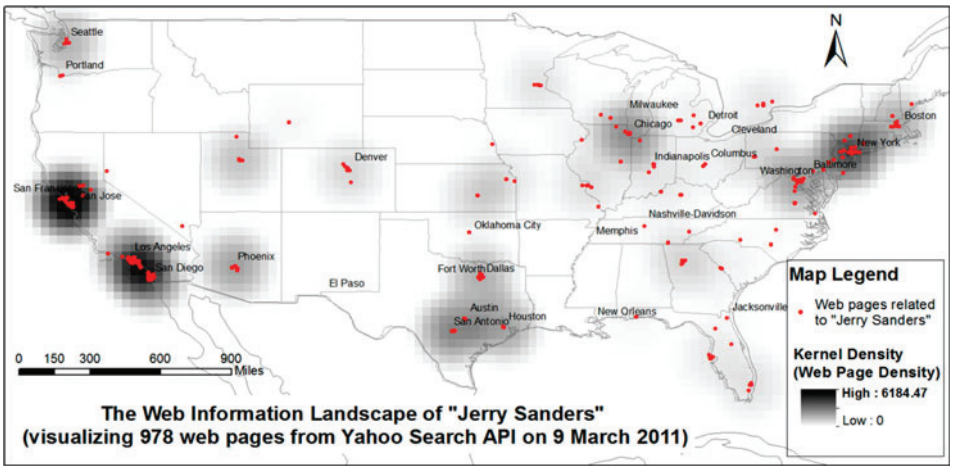

Figure 4. Creating a web information landscape (using kernel density method based upon the modified web page ranks) for the 'Jerry Sanders' search result web pages (red dots). GIS parameters: 3 map unit threshold (radius), 0.5 map unit output resolution (1 map unit equals 1 decimal degree).

(1) Many points (web pages) overlap (with the same server IP addresses, or geolocation coordinates). The kernel density method can better represent the 'density' of points in this case.
(2) It is easy and fast to calculate kernel density (available in ArcGIS Toolbox) for hundreds of points at the same time.
(3) Since the output results of kernel density maps are raster-based, we can use map algebra to calculate differential values between different keywords and in different dates.
(4) The general public is more familiar with the concepts of 'hot spots' or 'high density' created by the kernel density method. Other spatial statistic methods are less intuitive.

It is easy to produce different spatial output resolutions and generalization thresholds (radius) based on analysis needs. We then performed the kernel density function in ArcGIS, specifying a 3 map unit threshold (radius) and 0.5 map unit output scale. Map unit is defined by the data frame used in GIS software (ArcGIS 10, 2012). In this example, one map unit represents one decimal degree, approximately 80 km (50 miles) in California. We used 3 map units (around 150 miles) to reflect the average size of metropolitan areas (cities). In our testing results, we found out that the following settings of kernel density thresholds can visualize more meaningful spatial patterns at different map scales.

- 6–8 map units (480 km) for detecting the state-level geospatial patterns.
- 2–3 map (160 km) units for detecting the county-level geospatial patterns.
- 0.5–1 map units (40 km) for detecting the city-level geospatial patterns.
- 0.1–0.2 map units (8 km) for detecting the zip code-level geospatial patterns.

The red dots indicate the locations of IP addresses associated with the web pages searched by the keyword (Jerry Sanders). In our design, the ranking numbers of search results were considered as the 'popularity' or the 'population' in the kernel density algorithm. A higher-ranked web page is more 'popular' and more 'visible' comparing to a lower-ranked web page. Therefore, we converted the ranking numbers into the population parameter:

$$\text{Population} = (\text{Total number of web pages} + 1) - \text{rank\#} \quad (1)$$

For example, a web page ranked #1 in a set of 1000 web pages was assigned to '1000' $(1000 + 1 - 1)$ for its population parameter. A web page ranked # 900 was assigned to '101' $(1000 + 1 - 900 = 101)$ for its population parameter. We used a black–white color scheme to represent unclassified kernel density from the minimum population (density) value (0) to the maximum population (density) value (6184.47; Figure 4).

