(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 August 2013 (22.08.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/123392 A1

(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/US2013/026433

(22) International Filing Date:
15 February 2013 (15.02.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/599,648 16 February 2012 (16.02.2012) US

(71) Applicant: **SAN DIEGO STATE UNIVERSITY RESEARCH FOUNDATION** [US/US]; 5250 Campanile Drive, San Diego, California 92182 (US).

(72) Inventor; and
(71) Applicant : **TSOU, Ming-Hsiang** [US/US]; 8392 Lake Artemus Ave., San Diego, California 92119 (US).

(74) Agents: **BILLION, Richard E.** et al.; 7401 Metro Blvd., Suite 425, Minneapolis, Minnesota 55439 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) Title: METHOD AND APPARATUS FOR VISUALIZING GEOSPATIAL FINGERPRINTS ON WEB INFORMATION LANDSCAPES

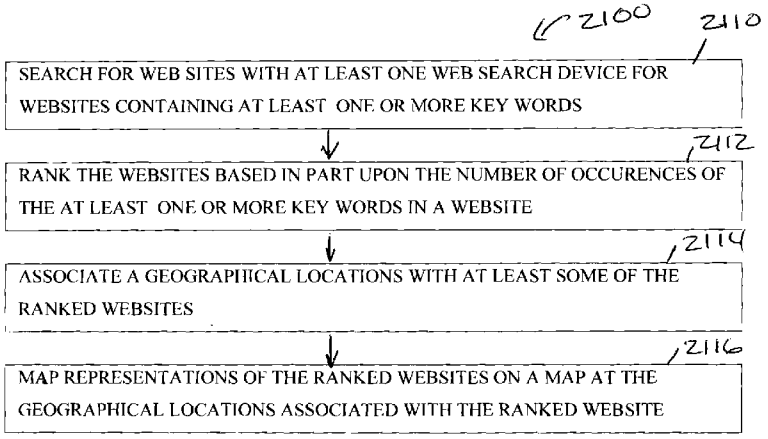


FIG. 21

(57) Abstract: A method includes searching for web sites with at least one web search device for websites containing at least one or more key words, and ranking the websites based in part upon the number of occurrences of the at least one or more key words in a website. The method also includes associating a geographical location with at least some of the ranked websites, and mapping representations of the ranked websites on a map at the geographical locations associated with the ranked website. Also disclosed is an apparatus for carrying out the above method.

WO 2013/123392 A1

METHOD AND APPARATUS FOR VISUALIZING GEOSPATIAL FINGERPRINTS ON WEB INFORMATION LANDSCAPES

Related Applications:

This application claims the benefit under 35 U.S.C. §119(e) of prior U.S. Provisional Patent Application No. 61/599,648 filed February16, 2012, which is incorporated herein by reference.

Government Support Clause:

This invention was made with government support under Grant Number 1028177 awarded by National Science Foundation. The United States government has certain rights in the invention.

Field of the Invention:

Various embodiments described herein relate to a Method and Apparatus for Visualizing Geospatial Fingerprints on Web Information Landscapes.

Summary of the Invention:

A new method for visualizing and analyzing information landscapes of ideas and events posted on public web pages through customized web search engines is disclosed. This research integrates GIScience, computational linguistics, and web search engines to track and analyze public web pages and associated web contents. 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 with time stamps. By creating a Spatial Web Automatic Reasoning and Mapping System (SWARMS) prototype, one can can visualize the spread of concepts, ideas and news on the Web over time and space. The Web, as the collective thought of human communication, can provide valuable insight into the spread of diseases, controversial concepts, or radical movements. By analyzing multiple web information landscapes with kernel density methods and map algebra tools, web information landscapes can be created showing the density of web pages. These maps reveal important "geospatial fingerprints" for selected keywords reflecting semantic constructs. The revealed geospatial fingerprints and unique spatial patterns can illustrate hidden semantic or contextual meanings associated with different keywords. This approach can provide a new research

direction for geographers to study human thought and behavior, global web content, and internet communication theories. The spatial-temporal analysis of web content can help people understand the diversity of human concepts in a global scale and can be applied in many applications, including business marketing, homeland security, public policy making, and public health.

Brief Description of the Drawings

The embodiments will be readily understood by the following detailed description in conjunction with the accompanying drawings, wherein like reference numerals designate like structural elements, and in which:

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

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).

Figure 4. The web information landscape (a kernel density map based upon the modified website ranks) for the top 978 "Jerry Sanders" search web pages (red dots) using a 3 map unit threshold (radius) and 0.5 map unit output scale (one map unit equals one decimal degree).

Figure 5. The differential popularity between the San Diego mayor (red dots) and the Los Angeles mayor (blue dots) using a 3 map unit threshold (radius) and 0.5 map unit output scale.

Figure 6. The popular web pages of "Antonio Villaraigosa" located around Denver, Colorado.

Figure 7. Comparing six different settings of radius distances (thresholds) and output grids in the differential maps between "Jerry Sanders"(Red) and "Antonio Villaraigosa" (Blue).

Figure 8. The kernel density of "burn Koran" keyword search results (top) and the U.S. city population.

Figure 9. The differential information landscape of "burn Koran" versus U.S. city population density, with the location of two event centers, Topeka, Kansas and Gainesville, Florida (web search results on 30 January 2011).

Figure 10. Yahoo Background Maps (with 56,000 web pages).

Figure 11. The differential map between "burn Koran" on Jan 30 (1000 pages) and Yahoo background web density map (baseline map).

Figure 12. The differential information landscape of "burn Koran" between 30 January 2011 and 3 April 2011. Red color indicates the increasing population of "burn Koran" web pages in April 2011 compared to the popularity in January 2011. Blue color indicates the decreasing population of "burn Koran" web pages in April 2011 compared to January 2011.

Figure 13. Categorizing web pages for different search engine comparisons.

Figure 14. The comparison of different search engines and different keywords.

Figure 15. The identical web pages between Yahoo and Bing using the keyword "Jerry Sanders".

Figure 16. Comparison of web information differential landscape: Yahoo API results (top), Bing API results (bottom) (standardized by the Yahoo background map).

Figure 17. The temporal change comparison of Yahoo API search results for different dates and with different keywords.

Figure 18. The global distribution patterns of the keyword search "Osama bin Laden" in three different languages (English, Chinese (simplified), and Arabic).

Figure 19 is a schematic diagram of computer that includes several computer subsystems, according to an example embodiment.

Figure 20 shows a diagrammatic representation of a computing device for a machine in the example electronic form of a computer system, according to an example embodiment.

Figure 21 is a flow chart of a method of gathering and displaying data from websites, according to an example embodiment.

Figure 22 is a flow chart of a method of combining and reranking websites after conducting key word searches using two search engines, according to an example embodiment.

Figure 23 is a flow chart of a method of gathering and displaying data from websites, according to an example embodiment.

Additional FIGs describe further example embodiments of the SWARMS system.

Detailed Description

In the following description, numerous specific details are set forth to provide a thorough understanding of the concepts underlying the described embodiments. It will be apparent, however, to one skilled in the art that the described embodiments may be practiced without some or all of these specific details. In other instances, well known process steps have not been described in detail in order to avoid unnecessarily obscuring the underlying concepts.

Introducing Web Information Landscapes

The world today is constantly awash with a flood of ideas, and the diffusion of these ideas now leaves measurable traces in cyberspace that can be mapped onto realspace and in near real time. Introduced herein is a new research framework for web keyword searches and web page content analysis, called Spatial Web Automatic Reasoning and Mapping System (SWARMS), to track ideas, events, and trends disseminated in cyberspace (the web and social media). In this article, we define "the web" as the connected Internet and its broader network-based applications, include the World Wide Web, instant messengers, FTP servers, social media, web services, etc. 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 new SWARMS prototype can helps visualize and analyze the space-time dimensions of the spread of information, concepts, and ideas posted on the publically-accessible web pages. Hundreds of web pages were geocoded with real world coordinates and represented in the form of web information landscapes (web page density maps). These web information landscapes can help us monitor the spatial and temporal distribution patterns of web pages and reveal the nature of significant events, controversial concepts or epidemics. Understanding the diffusion and acquisition patterns of web information landscapes in response to disasters, terrorism, and epidemics has the potential to facilitate intervention and response, and eventually, prevention.