Although these web page search results from Yahoo search engine cover the whole world, most web pages in such a query are located in the United States due to the language of keywords (in English). We are planning to study international patterns and keywords in other languages in the future. However, most SWARMS mapping and analyses thus far have only focused on the spatial distribution patterns in the United States with English keywords. In the case of the mayor names, two interesting spatial patterns emerge (Figure 4). First, two major hotspots of 'Jerry Sanders' are located in the [San Jose-San Francisco] and [Los Angeles-San Diego] metropolitan areas. The hotspot near San Diego is significant as this is the pattern we are looking to find; cyberspace activity is reflective of real-world phenomena. The hotspot near San Francisco is significant as it is unexpected. Is the San Francisco result due to noise within the SWARMS framework, or is this result new information about the impact of Jerry Sanders in the San Francisco area? We believe that the pattern is due to large number of hosting businesses in the San Francisco area, as will be explored later in the discussion of a 'background' map and displayed in Figure 5.

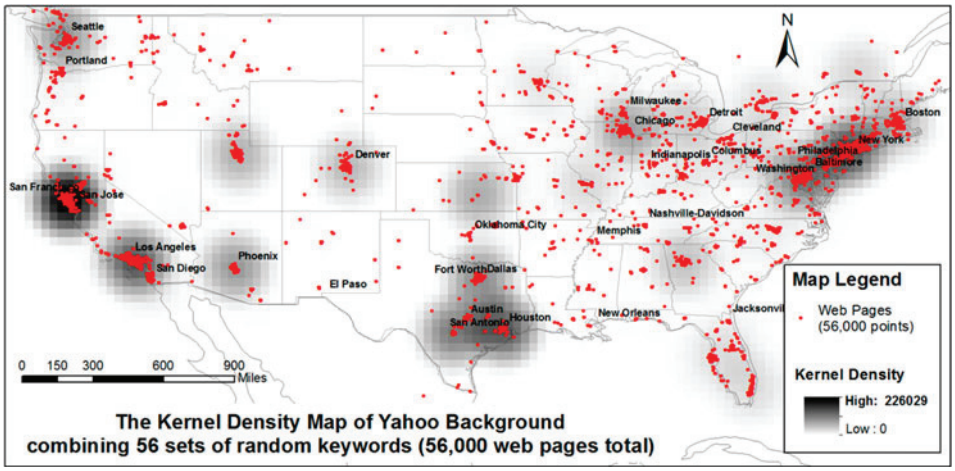

Figure 5. The background web information landscape (by using 56,000 web pages for kernel density map).

Second, most web page locations are associated with major US cities. This indicates that the density of web pages may be closely related to the size of city populations. Bigger cities will have higher density of web pages associated with selected keywords. The next sections will discuss the spatial relationship between web pages and IP addresses and illustrate some prominent GIS analysis methods we developed for the visualization of web information landscapes.

Understanding the spatial relationships between web page contents and IP address geolocations

One important aspect of our research is to identify the spatial relationship between the content of web pages (such as city mayors, breaking news, or special events) and the geolocations of IP addresses associated with the web pages (web servers). To analyze the spatial relationship between the content of ranked web pages and the locations of IP addresses, our research team developed a series of manual classification processes to identify various spatial relationships between the contents of web pages, associated companies/organization, and the geolocation of IP addresses. We applied this manual coding scheme to reclassify top 500 web pages from the 'Jerry Sanders' keyword search results and found out that 65% web pages have 'explicit' spatial information in their contents and 20% web pages have 'implicit' spatial information requiring readers to use other resource to figure out the spatial information of associated organization. We were unable to find any content-related spatial information for only 15% of the web pages. However, within the web pages having either explicit and implicit spatial information (85% total), only 35% of the web pages have the same spatial locations between the content of web pages (such as companies or authors) and the IP address geolocations. Sixty-five percent of the web pages do not have any spatial association between their web contents and their IP address geolocations.

One emerging research question from this observation is whether we visualize any unique spatial pattern in web information landscapes by using data with mixed geolocation accuracy. If the search engine generates 1000 web pages from one selected keyword, only 350 web pages (points in the map) have meaningful spatial association with their IP geolocation and the rest of 650 web pages (points in the map) have inaccurate spatial association with IP geolocation or missing spatial information. Can the 350 meaningful web pages illustrate significant spatial pattern in our web information landscape? To test this, we compared the visualization map for 'Jerry Sanders' with all of the records to the visualization map of 'Jerry Sanders' with only the meaningful spatial association. For both of the images, the location with the highest density was San Diego, CA. This implies that while there may be some 'noise' in the 65% of the web pages that have inaccurate spatial association, the spatial 'signal' may still be visible and meaningful for further study.

Our design is to create a 'background' web information landscape similar to the 65% meaningless (spatially) web pages. Then we can compare the differences between the keyword-generated web information landscape and the background landscape. By using 168 randomly chosen keywords and removing some stop words, we create 56 sets of random keywords (such as 'most', 'As', 'possible', 'himself', 'Sue', 'young', 'so', '61', 'sort', 'the', 'so', 'B', 'too', 'age'). We started with the British National Corpus and parsed the corpus to obtain a dependency database in order to develop a

general background model. Around 200 keywords were randomly chosen from the database and were cleaned of punctuation noise (no results of '?' or '.'). From this processed set of 168 words, we strung together 56 sets. Fifty-six thousand web pages were created by Yahoo Search engines and we used their IP geolocations to create the 'background' web information landscape associated with Yahoo search engine (Figure 5).

Our next step is to calculate the differences between the keyword map and the background map. A raster-based map algebra tool from ArcGIS was used with the following formula:

$$\text{Differential Value} = (\text{Keyword-A}/\text{Maximum-Kernel-Vlue-of-Keyword-A}) - (\text{Background}/\text{Maximum-Kernel-Value-of-Background}) \quad (2)$$

Figure 6 illustrates the map algebra results showing the differential value (ratio) between the keyword map ('Jerry Sanders') and the background map. The red color indicates that the keyword web pages have higher kernel density ratio in the region comparing to the background ratio. The blue color indicates that the keyword web pages have lower kernel density ratio in the region comparing to the background ratio. The spatial pattern in Figure 6 clearly reflects the spatial context of 'Jerry Sanders' – the City of San Diego.

We also select three additional city mayors' names ('Michael McGinn' in Seattle, 'Rahm Emanuel' in Chicago, and 'Sam Adams' in Portland) to perform the same differential analysis. The differential maps in all three keywords show strong spatial contexts of the mayor's names: highest kernel density ratio (red hot spots) in each mayor's city (Figure 7).

Although the three differential maps all illustrated the red hot spots (higher kernel density ratios comparing to the background ratio), each map has its unique spatial pattern. For example, McGinn's map only has one major hot spot (Seattle). Emanuel's map has several hot spots, including Chicago, Dallas, and Washington

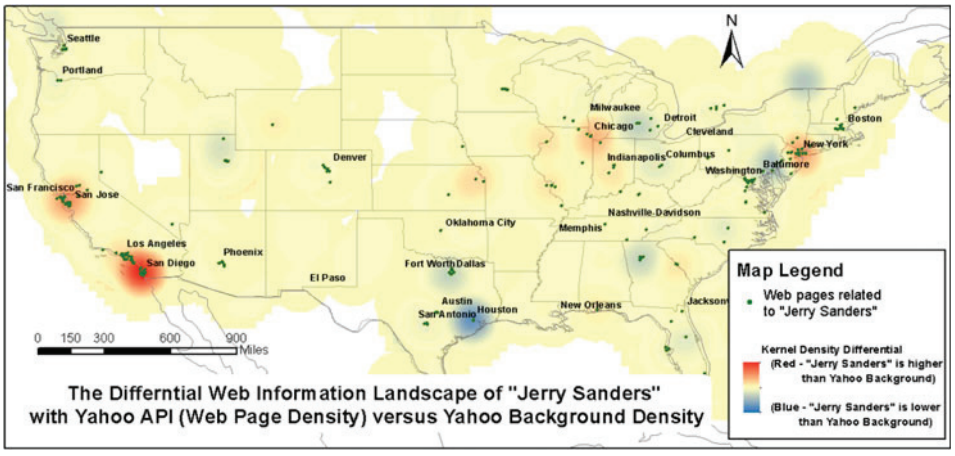

Figure 6. The differential web information landscape of 'Jerry Sanders' versus Background Information Landscape (with Yahoo Search Results).