The SWARMS prototype is designed to track spatial patterns of publically-accessible web pages based upon searching clusters of keywords determined by domain experts. In one embodiment, clusters of keywords are predefined. The Web pages and web content associated with the same keywords are 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 including 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, the SWARM prototype can also track the social impact of significant events over time as they are reflected in cyberspace. The Web, linking millions of networks and billions of people, has become an important base of computer-supported social networks (CSSNs). This concept extends the scope of spatial analysis from physical world phenomena to cyberspace contents.

Discussed first are the methods used by a computer system to implement a system capable of identifying concepts and mapping those concepts to produce visual representations of these various concepts, ideas, and the like. The methods include conducting searches using one or more search engines and will be detailed below. This discussion will be followed by more specific examples, such as an example of SWARMS with a few selected keywords, including "Jerry Sanders" (the mayor of San Diego, California), "Antonio Villaraigosa" (the mayor of Los Angeles, California), "burn Koran", "Osama bin Laden", etc. The web search results and associated maps between two popular search engines, Yahoo and Bing. 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), the Google search results were not

included in this paper. Detailed spatial analysis with web information landscapes will be discussed later.

Methods

Figure 21 is a flow chart of a method 2100 of gathering and displaying data from websites, according to an example embodiment. The method 2100 includes searching for web sites with at least one web search device for websites containing at least one or more key words 2112, and ranking the websites based in part upon the number of occurences of the at least one or more key words in a website 2114. The method 2100 also includes associating a geographical locations with at least some of the ranked websites, and mapping representations of the ranked websites on a map at the geographical locations associated with the ranked website 2116.

Figure 22 is a flow chart of a method 2200 of combining and reranking websites after conducting key word searches using two search engines, according to an example embodiment. In one embodiment, searching for web pages includes searching web pages for at least one or more key words using a first search engine and a second search engine 2210. The search engine can include any type of search engine. For example, there are search engines available from Yahoo of Sunnyvale,CA; Bing of Bellvue, WA; and Google of Mountain View, CA as well as others. Generally, these search engines produce a limited number of top results. Several of these have an Application Programming Interface(API) that is available to increase the number of top results. For example, the Yahoo and Bing search engines each have an API that can increase the number of top search results to 1000. In one embodiment, the APIs that extend the number of top search results that are delivered are used in the key word searches of websites. In this embodiment, the websites are ranked based upon the number of occurences of the key word or words found by the search engine. Ranking the websites further includes ranking the websites found by a first search engine 2212, ranking the websites found by a second search engine 2214, and combining the websites found by the first search engine and the second search engine and reranking the combined list of websites found by the two websites 2216. In one example embodiment, the top 1000 occurences of a particular key word search term for Bing is ranked by giving the top website a rank of 1000 and the bottom website a rank of 1. The same can be done for the top 1000 occurences of a particular key word search term from the Yahoo search engine. These ranking values can be added. The combined listing can be sorted from high to low based on

the new combined ranking score. Of course, this is but one way to rerank the websites found by two search engines. It is contemplated that there are other combining and reranking schemes. In addition, there may be instances where the number of websites found by the search engines may be reduced or increased from the top 1000 websites.

The result is a listing of the top websites that include the search terms sought. Geographical locations are then associated with the top websites. In one embodiment, all of the top websites have a geographical location associated with them. In another embodiment, less that all of the top websites have a geographical location associated with them. For example, in a particular application, the top 1000 websites for a Bing Search may include websites not found by a top 1000 websites found in a Yahoo Search. When combined and reranked, the number of websites will be over 1000. The geographical association may only be applied to 500 websites if a user feels the resulting map will be adequate to reveal the information that needs to be conveyed. Of course, there may be websites whose geographical location can not be determined. These would be unmappable to a geographical location on a map.

The locational information needed for associating a geographical locations with a ranked website can be obtained in many ways, as detailed in the specific examples that follow. Two of the ways include searching the website for locational information. For example, many websites have a tab for contact information. The contact information can include a physical address which can be associated with the website. The contact information can also include a telephone number which may be used to determine a physical area. It is recognized that telephone information may not be as accurate as many times the area code for a phone number may be in one physical location and the website could be in another. This is the similar to when an individual has a cell phone with an area code in Tempe, AZ and actually lives in Sioux Falls, SD which has a totally different area code. Another method of determining the geographical location of a website is by using the Internet Protocol (IP) address. Most IP addresses now have a geographical location associated with them. This information is available via another API, in one embodiment.

The websites found are then mapped to their geographical location on a map 2116 (See ***Figure 21***). In addition, the ranking value associated with the websites is represented on the map. In one embodiment, the ranking value of the websites is represented for the website using a Kernel point density function. In one embodiment, the ranking value represents the

height of a mountain depicted on the resulting map. So, if the top website is located in a city located between Los Angeles, CA and Newport Beach, CA and the top ranking is 1000, the city will include a mountain that is 1000 units high. Similarly, the locations of other websites will be mapped to geographical locations with a mountain corresponding to their ranking value. This is a very simple way to visualize one embodiment of this invention. Of course, the Kernel point density function may modify the above and vary the height based on a statistical distribution of the results or based upon other factors.

The resulting map is reviewed and analyzed to yield information therefrom. The map can be created at a first time and at a second time using the same or substantially the same set of keywords or search terms. The first map is then compared to the second map to reveal temporal changes or changes over time. A map that includes representations of rankings of searches of websites can also be compared to a base map. The base map is a map produced by a search of one or more standardized words. In one embodiment, the base map is a kernel density map of 50,000 web pages from 300 random keywords. This kernel density base map is subtracted from a kernel density map to remove "noise" from the map produced by the search of one or more keywords of the selected concept. In this way, the background noise associated with web pages is essentially removed. What remains is a measure of how much over the noise the selected map formed by the above methods is. If the searched and mapped concept is amongst the noise, removing the noise from the searched and mapped concept represented by the key word search will result in a kernel density map showing very little, if anything at all.

Figure 23 is a flow chart of a method 2300 of gathering and displaying data from websites, according to an example embodiment. The method 2300 includes the comparison of two similar key word searches or two similar ideas. The key words are generally related. For example, the key words "Obama" and "Romney" are related in that they are both presidential candidates. Other key words could be products related to a same sector of the economy. Related words are part of an ontology or heuristic ontology. This method 2300 includes searching for web sites with at least one web search device for websites containing a first set or at least one or more key words 2310, and searching for web sites with at least one web search device for websites containing a second set or at least one or more key words 2312. Of course, the first set of one or more key words is related to the second set of at least one or more key words. The relationship has to be more than that they were both searched using this

method. There has to be a relationship between the words to make the comparison meaningful. The method 2300 also includes ranking the websites found using the first set of at least one or more key words based in part upon the number of occurences of the at least one or more key words in a website 2314, and ranking the websites found using the second set of at least one or more key words based in part upon the number of occurences of the at least one or more key words in a website 2316. The method 2300 also includes associating geographical locations with at least some of the ranked websites 2318 and determining a differential value between a first keyword search and a second keyword search in a geographic area 2320. The differential values in the rankings are then mapped 2322. Mapping representations 2322 of the differential value of the first keyword search and the second keyword search is done for at least one geographical location on the map. In one embodiment, the differential value between the first keyword and the second keyword is reflected in the representation of the website using a Kernel point density function. In some embodiments of this method, searching for web pages includes searching web pages for at least one or more key words using a first search engine and a second search engine. In such an embodiment, ranking the websites further includes ranking the websites found by a first search engine, ranking the websites found by a second search engine, and combining the websites found by the first search engine and the second search engine and reranking the combined list of websites found by the two websites.