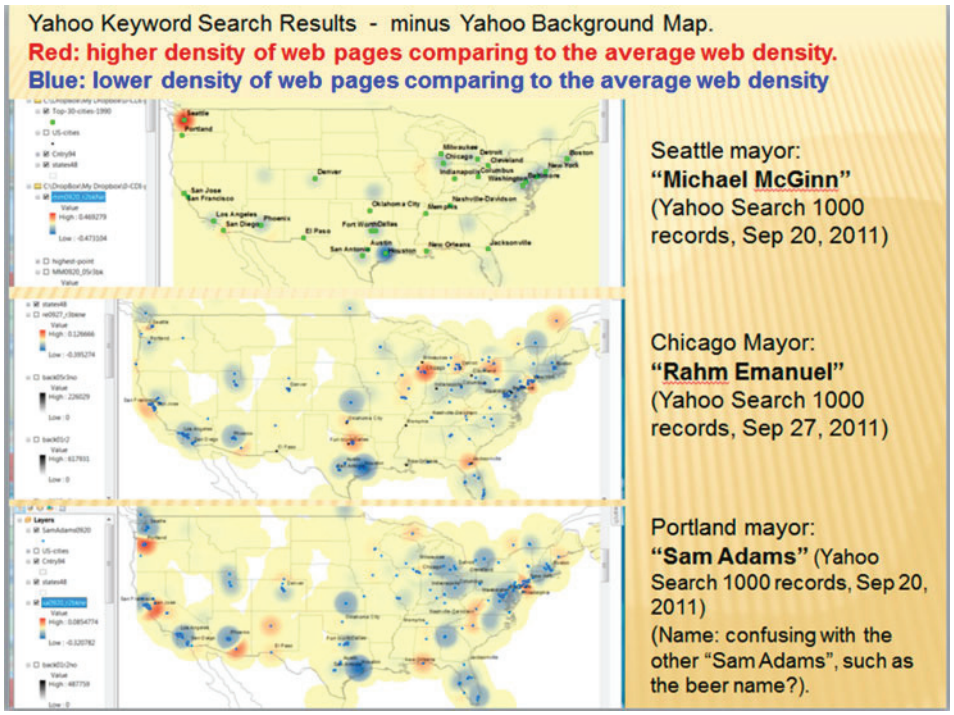

Figure 7. The Differential Web Information Landscapes of three city mayors (McGinn, Emanuel, and Adams).

DC. These spatial patterns may reflect the characteristics of these mayors since Emanuel was the former White House Chief of Staff in Washington DC.

This SWARMS framework can be used to query keywords in multiple languages (e.g. Chinese, Arabic, Spanish, or Japanese) and be used in multiple web-search engines. In our early tests, we only used English in our keyword search. Different languages may create significantly different web information landscapes. Figure 8 illustrates the global distribution pattern of the 'Osama bin Laden' keyword search in three different languages (English, Chinese (simplified), and Arabic). The global distributions of web pages about 'Osama bin Laden' are quite different between the three maps. Further language-specific analysis will be required for understanding the meaning of these spatial pattern language variations.

Conclusion

We present a new methodology and a research framework for analyzing the dynamic web information landscape and tracking the spread of ideas through web-based keyword searches. The SWARMS prototype can convert traditional text-based web-search results into web information landscapes. The acquired geospatial fingerprints and spatial patterns in differential web information landscapes may illustrate hidden semantic or contextual meanings associated with different keywords and concepts. For the first keyword example, 'Jerry Sanders' has a strong semantic link to the City

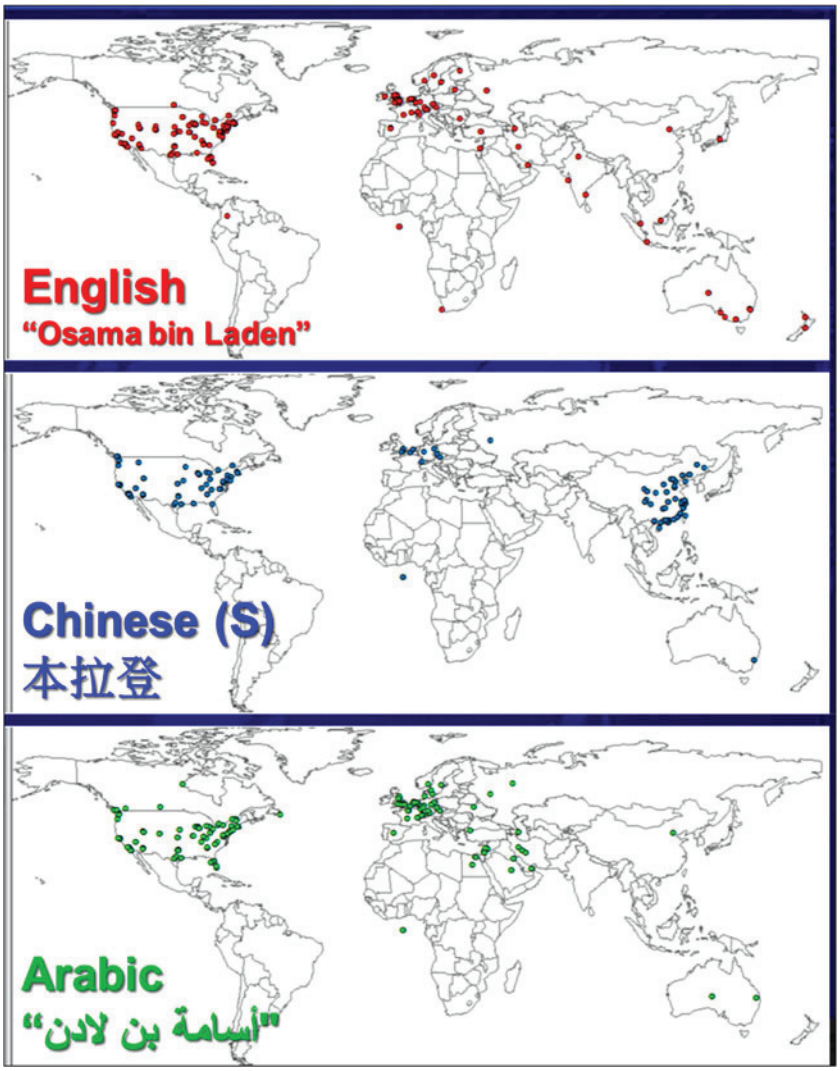

Figure 8. The global distribution patterns of the keyword search 'Osama bin Laden' in three different languages (English, Chinese (simplified), and Arabic).

of San Diego. This approach may provide a new research direction for studying human thought, web content, and communication theories.

One major motivation of this research project is to test the First Law of Geography – that *'everything is related to everything else, but near things are more related than distant things'* (Tobler 1970), in the arena of cyberspace. Our research team aimed to validate the first law of geography with our SWARMS prototype and the differential web page density maps. We also realize that our SWARM prototype contained several shortcomings:

(1) Our prototype relied on commercial search engine APIs, such as the Yahoo BOSS API and Bing API. The actual searching and ranking algorithms in

their engines cannot be revealed due to commercial value concerns. Although we can know their approximate methodologies based on public technical reports or announcements, the actual procedures are still in a 'black box' for researchers. Moreover, these search engines frequently change algorithms and APIs.

(2) It is difficult to find appropriate 'identifiers' or keywords to represent one concept or one item. We might need to search for all related combinations of keywords in order to obtain comprehensive coverage of web information landscapes. Moreover, these different terms may have deeper meanings and may be used by different groups or societies. We need to understand these variations from a computational linguistics perspective and perform more testing and comparison to help us choose the right combination of keywords.

(3) We need to use real-world data to validate our methods and research frameworks. The algorithms and equations introduced in this paper may need to be calibrated to fit different scenarios and different applications. We are currently working on epidemiology cases (flu and whooping cough) and political election cases (2012 Presidential Election).

To summarize, our preliminary maps indicate that there are strong spatial relationships between the activities in cyberspace and real-world locations of related web pages. More advanced spatial analysis methods and keyword search methods will need to be applied to help us understand deeper relationships, spatial patterns, and spatial statistics interpretations. For instance, we may apply survival analysis to calculate the hazard (risk) of a certain location being influenced by a certain event or idea, and link such ideas to various biophysical, socioeconomic, and demographic factors to better understand the mechanisms behind the observed information patterns over space and time (An and Brown 2008). In addition, a set of new metrics and analytical methods needs to be developed to better characterize, analyze, and understand the space–time trajectories of the related events/ideas diffusing over the web.