Specific Example Methods of Mapping Ideas in Cyberspace

With the high popularity of web search engines and social networks, many research projects have focused on the spatial analysis and visualization of web information and keyword searches, such as Google Flu Trend (Varian and Choi 2009; Ginsberg et al. 2009), the Healthmap project (Brownstein et al. 2008), and BioCaster (Collier et al. 2010). In contrast to our SWARM 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). Various research projects have pursued data mining projects investigating word co-occurrence (e.g., Ohsawa et al. 2002), centrality (e.g., Corman et al. 2002), and sentiment analysis (e.g., Chute 2008; Li and Wu 2010; 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 (e.g., Monge and Contractor 1998), social networks (e.g., Kempe, Kleinberg, and Tardos 2003; Papacharissi 2009; Cupples 2010; Perez et al. 2010; Singh, Gao, and Jain 2010), emails (e.g., Matsumura and Sasaki 2007), and websites (e.g., Elmer 2006). In particular, projects such as those focused on algorithms and structural topographical configurations and calculations (e.g., Sen and Davulcu 2010; Shekhar and Oliver 2010) suggest that unique patterns may provide unique geospatial network "fingerprints" that characterize the evolution of different social dynamic processes (e.g., Worboys 2010; Zook 2010). Several scholars have suggested there may be narrative markers of health-based (Little, Jordens, and Sayers 2003) and hate-based or terrorist groups (e.g., Leets and Bowers 1999; Hoffman 2005; Brown 2009). Such markers may be discernible through various data-mining techniques. Web-related projects to date, 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 geographic information retrieval 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. The 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).

There are many research projects focusing on mapping cyberspace and the visualization of web content. Most projects, however, use multi-dimensional coordinates or abstract distances to create visualization maps rather than adopting realspace coordinates in the real world. Compared to previous mapping cyberspace research projects, SWARMS prototype includes at least four unique features listed below. It should be noted that other unique features also exist.

1. SWARMS utilizes powerful commercial Web search engine APIs (such as Yahoo BOSS APIs and Bing APIs) rather than developing our own 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. SWARMS 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 multi-dimensional 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 advance 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. SWARMS maps 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).

Geolocation Methods

Recently, the World Wide Web Consortium (W3C) developed a standardized specification for Geolocation APIs (W3C 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, WiFi and Bluetooth MAC addresses, and GSM/CDMA cell IDs, as well as user input*" (W3C 2010). So far, however, few web pages have adopted W3C geolocation APIs. We hope that more web application and web server administrators will adopted the W3C specification in the future.

SWARMS 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 (gazatteers, geographic names, and spatial reasoning).

The first method, IP geolocation (Muir and van 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 (LBS), target marketing, epidemiology, and criminal investigation (Choi and Tekinay 2003; Lee 2008; Tsou and Kim 2010).

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

Methods	**Accuracy / Spatial Resolution**	**Spatial Data Availability**	**Implementation Requirement**	**Privacy Concerns**
IP geolocation	U.S.: Zip code level. (92119) or city-level. (10km-100km). International: City-level (Taipei, New York, etc.) or country-level	All web pages have IP addresses. Web Page IP Geocoding usually has 90 percent successful rate.	WHOIS databases (public), Commercial or free IP geolocation databases or web services (IPPAGE.org, MaxMind.com)	Medium (many IP geolocation databases are public)
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 percent 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	High (attached to individual users– mobile phones)
Geographic context analysis (Gazetteers, geographic names)	Various spatial footprint resolutions ranging from 1km to 1000km, to 3000km	10 percent web pages have U.S. zip code information; 20 percent have geographic identifiers. (Himmelstein 2005).	Challenging. Require a comprehensive framework with geographic name ontology, gazetteers, spatial reasoning tools.	Low (all gazetteers databases and ontologies are public)

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 complicated 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, zip codes, and real world coordinates (Poese et al. 2011). For example, the database can convert the IP address, 130.191.118.3 to the U.S. 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 percent to 73 percent to place an IP location within 40km 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 research 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 provide spatial information for over 90 percent of web pages. The privacy concerns are medium because most IP geolocation databases are publically available.

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. Only 1 percent of public social media content (such as Tweets) contains GPS or Wi-Fi coordinates. Most smart phone users do not turn on the GPS-geolocation functions in their social media applications (Twitter or Facebook). Privacy is another major concern for this method. The geolocation data collected by smart phones are highly personal and should be protected by laws, because they can be used to identify individuals or specific human behaviors.

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 the SWARMS prototype. First, the spatial resolution of this method varies ranging from 500 meters (such as Sea World in San Diego) to 800KM (the State of California). Some web pages may not include geographic identifiers in their contents (only 20 percent of web pages have geo-name identifiers) (Himmelstein 2005). This method cannot efficiently handle large amounts of records. On the other hand, privacy concerns for this method are low because most geographic names are public knowledge.

After comparing three different geolocation methods, IP geolocation was adopted, in this particular embodiment, as the major SWARMS prototype geolocation method. The SWARMS prototype performed geolocation procedures for approximately 90 percent 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 percent 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 percent) in geocoding tests. Even with an assumption that the IP geolocation accuracy is low (60 or 70 percent of web pages have accurate locations), the methods can still successfully detect web page clusters and the increasing/decreasing web page density with point kernel density methods.

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. SWARMS 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 usually are 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.

Most web search engines rely on web crawlers (or web robots) to collect and index web page content into a centralized database. For example, Google collects millions of web page indexes in its search engine databases daily by deploying thousands of web crawlers from their server farms. Web crawlers are dynamic network programs designed for collecting and duplicating targeted website contents (remotely) into web index databases. Each web crawler can switch its targeted websites by examining the hyperlinks found in the original web pages, often HyperText Markup Language (HTML) documents. Therefore, the crawler can perform very comprehensive web page indexing tasks for web search engines (Brin and Page 1998). After the creation of web page index databases, the next step is to decide the ranking of web pages based on specific keywords. Different search engines have adopted different ranking algorithms and methods. For example, to determine the importance of web pages, Google developed its famous PageRank method, "*a global ranking of all web pages, regardless of their content, based solely on their location in the Web graph structure*" (Page et al. 1999, 15). PageRank relies on the external referred pages (other pages linked to the targeted web page) to calculate the ranks. For example, the SDSU web page will be more important if it was referred by two important web pages (the California State University System web page and CNN.com). Another web page will be ranked lower than the SDSU web page if it were referred by two less important web pages. The referring structure of web pages will determine their ranking numbers in the Google search engine.

Currently, the Google search engine combines PageRank with other content-based analysis methods to make the keyword webpage search in Google more accurate and more effective. However, one major limitation of Google search engine is the restriction of its application programming interfaces (APIs). Current Google search APIs can only be used to retrieve up

to 64 web pages from the Google search engine each time. Therefore, SWARMS uses the 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 is similar to Google's. It 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 one embodiment of a SWARM prototype. However, after comparing the top 1000 search web pages between Yahoo and Bing, the Yahoo search engine was selected 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. At the end of this specification, a comparison between the Yahoo API search results and Bing API search results is made. It should be noted that any search engine could be used in the invention and that other search engines are contemplated. Each of the search engines could be used as a platform for a SWARMS device.

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 (DNS) tree structures or IP addresses rather than real world latitudes and longitudes. Without the linkage to real places, it is difficult to performance 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 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 U.K. 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 density landscapes showing different types of websites (commercial sites, government organizations, and non-profit 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 literature includes 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 seems 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 Spatial Web Automatic Reasoning and Mapping System (SWARMS).