Thanks to the massive power of computers and the Internet to copy and transform data across the globe and facilitate the rapid spread of movements and ideas, the world is facing both challenges and opportunities. The existence of such movements is not new, nor is their capacity for finding receptive audiences in new locales, but the rapidity with which they spread and take root in new soil may well be a unique feature of the information age. Quantitative changes in the speed of audience growth or turnover may be accompanied by qualitative changes in the audiences themselves. Fortunately, the very technology that promotes the rapid spread of ideas is also providing the tools to understand them. A better understanding of the spatial and temporal dynamics of the 'collective thinking of human beings' over the Internet could lead to improved comprehension of the factors behind those ideas. Such insight is important in reducing misunderstandings and strategizing how to address controversies and conflicts.

References

Amitay, E., N. Har'el, R. Sivan, and A. Soffer. 2004. "Web-a-where: Geotagging Web Content." In *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR04)*, 273–280. Sheffield: ACM Press.

An, L., and D. G. Brown. 2008. "Survival Analysis in Land-change Science: Integrating with GIScience to Address Temporal Complexities." *Annals of the Association of American Geographers* 98 (2): 323–344. Accessed February 3, 2012. http://dx.doi.org/10.1080/00045600701879045

ArcGIS 10. 2012. Redlands, CA: Environmental Systems Research Institute (ESRI).

Bai, X. 2011. "Predicting Consumer Sentiments from Online Text." *Decision Support Systems* 50 (4): 732–742. Accessed May 8, 2012. http://www.sciencedirect.com/science/article/pii/S016792361000148X

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. "The Semantic Web." *Scientific American* 284 (5): 34–43. Accessed May 8, 2012. http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Börner, K., C. Chen, and K. W. Boyack. 2003. "Visualizing Knowledge Domains." *Annual Review of Information Science and Technology* 37 (1): 179–255. Accessed May 8, 2012. http://dx.doi.org/10.1002/aris.1440370106

Brown, C. 2009. "WWW.HATE.COM: White Supremacist Discourse on the Internet and the Construction of Whiteness Ideology." *Howard Journal of Communications* 20 (2): 189–208. Accessed February 3, 2012. http://dx.doi.org/10.1080/10646170902869544

Brownstein, J. S., C. C. Freifeld, B. Y. Reis, and K. D. Mandl. 2008. "Surveillance Sans Frontières: Internet-based Emerging Infectious Disease Intelligence and the HealthMap Project." *PLoS Med* 5 (7): e151. Accessed May 8, 2012. http://dx.doi.org/10.1371%2Fjournal.pmed.0050151

Buyukokkten, O., J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. 1999. "Exploiting Geographical Location Information of Web Pages." In ACM SIGMOD Workshop on the Web and Databases (WebDB'99). Philadelphia, PA: ACM Press.

Choi, W.-J., and S. Tekinay. 2003. "Location-based Service Provisioning for Next Generation Wireless Networks." *International Journal of Wireless Information Networks* 10 (3): 127–139. Accessed May 8, 2012. http://dx.doi.org/10.1023/B:IJWI.0000007816.59854.f2

Chute, C. G. 2008. "Biosurveillance, Classification, and Semantic Health Technologies." *Journal of the American Medical Informatics Association* 15 (2): 172–173. Accessed May 8, 2012. http://jamia.bmj.com/content/15/2/172.short

Collier, N., R. M. Goodwin, J. Mccrae, S. Doan, A. Kawazoe, M. Conway, A. Kawtrakul, K. Takeuchi, and D. Dien. 2010. "An Ontology-driven System for Detecting Global Health Events." In *23rd International Conference on Computational Linguistics*, 215–222. Beijing: Association for Computational Linguistics.

Corman, S. R., T. Kuhn, R. D. Mcphee, and K. J. Dooley. 2002. "Studying Complex Discursive Systems: Centering Resonance Analysis of Communication." *Human Communication Research* 28 (2): 157–206. Accessed May 8, 2012. http://dx.doi.org/10.1111/j.1468-2958.2002.tb00802.x

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

Elmer, G. 2006. "Mapping the Cyber-stakeholders: U.S. Energy Policy on the Web." *The Communication Review* 9 (4): 297–320. Accessed May 8, 2012. http://dx.doi.org/10.1080/10714420600957282

Fabrikant, S. I., and B. P. Buttenfield. 2001. "Formalizing Semantic Spaces for Information Access." *Annals of the Association of American Geographers* 91 (2): 263–280. Accessed May 8, 2012. http://dx.doi.org/10.1111/0004-5608.00242

Fabrikant, S. I., D. R. Montello, and D. M. Mark. 2010. "The Natural Landscape Metaphor in Information Visualization: The Role of Commonsense Geomorphology." *Journal of the American Society for Information Science and Technology* 61 (2): 253–270. Accessed May 8, 2012. http://dx.doi.org/10.1002/asi.21227

Fleishman, G. 1996. "Web Log Analysis: Who's Doing What, When?" *Web Developer* 2 (2). Accessed February 6, 2008. http://www.webdeveloper.com/management/management_log_analysis.html

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (7232): 1012–1014. Accessed May 8, 2012. http://dx.doi.org/10.1038/nature07634

Himmelstein, M. 2005. "Local Search: The Internet is the Yellow Pages." *Computer* 38 (2): 26–34. doi:10.1109/MC.2005.65.

Hoffman, D. 2005. "Violent Events as Narrative Blocs: The Disarmament at Bo, Sierra Leone." *Anthropological Quarterly* 78 (2): 329–353. Accessed May 8, 2012. http://www.jstor.org/stable/4150837

Joho, H., and M. Sanderson. 2004. "The SPIRIT Collection: An Overview of a Large Web Collection." *ACM SIGIR Forum* 38 (2): 57–61. doi:10.1145/1041394.1041395.

Jones, C. B., and R. S. Purves. 2009. "Geographical information retrieval." In *Encyclopedia of Database Systems*, edited by L. Liu and M. T. Zsu. Springer Publishing Company.

Kempe, D., J. Kleinberg, and É. Tardos. 2003. "Maximizing the Spread of Influence Through a Social Network." In *9th ACM Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining*, 137–146. Washington, DC: ACM.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–723. Accessed May 8, 2012. http://www.sciencemag.org/content/323/5915/721.short

Lee, A. 2008. "Internet Press Freedom and Online Crisis Reporting: The Role of News Websites in the SARS Epidemic." In *The Social Construction of SARS: Studies of a Health Communication Crisis*, edited by J. H. Powers and X. Xiao, 69–90. Amsterdam: John Benjamins.

Leets, L., and P. J. Bowers. 1999. "Loud and Angry Voices: The Insidious Influence." *Communication Monographs* 66 (4): 325–340. Accessed May 8, 2012. http://dx.doi.org/10.1080/03637759909376483

Little, M., C. F. C. Jordens, and E.-J. Sayers. 2003. "Discourse Communities and the Discourse of Experience." *Health: An Interdisciplinary Journal for the Social Study of Health, Illness and Medicine* 7 (1): 73–86. Accessed May 8, 2012. http://hea.sagepub.com/content/7/1/73.abstract

Markowetz, A., T. Brinkhoff, and B. Seeger. 2003. "Exploiting the Internet as a Geospatial Database." In *Post-Workshop Book of International Workshop on Next Generation Geospatial Information*, Cambridge, MA.

Matsumura, N., and Y. Sasaki. 2007. "Human Influence Network for Understanding Leadership Behavior." *International Journal of Knowledge-based and Intelligent Engineering Systems* 11 (5): 291–300. Accessed May 8, 2012. http://iospress.metapress.com/content/A91300694N378M45

Miller, G. 2011. "Social Scientists Wade into the Tweet Stream." *Science* 333 (6051): 1814–1815. Accessed May 8, 2012. http://www.sciencemag.org/content/333/6051/1814.short

Monge, P. R., and N. S. Contractor. 1998. "Emergence of Communication Networks. In *Handbook of Organizational Communication*. 2nd ed, edited by F. M. Jablin and L. L. Putnam, 440–502. Thousand Oaks, CA: Sage.

Muir, J. A., and P. C. V. Oorschot. 2009. "Internet Geolocation: Evasion and Counter-evasion." *ACM Computing Surveys* 42 (1): 1–23. doi:10.1145/1592451.1592455.

Ohsawa, Y., H. Soma, Y. Matsuo, N. Matsumura, and M. Usui. 2002. "Featuring Web Communities Based on Word Co-occurrence Structure of Communications." In *11th International World Wide Web Conference*. Honolulu, HI: ACM.