We designed and implemented the Spatial Web Automatic Reasoning and Mapping System (SWARMS) prototype for creating visual maps and web information landscapes. Figure 1 illustrates the overall conceptual framework. Initial searches are conducted by using pre-defined keywords on specific topics (e.g., infectious diseases or radical concepts) provided by domain experts to search from 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 10), we convert the geocoded databases (created by a Microsoft SQL server) to *[Visualization Maps]* showing the information 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 visualization 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 language-independent 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 4 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 64 web pages from Google due to the limitation of Google APIs).

Sometimes the system might not be able to return 1000 web pages due to index problems 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. In following tests, it became clear that using the same keywords with the same search engine on different dates may return different numbers

of web pages due to the dynamic updates of web index databases. Nevertheless, most returns include over 950 records in Yahoo API (version 1). One important function of the SWARM prototype is the capability of multi-temporal search and comparison for the same keywords. Each keyword search result table includes both the keyword and the search date. We may be able to use this information and analysis to visualize and study the dynamic spread of radical concepts among different days, weeks, or months.

To demonstrate our method, we first used the keyword search of "Jerry Sanders" on 9 March 2011 with Yahoo API and generated 978 web pages with ranks (Figure 2). Since “Jerry Sanders” is the name of San Diego mayor in California. We can use the test results to verify if the spatial pattern associated with the spatial context of keywords. We converted 978 search results from the SWARMS prototype into geocoded web information databases and visualization maps using the coordinates associated with each web page (Figure 3). The geocoding of these 978 web pages utilized Simple Object Access Protocol (SOAP) and IP Address Lookup Service from the IPPage.com (http://www.ippages.com/lookups/). The IPPage Lookup Service was limited to 5000 records per day for free. 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 percent 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 and the density of points might not be easily recognizable due 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.

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. Calculating kernel density (available in ArcGIS Toolbox) can be done 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.

Initially, different spatial output resolutions and generalization thresholds (radius) based on analysis needs are produced. We then performed the kernel density function in ArcGIS, specifying a 3 map unit threshold (radius) and 0.5 map unit output scale (we will discuss radius choice later). Map unit is defined by the data frame used in a GIS software (ArcGIS 10). In this example, one map unit represents one decimal degree in the map -- approximately 80 km (50 miles) in California. The red dots indicate the locations of websites associated with the keyword (Jerry Sanders). In this design, the ranking numbers of search results were considered as the "popularity" or the "population" in the kernel density algorithm. A higher ranked website is usually more "popular" and more "visible" comparing to lower ranked websites. Therefore, we converted the ranking numbers into the population parameter:

population = (Total number of web pages + 1) - rank# (Equation 1)

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.

With the consideration of web search ranking numbers, the web information landscape can provide more meaningful information for our analysis. 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 search results cover the whole world, most of web pages in such a query are located in the United States due to the language of keywords (in English). 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. Second, most web page locations are associated with major U.S. cities. This indicates that the density of web pages may be closely related to the size of city populations. The next sections will illustrate some prominent GIS analysis methods we developed for the further analysis of information landscapes.

Revealing Hidden Geospatial fingerprints of Web Information Landscapes

During our early tests, we found that a single web information landscape may only provide limited information for spatial analysis. Comparing multiple web information landscapes and standardizing kernel density maps can reveal important spatial patterns and "geospatial fingerprints" for selected keywords and concepts. Three types of comparison methods can be applied for the comparison analysis of web information landscapes.

1. Comparison of two maps with similar keywords, such as “Jerry Sanders” versus “Antonio Villaraigosa” (both are U.S. city mayors).
2. Comparison of one keyword map versus standardized background maps (such as population density maps, or the average web page density maps). The background maps can be created by combining multiple randomized keyword search results.
3. Comparison of the temporal changes of maps with a single keyword. For example, we can compare the “burn Koran” keyword search on 30 January 2011 versus the “burn Koran” search on 03 April 2011.

Figure 5 illustrates a differential web landscape map by comparing two information landscapes and visualizing the differences between two keyword search results: "Jerry Sanders" (the mayor of San Diego) versus "Antonio Villaraigosa" (the mayor of Los Angeles). The creation of differential maps involved a series of GIS analysis operations. First, we generated point kernel density maps from the two keyword search results with the same kernel threshold (3 map units) and the same output scale (0.5 map units). The next step is to calculate the differences between the two maps. A raster-based map algebra tool from ArcGIS was used with the following formula:

Differential Value = (Keyword-A / Maximum-Kernel-Value-of-Keyword-A) - (Keyword-B / Maximum-Kernel-Value-of-Keyword-B) (Equation 2)

We use the maximum kernel values (6184 in the "Jerry Sanders" database, 5540 in the "Antonio Villaraigosa" database) from each original information landscape to standardize the kernel density values. The map algebra result shows the differential popularity (density) between the web pages related to San Diego mayor and to the Los Angeles mayor (Figure 5). Using the blue-red color scheme and a Minimum-Maximum stretch, the red hotspots in the new map indicate areas where the web page density of "Jerry Sanders" was higher than the web page density of "Antonio Villaraigosa", and the blue color areas indicate that the web page density of "Antonio Villaraigosa" was higher than "Jerry Sanders". The differential map (Figure 5) clearly demonstrates the strong spatial relationship associated with the two different keywords. The web pages related to San Diego mayor (Jerry Sanders) are much more "popular" (high density) than the other mayor in the areas around San Diego. The web pages related to the Los Angeles major (Antonio Villaraigosa) is more popular in the areas around Los Angeles and Denver areas. Such data suggest the discriminant validity in the SWARMS methodology—if it were swamped with error variance, such intuitive distinctions would be unlikely to appear.

The differential information landscape map illustrates important ***geospatial fingerprints*** hidden in the text-based web search results depending on the context of selected keywords. In this article, geospatial fingerprints are defined as *the unique spatial patterns (e.g., clusters) of web information landscapes associated with different keywords or concepts.* In our demonstration, the contexts of "Jerry Sanders" and "Antonio Villaraigosa" have implicit spatial relationships with the City of San Diego and the City of Los Angeles. These implicit spatial relationships can be visualized in their geospatial fingerprints when comparing the differences between the two web information landscapes.

In Figure 5, one unusual popularity hotspot of "Antonio Villaraigosa" is located around the City of Denver. To further investigate the hidden spatial relationships between "Antonio Villaraigosa" and Denver, we used the spatial selection function in ArcGIS to select all web pages located around the City of Denver. The "Antonio Villaraigosa" search produced 58 out of 989 web pages (5.86 percent) located around the City of Denver. On the other hand, "Jerry

Sanders" search only produced 22 out of 978 (2.25 percent) web pages located around Denver. Figure 6 lists 14 highly ranked web pages of "Antonio Villaraigosa" located around the City of Denver. After reviewing these individual web pages, we determined that many web pages located in Denver were created by very conservative Republicans or anti-illegal immigration groups. These web pages created a “negative popularity" hotspot in the information landscape. These anti-illegal immigration groups strongly dislike Villaraigosa because he is one of the few big city Hispanic mayors in the United States. With this example, the diagnostic value of geospatial fingerprint comparisons displayed their potential for identifying the spatial distributions of radical social movement support groups and their web pages. This example also indicated a potential methodological issue. The web page rank numbers assigned by commercial search engines only describe the “popularity” of web pages. But the number will not tell us whether the web page is “positive popular” or “negative popular”. To address this, our research team is currently pursuing sentiment analysis (using computational linguistic methods) to identify the “pro” web pages and “con” web pages.

Another important aspect of the creation of information landscapes is the selection of the kernel density threshold (radius). Changing threshold distances adopted in kernel density operations can result in drastically different spatial patterns and relationships at various map scales. For example, in the previous example, the differential map with a 3 map unit radius illustrates the red and blue hotspots in San Diego and Los Angeles. Figure 7 compares six different setting of radius distances (thresholds) and output grids in the differential maps between "Jerry Sanders" and "Antonio Villaraigosa". The spatial signature of the two keywords disappears in the differential map using 6 map units for the radius distance in Figure 9A. Figures 9B and 9C clearly identify the clusters of San Diego and Los Angeles. Figure 9C, with the 2 map unit threshold, illustrates the "boundary line" between the two keywords. Figure 9D, with a 1 map unit radius, identifies the San Diego and the Los Angeles clusters but misses the boundary line between the two keywords. When we reduce the kernel radius to 0.5 (Figure 9E) and 0.1 (Figure 9F) map units, the identified hot spots become much smaller than the county boundaries in California.

We suggest the following settings of kernel density thresholds for detecting geospatial fingerprints at different map scales.

- 6 - 8 map units for detecting the State level geospatial fingerprints.
- 2-3 map units for detecting the County level geospatial fingerprints.

- 0.5 - 1 map units for detecting the City level geospatial fingerprints.
- 0.1 - 0.2 map units for detecting the Zip Code level geospatial fingerprints.

The spatial scale dependency observed in Figure 7 reflects the nature of geospatial fingerprints and the spatial characteristics of web information landscapes. Similar to other concerns in spatial analysis, such as autocorrelation and the modifiable areal unit problem (MAUP), map scale plays a significant role in the visualization and detection of spatial patterns and spatial relationships for web information landscapes.

Searching for Hotspots of Radical Concepts and Linking Information Landscapes to Real World Census Data

In September 2010, an obscure preacher's intention to burn the Koran spread like wildfire in various media throughout much of the world. This singular announcement by a solitary person touched off violent protests that took many lives and threatened further escalation of tensions and rifts between the West and the Islamic world. Following this event, we used "burn Koran" as our keyword to search for the top 1000 web pages from Yahoo's search engine and analyzed the spatial distribution of the keyword and associated web pages. The keyword search was conducted on 30 January 2011 (four months into the debate over the selected event). The four month delay of this keyword search resulted from technical problems with our SWARMS prototype in 2010. Figure 8A (top) illustrates the information landscape of the "burn Koran" keyword search results (1000 web pages) with the setting of 3.0 map units for the kernel density threshold and 0.5 map units for the output grid resolution.

In this example, we used another method to detect geospatial fingerprints by comparing the "burn Koran" information landscape with the real population density in the United States (as a baseline map) rather than comparing to other keywords. Figure 8B (bottom) illustrates the kernel density map of city population based on 3,149 U.S. cities (small black dots) with a weighted population value (radius: 3.0 map units, output grid: 0.5 map units). The U.S. cities dataset was provided by ESRI (its Data and Map product). A reasonable assumption is that the population of a city is correlated with the number of web servers located around the city. Bigger cities will have more websites hosted by their residents. Comparing Figure 8A and 8B, there are similar density patterns around the East Coast (New York, Philadelphia, and Baltimore) and the West (California). Nevertheless, there are differences between the two maps. To clarify the pattern variations, we again applied the map-algebra function to

calculate the differences between the two density maps. Figure 9 illustrates the differential map between the "burn Koran" popularity and the real world population density—that is, the map controls or adjusts for population density.

The U.S. population density map was used to standardize the popularity density map of “burn Koran”. After the standardization, the red color hot spots indicate that San Jose, Houston, and the middle of Kansas State have higher web page densities associated with the "burn Koran" keyword. The blue color hot spots indicate the negative value (lower density) of "burn Koran" web pages standardized by city population density.

One interesting finding in Figure 9 is the unusually high density of “burn Koran” web pages in the middle of Kansas. In fact, after the original event happened in the church located in Gainesville, Florida (green symbol), another church in the city of Topeka, Kansas claimed that they would continue the action of “burn Koran”. The spatial change of radical event centers may be reflected in our differential information landscape. The red hotspots can be found around the city of Topeka, Kansas rather than around the city of Gainesville, Florida.

There are some drawbacks to using the U.S. population density map as our standardization map. The web page density is usually associated with the density of websites and web servers rather than the actual population. For example the web page density in San Jose will be much higher than other places due to the cluster of IT and web service companies in Silicon Valley. Therefore, we created another standardization map by using 300 randomly chosen keywords. Each set was used as keywords to search in Yahoo and Bing APIs. Some examples of random keywords are “'most', 'As', 'possible', 'himself', 'Sue', '.', 'young', 'so', '61', 'sort', 'the', 'so', 'B', 'too', 'age'. Since many search engines ignore some stop words, like 'the,' 'a,' and 'for', we removed over 130 stop words from the list and then combined the rest of random keywords into 56 sets (three words per set). 56,000 web pages were created and combined as the “average background” of Yahoo search engine results (Figure 10).

Figure 11 illustrates the differential map between the “burn Koran” and the Yahoo background web density maps. The geospatial fingerprint in Figure 11 is still similar to Figure 9, but the web page density in San Jose in Figure 11 is not higher compared to the

background map. Therefore, we think the Yahoo background map (using random keywords) is a better standardization map than the city population density map.

On 1 April 2011, fourteen people were killed in Afghanistan due to the controversial burn Koran incident in Florida. We used the SWARM prototype to compare the web information landscapes between 30 January 2011 and 3 April 2011 to demonstrate the dynamic temporal changes of web information landscapes. Figure 12 illustrates the differences between web page densities on the two dates. Red indicates increased web page density of "burn Koran" in April 2011 compared to the web page density of "burn Koran" in January 2011. Blue highlights the decreasing web page density of "burn Koran" in April 2011. There was a significant decrease in "burn Koran" web pages in the April map around the city of Topeka, Kansas. On the other hand, Saint Louis, Pittsburgh, and Philadelphia have more web pages (increasing web page densities) related to "burn Koran" on 3 April 2011.

Sensitivity Assessment: Comparing Search Results between Bing and Yahoo Engines

Although our preliminary web information maps illustrated strong spatial patterns related to search keywords, we need to gain more knowledge about commercial search engines and analyze their keyword search results. To do so, we first compared the search results between Bing search engine API and Yahoo search engine API by using the same keywords. Through an internal research team discussion, we created twelve ad-hoc definitions of web page categories:

1. Blog (personal or group blogs): personal journals or small group diaries with clear authors/writers information and dates at the top of web pages. Blogs usually have very strong personal opinions.
2. Commercial: web pages created for selling products or services, or explaining information related to commercial products.
3. Educational: school, university, and educational institute web pages. Educational web pages usually come within the EDU domain in their URL.
4. Entertainment + videos: web pages provide multimedia media or on-line videos for entertainment purposes, such as YouTube.
5. Forum: websites allow a group of users to post and share their opinions and comments.

6. Governmental websites (local, state and federal governments, usually associated with .gov, .us, .org etc.).
7. Informational: Wikipedia-type web pages, such as About.com, Wikipedia.org and other similar online Yellow Info pages.
8. News: web pages include local news sites or national news, such as ABC, NBC, CNN, FOX, KUSI, etc.).
9. NGO (non-governmental organizations): web pages associated with NGOs and related activities (such as Red Cross or Rotary Club).
10. Social media: personal or group twitter sites or public Facebook Pages. User can create personal content or messages easily.
11. Special Interest Groups: webs pages are created to promote specific concepts or items - such as supporting political parties or controversial opinions).
12. Offline: web pages cannot be found or became broken links.

One researcher was selected to classify search results web pages from both Yahoo and Bing APIs by using the twelve definitions. Figure 13 illustrates an example of our classified web pages represented by different colors. For example, blue indicates the "News" category and yellow indicates "Governmental" websites. To compare different web page search engines (Yahoo API and Bing API), we used "Jerry Sanders" as the keyword to search in both APIs. The Yahoo API returned 1000 records (on 08 September 2011) and the Bing API returned 668 records (on 08 September 2011). After the classification, we found out that "Blogs" and "News" web pages are the major elements in both Yahoo and Bing search results (Figure 14A). Bing search engine, however, tends to return commercial, informational (wiki), and social media web pages. Yahoo search engine preferentially returns blogs, news, and educational web pages (Figure 14A). We also compared the top 100 search results from both engines. The Yahoo top 100 results contain more news and blogs than the Bing top 100 results, which have more information and social media. Larger differences exist between Yahoo and Bing in the top 100 rather than the top 1000 results.