Papacharissi, Z. 2009. "The Virtual Geographies of Social Networks: A Comparative Analysis of Facebook, LinkedIn and ASmallWorld." *New Media & Society* 11 (1–2): 199–220. Accessed May 8, 2012. http://nms.sagepub.com/content/11/1-2/199.abstract

Poese, I., S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. 2011. "IP Geolocation Databases: Unreliable? *ACM SIGCOMM Computer Communication Review* 41 (2): 53–56. doi:10.1145/1971162.1971171.

Purves, R. S., P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, et al. 2007. "The Design and Implementation of SPIRIT: A Spatially Aware Search Engine for Information Retrieval on the Internet." *International Journal of Geographical Information Science* 21 (7): 717–745. Accessed May 8, 2012. http://dx.doi.org/10.1080/13658810601169840

Schouten, B., and Y. Engelhardt. 2006. "Network Nations." In *Else/where: Mapping – New Cartographies of Networks and Territories*, edited by J. Abrams and P. Hall, 64–67. Minneapolis: University of Minnesota Press.

Shavitt, Y., and N. Zilberman. 2010. A Study of Geolocation Databases. Accessed February 5, 2012. http://arxiv.org/abs/1005.5674

Shekhar, S., and D. Oliver. 2010. "Computational Modeling of Spatio-temporal Social Networks: A Time-aggregated Graph Approach." *Spatio-Temporal Constraints on Social Networks Specialist Meeting*. Santa Barbara: University of California.

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

Silva, M. J., B. Martins, M. Chaves, A. P. Afonso, and N. Cardoso. 2006. "Adding Geographic Scopes to Web Resources." *Computers, Environment and Urban Systems* 30 (4): 378–399. Accessed May 8, 2012. http://www.sciencedirect.com/science/article/pii/S0198971505000608

Singh, V. K., M. Gao, and R. Jain. 2010. "Social Pixels: Genesis and Evaluation." In *International Conference on Multimedia (MM'10)*, 481–490. Firenze: ACM.

Svantesson, D. J. B. 2005. "Geo-identification: Now They Know Where You Live." *Privacy Law & Policy Reporter* 11 (6): 171–174.

Takhteyev, Y., A. Gruzd, and B. Wellman. 2012. "Geography of Twitter Networks." *Social Networks* 34 (1): 73–81. doi:10.1016/j.socnet.2011.05.006.

Tobler, W. R. 1970. "A computer movie simulating urban Growth in the Detroit Region." *Economic Geography* 46 (2): 234–240. doi:10.2307/143141.

Tsou, M.-H., and I. H. Kim. 2010. "Increasing Spatial Awareness by Integrating Internet Geographic Information Services (GIServices) with Real Time Wireless Mobile GIS Applications." *International Journal of Strategic Information Technology and Applications* 1 (4): 42–54. doi:10.4018/jsita.2010100103.

Turner, A. 2004. "Geolocation by IP Address." *Linux Journal* 6. Accessed May 8, 2012. http://www.linuxjournal.com/article/7856

Varian, H. R., and H. Choi. 2009. "Predicting the Present with Google Trends." *Google Research Blog*. Accessed May 8, 2012. http://googleresearch.blogspot.com/2009/04/predicting-present-with-google-trends.html

Wellman, B., J. Salaff, D. Dimitrova, L. Garton, M. Gulia, and C. Haythornthwaite. 1996. "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community." *Annual Review of Sociology* 22: 213–238. Accessed May 8, 2012. http://www.jstor.org/stable/2083430

Worboys, M. 2010. "Responsive Social Networks." *Spatio-Temporal Constraints on Social Networks Specialist Meeting*. Santa Barbara: University of California.

World Wide Web Consortium. 2010. "Geolocation API Specification, W3C Candidate Recommendation 07 September 2010." *World Wide Web Consortium (W3C)*. Accessed March 29, 2011. http://www.w3.org/TR/2010/CR-geolocation-API-20100907

Youn, I., B. L. Mark, and D. Richards. 2009. "Statistical Geolocation of Internet Hosts." In *IEEE International Conference on Computer Communications and Networks (ICCCN)*, 1–6. San Francisco, CA: IEEE.

Zandbergen, P. A. 2009. "Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning." *Transactions in GIS* 13: 5–25. Accessed May 8, 2012. http://dx.doi.org/10.1111/j.1467-9671.2009.01152.x