We then compared the Yahoo search engine API activity between two different keywords, "Jerry Sanders" and "Texans". The Texans are a U.S. NFL team located in Houston, Texas. Figure 14B illustrates different search results from the Yahoo API (both keywords returned 1000 records on 08 September 2011). With "Texans", 40.3 percent of the top 1000 results included commercial web pages, compared to 9 percent of the "Jerry Sanders" results. On the

other hand, 31 percent of "Jerry Sanders" results were news pages, compared with 18.2 percent from "Texans" (Figure 14B).

One surprising finding in our comparison is the significant differences between individual web page URLs. When we compared the 1000 web pages from Yahoo results and the 688 pages from Bing, only 40 out of 688 pages have identical URLs (Web addresses). Most people may not realize that the top 1000 (or 688) web pages from two search engines are quite different, because most web search users only focus on the top 10 results rather than the top 1000 results. The top 10 results between Yahoo and Bing are similar (five out of ten pages are the same). Of significance however, only 5.8 percent of the web pages are identical in their top 688 records (Figure 15). In the "Texans" search result comparison (Yahoo versus Bing), only 36 web page URLs are identical between the 1000 Yahoo records and 780 Bing records.

Although the two search engines returned very different web pages using the same keyword, the web page density differential maps we created (compared to the Yahoo background density map, Figure 10) still display strong spatial patterns for the keywords in both the Yahoo results and Bing results. Figure 16A shows the comparison between the Yahoo background versus Yahoo search results for "Jerry Sanders" (top). Figure 16B shows the Yahoo background versus Bing search results for "Jerry Sanders" (bottom). Ideally, we should use Bing background versus the Bing search results for "Jerry Sanders". But due to the instability of Bing API recently, we have not created the Bing background map yet. The regions in red indicate more web page density than average; blue indicates less web pages/density than average. The web density maps created by Yahoo API (top one) show a hotspot for "Jerry Sanders" in San Diego, California. The density map created by Bing API (bottom one) also shows a hotspot for both Los Angeles and San Diego even though only 5.8 percent of the Bing web page records are identical to the Yahoo API results (Figure 16). We did the same comparison maps with the keyword "Texans" and got similar results. Both Yahoo and Bing API maps show a higher density of web pages related to "Texans" in major Texas cities, including Houston and San Antonio. The Bing map show a higher concentration of web page density in Houston than the Yahoo map.

In addition to the map keyword comparison, we also did a temporal change comparison to reveal the dynamics of search results between different days, weeks, and months. For

example, with the keyword "burn Koran", 962 out of 1000 search results on 03 April 2011 are identical compared with the 1000 search results on 04 April 2011. Only 38 web pages are new on 04 April (Day 1 comparison) and 136 web pages have exactly the same rank in both dates. After 30 days, only 662 out of 1000 records returned on 03 May 2011 are identical to the search results from 03 April 2011 (10 web pages have the same rank in both dates). Figure 17A illustrates the dynamics of search results on different dates. In addition to compare the identical URLs, we also compared temporal changes using IP addresses and host names (Figure 17A).

In the course of our analysis, we also observed that different keywords may have markedly different temporal change rates. We tested the search of “Osama Bin Laden” on 04 May 2011, two days after he was killed in Pakistan. Figure 17B illustrates the dramatic change of "Osama Bin Laden" search results on Day 1 (625 URLs out of 1000 are different between 04 May 2011 and 05 May 2011) and Day 2 (793 URLs are different between 04 May 2011 and 06 May 2011) compared to the slower change rate for “burn Koran” on Day 1 (38 URLs out of 1000 are different between 03 April 2011 and 04 April 2011) and Day 2 (67 URLs are different between 03 April 2011 and 05 April 2011).

Our sensitivity test results indicate that our methods may be effectively applied to different web search engines and can provide useful spatial information for selected keywords. Even though the search and ranking algorithms in Yahoo API and Bing API differ strongly (with only 5.8 percent overlap in URLs), our methods can still detect the change of web page density in the differential maps. For temporal search comparisons, different keywords will have varying temporal change rates depending on the contexts of selected keywords.

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. In our early tests, we only used English in our keyword search. Different languages may create significantly different web information landscapes. Figure 18 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.

FIG. 20 shows a diagrammatic representation of a computing device for a machine in the example electronic form of a computer system 2000. In various example embodiments, the machine operates as a standalone device or can be connected (e.g., networked) to other machines. In a networked deployment, the machine can operate in the capacity of a server or a client machine in a server-client network environment, or as a peer machine in a peer-to-peer (or distributed) network environment. The machine can be a personal computer (PC), a tablet PC, a set-top box (STB), a Personal Digital Assistant (PDA), a cellular telephone, a portable music player (e.g., a portable hard drive audio device such as an Moving Picture Experts Group Audio Layer 3 (MP3) player, a web appliance, a network router, a switch, a bridge, or any machine capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine. Further, while only a single machine is illustrated, the term "machine" shall also be taken to include any collection of machines that individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

The example computer system 2000 includes a processor or multiple processors 2002 (e.g., a central processing unit (CPU), a graphics processing unit (GPU), arithmetic logic unit or all), and a main memory 2004 and a static memory 2006, which communicate with each other via a bus 2008. The computer system 2000 can further include a video display unit 2010 (e.g., a liquid crystal displays (LCD) or a cathode ray tube (CRT)). The computer system 2000 also includes an alphanumeric input device 2012 (e.g., a keyboard), a cursor control device 2014 (e.g., a mouse), a disk drive unit 2016, a signal generation device 2018 (e.g., a speaker) and a network interface device 2020. The data storage apparatus 300 is also attached to the bus 2008.

The disk drive unit 2016 includes a computer-readable medium 2022 on which is stored one or more sets of instructions and data structures (e.g., instructions 2024) embodying or utilized by any one or more of the methodologies or functions described herein. The instructions 2024 can also reside, completely or at least partially, within the main memory 2004 and/or within the processors 2002 during execution thereof by the computer system 2000. The main memory 2004 and the processors 2002 also constitute machine-readable media.

The instructions 2024 can further be transmitted or received over a network 2026 via the network interface device 2020 utilizing any one of a number of well-known transfer protocols (e.g., Hyper Text Transfer Protocol (HTTP), CAN, Serial, or Modbus).

While the computer-readable medium 2022 is shown in an example embodiment to be a single medium, the term "computer-readable medium" should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers) that store the one or more sets of instructions and provide the instructions in a computer readable form. The term "computer-readable medium" shall also be taken to include any medium that is capable of storing, encoding, or carrying a set of instructions for execution by the machine and that causes the machine to perform any one or more of the methodologies of the present application, or that is capable of storing, encoding, or carrying data structures utilized by or associated with such a set of instructions. The term "computer-readable medium" shall accordingly be taken to include, but not be limited to, solid-state memories, optical and magnetic media, tangible forms and signals that can be read or sensed by a computer. Such media can also include, without limitation, hard disks, floppy disks, flash memory cards, digital video disks, random access memory (RAMs), read only memory (ROMs), and the like.

A machine-readable medium 2022 (See in ***Figure 20***) provides instructions that, when executed by a machine, cause the machine to perform operations including searching for web sites with at least one web search device for websites containing at least one or more key words, and ranking the websites based in part upon the number of occurences of the at least one or more key words in a website, associating a geographical locations with at least some of the ranked websites, and mapping representations of the ranked websites on a map at the geographical locations associated with the ranked website. The machine-readable medium also having instructions, that when executed by a machine, cause the machine to search for web pages using a first search engine and a second search engine. In some embodiments, the instructions cause the machine to rank the websites found by a first search engine, rank the websites found by a second search engine, and combine the websites found by the first search engine and the second search engine and rerank the combined list of websites found by the two websites. The instructions, when executed by a machine, cause the machine to map the ranked websites a Kernel point density function. In some embodiments, the instructions, when executed by a machine, cause the machine to compare a map formed at a first time to a

map formed at a second time. In still another embodiment, the instructions, when executed by a machine, cause the map to be compared to a standardized map of a number of randomly selected key word searches. In yet another instance, the instructions when executed on the machine compares a map formed by searching a first set of at least one or more key words to a map formed by searching a second set of at least one or more key words.

The example embodiments described herein can be implemented in an operating environment comprising computer-executable instructions (e.g., software) installed on a computer, in hardware, or in a combination of software and hardware. Modules as used herein can be hardware or hardware including circuitry to execute instructions. The computer-executable instructions can be written in a computer programming language or can be embodied in firmware logic. If written in a programming language conforming to a recognized standard, such instructions can be executed on a variety of hardware platforms and for interfaces to a variety of operating systems. Although not limited thereto, computer software programs for implementing the present method(s) can be written in any number of suitable programming languages such as, for example, Hyper text Markup Language (HTML), Dynamic HTML, Extensible Markup Language (XML), Extensible Stylesheet Language (XSL), Document Style Semantics and Specification Language (DSSSL), Cascading Style Sheets (CSS), Synchronized Multimedia Integration Language (SMIL), Wireless Markup Language (WML), Java™, Jini™, C, C++, Perl, UNIX Shell, Visual Basic or Visual Basic Script, Virtual Reality Markup Language (VRML), ColdFusion™ or other compilers, assemblers, interpreters or other computer languages or platforms. The methods discussed above can be implemented on one or more computing systems or machines. The invention also contemplates media which includes an instruction set for causing one or more processors to implement the method. Media includes physical media such as various disks, memory associated with computer systems attached to the internet or attached in other networked configurations.

The computer system 2000 shown in FIG. 20 can also be shown alternatively. FIG. 19 is a schematic diagram of another computer system 1900. The computer system 1900 will generally have most of the same components as the computer system 2000. The computer system 1900 can be thought of as interactive computer subsystems, modules or the like, that can be entirely made of computer hardware, entirely made of computer software (instruction sets), or can be a combination of computer hardware and software. The computer system

1900 includes a search subsystem 1910 for searching for web sites with at least one web search device for websites containing at least one or more key word , and a rank subsystem 1920 for ranking the websites based in part upon the number of occurences of the at least one or more key words in a website. The computer system 1900 also includes an association subsystem 1930 for associating a geographical locations with at least some of the ranked websites, and a mapping subsystem 1940 for mapping representations of the ranked websites on a map at the geographical locations associated with the ranked website. Each of these subsystems is attached to a computer bus 1950 over which computer instructions and data move between the various components coupled to the bus. Also attached to the bus 1950 is one or more processors 2002, a main memory 2004, and a static memory 2006. Various input and output devices, such as a video display 2010, are also attached to the bus 1950. The computer system 1900 also includes a network interface device 2020. When a computer executes a set of instructions, such as an instruction set found in a portion of software, the computer becomes a specialized machine. The technological advancement or technological result of the methods, hardware and software, discussed herein include producing maps representing analysis of social activities, ideas, and human communications. The mapping also allows monitoring of dynamic changes of social activities, ideas, and human communications, and the identification of geographical hot spots for various ideas, events or concepts. Some of the technological advancements are in the methods used to turn data into usable information, as well as providing a visual way to portray the usable information.

Conclusion

We present a new methodology and a multidisciplinary 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 of San Diego. The example of “Burn Koran” also demonstrated a strong linkage between the radical concept and the City of Topeka. 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. 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.

In order to better understand the deeper meanings of web information landscapes and the differential maps, we need to focus on the following questions as our research agenda:

1. How can we explain various relationships between and impacts of cyberspace activities and real-world events?
2. How does virtual space differ from, and interact with, real space? How can we reconcile differing spatial and temporal measurement units across these dimensions?
3. What are appropriate map and temporal scales for various types of cyberspace activities, and by what criteria should we select scales in order to effectively analyze spatial and temporal relationships and patterns?
4. What should a comprehensive space-time analysis framework look like for different scenarios or questions?
5. What types of ethical and civil-liberty implications (in terms of privacy, security and human rights) will we need to consider in the context of Web surveillance technologies and the analytical tools of social media?

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 confronted by great dangers 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.

The foregoing description, for purposes of explanation, used specific nomenclature to provide a thorough understanding of the invention. However, it will be apparent to one skilled in the art that the specific details are not required in order to practice the invention. Thus, the foregoing descriptions of specific embodiments of the present invention are presented for purposes of illustration and description. They are not intended to be exhaustive or to limit the invention to the precise forms disclosed. It will be apparent to one of ordinary skill in the art that many modifications and variations are possible in view of the above teachings.

The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications, to thereby enable others skilled in the art to best utilize the invention and various embodiments with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalents.

While the embodiments have been described in terms of several particular embodiments, there are alterations, permutations, and equivalents, which fall within the scope of these general concepts. It should also be noted that there are many alternative ways of implementing the methods and apparatuses of the present embodiments. It is therefore intended that the following appended claims be interpreted as including all such alterations, permutations, and equivalents as fall within the true spirit and scope of the described embodiments.

What is Claimed:

1. A method comprising:

searching for web sites with at least one web search device for websites containing at least one or more key words;

ranking the websites based in part upon the number of occurrences of the at least one or more key words in a website;

associating a geographical locations with at least some of the ranked websites; and

mapping representations of the ranked websites on a map at the geographical locations associated with the ranked website.

2. The method of claim 1 wherein searching for web pages includes searching web pages for at least one or more key words using a first search engine and a second search engine.

3. The method of claim 1 wherein

searching for web pages includes searching web pages for at least one or more key words using a first search engine and a second search engine, and wherein ranking the websites further comprises:

ranking the websites found by a first search engine;

ranking the websites found by a second search engine; and

combining the websites found by the first search engine and the second search engine and reranking the combined list of websites found by the two websites.

4. The method of claim 1wherein the ranking of the websites is reflected in the representation of the website as mapped.

5. The method of claim 4 wherein the ranking of the websites is reflected in the representation of the website as mapped using a Kernel point density function.

6. The method of claim 1 further comprising comparing a map produced by searching at least two key words to a map produced by a search of one or more standardized words.

7. The method of claim 1 further comprising comparing a first map produced in response to a search on at least one or more key words a first time to a second map produced in response to a search using a substantially similar set of key words at a second time.

8. The method of claim 1 wherein associating a geographical locations with a ranked website includes searching the website for locational information.

9. The method of claim 1 wherein associating a geographical locations with a ranked website includes using an IP address of the ranked website.

10. The method of claim 1 wherein associating a geographical locations with a ranked website includes using a geolocation API to determine locational information.

11. A method comprising:

searching for web sites with at least one web search device for websites containing a first set or at least one or more key words;

searching for web sites with at least one web search device for websites containing a second set or at least one or more key words, the first set of one or more key words related to the second set of at least one or more key words;

ranking the websites found using the first set of at least one or more key words based in part upon the number of occurrences of the at least one or more key words in a website;

ranking the websites found using the second set of at least one or more key words based in part upon the number of occurrences of the at least one or more key words in a website;

associating a geographical locations with at least some of the ranked websites;

determining a differential value between a first keyword search and a second keyword search in a geographic area; and

mapping representations of the differential value of the first keyword search and the second keyword search in at least one geographical location.

12. The method of claim 11 wherein the differential value between the first keyword and the second keyword is reflected in the representation of the website using a Kernel point density function.

13. The method of claim 11 wherein searching for web pages includes searching web pages for at least one or more key words using a first search engine and a second search engine.

14. The method of claim 13 wherein

searching for web pages includes searching web pages for at least one or more key words using a first search engine and a second search engine, and wherein ranking the websites further comprises:

ranking the websites found by a first search engine;

ranking the websites found by a second search engine; and

combining the websites found by the first search engine and the second search engine and reranking the combined list of websites found by the two websites.

15. A machine-readable medium providing instructions that, when executed by a machine, cause the machine to perform operations comprising:

searching for web sites with at least one web search device for websites containing at least one or more key words;

ranking the websites based in part upon the number of occurences of the at least one or more key words in a website;

associating a geographical locations with at least some of the ranked websites; and

mapping representations of the ranked websites on a map at the geographical locations associated with the ranked website.

16. The machine-readable medium of claim 15 wherein the instructions, when executed by a machine, cause the machine to search for web pages using a first search engine and a second search engine.

17. The machine-readable medium of claim 15 wherein the instructions, when executed by a machine, cause the machine to

search for web pages includes searching web pages for at least one or more key words using a first search engine and a second search engine, and wherein ranking the websites further comprises:

rank the websites found by a first search engine;

rank the websites found by a second search engine; and

combine the websites found by the first search engine and the second search engine and rerank the combined list of websites found by the two websites.

18. The machine-readable medium of claim 15 wherein the instructions, when executed by a machine, cause the machine to map the ranked websites a Kernel point density function.

19. The machine-readable medium of claim 15 wherein the instructions, when executed by a machine, cause the machine to compare a map formed at a first time to a map formed at a second time.

20. The machine-readable medium of claim 15 wherein the instructions, when executed by a machine, cause the machine to compare a map formed by searching at least one or more key words to a map formed by searching at least one or more standard words.

21. A computer system comprising:

a search subsystem for searching for web sites with at least one web search device for websites containing at least one or more key words;

a rank subsystem for ranking the websites based in part upon the number of occurences of the at least one or more key words in a website;

an association subsystem for associating a geographical locations with at least some of the ranked websites; and

a mapping subsystem for mapping representations of the ranked websites on a map at the geographical locations associated with the ranked website.

Figure 1

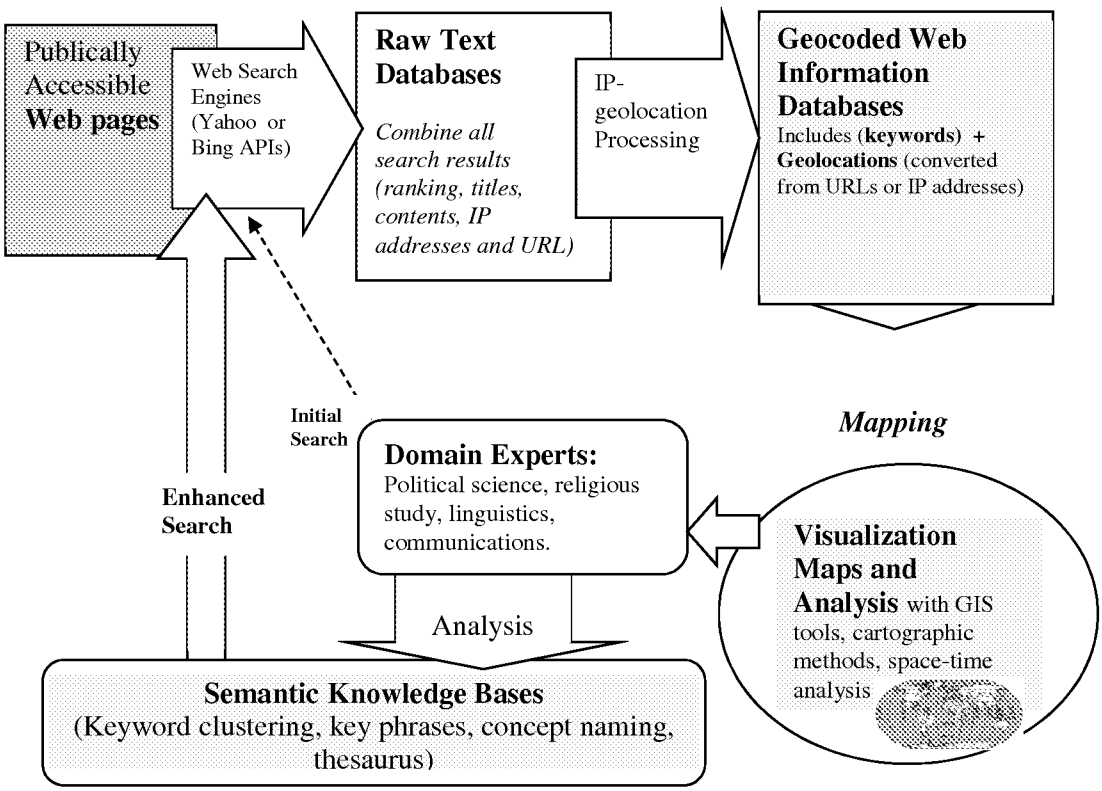
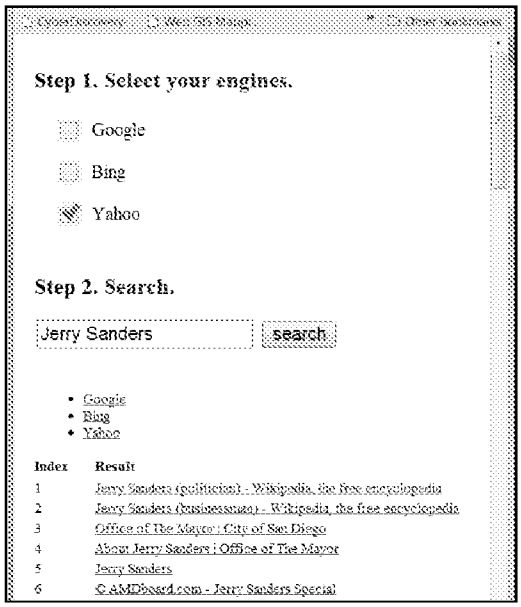

Publically Accessible **Web pages**
Web Search Engines (Yahoo or Bing APIs)
Raw Text Databases
Combine all search results (ranking, titles, contents, IP addresses and URL)
IP-geolocation Processing
Geocoded Web Information Databases
Includes (**keywords**) + **Geolocations** (converted from URLs or IP addresses)
Initial Search
Enhanced Search
Domain Experts:
Political science, religious study, linguistics, communications.
Mapping
Visualization Maps and Analysis with GIS tools, cartographic methods, space-time analysis
Analysis
Semantic Knowledge Bases
(Keyword clustering, key phrases, concept naming, thesaurus)

Figure 2


Step 1. Select your engines.
Google
Bing
Yahoo
Step 2. Search.
Jerry Sanders
search
Google
Bing
Yahoo
Index
Result
1
Jerry Sanders (politician) - Wikipedia, the free encyclopedia
2
Jerry Sanders (businessman) - Wikipedia, the free encyclopedia
3
Office of The Mayor | City of San Diego
4
About Jerry Sanders | Office of The Mayor
5
Jerry Sanders
6
© AMDboard.com - Jerry Sanders Special


Search
mappingideas.sdsu.edu/cyberdisc
CyberDiscovery ...
Web GIS Mappi...
966 Mayor Sanders & Daughter Featured in Nations
967 Jerry Brown - News - Road Runner
968 Sanders pitches alternative to Brown's redevelc
969 Talk:Jerry Sanders (businessman)
970 Comic-Con staying in San Diego | Reuters
971 Q&A: Jerry Sanders | Latest News (Republican
972 San Diego Minutemen - Mexican La Raza (The
973 MG JERRY SANDERS SCHOLARSHIP
974 Poseidon
975 Proposition 8 Trial: Day 6 Update
976 From: Jerry Sanders Date sent 06/19/2007 02:
977 Losing Money and Its President, AMD Will Tr
978 Cloak & Dagger Man Uncloaked | The Jewish V

Figure 3

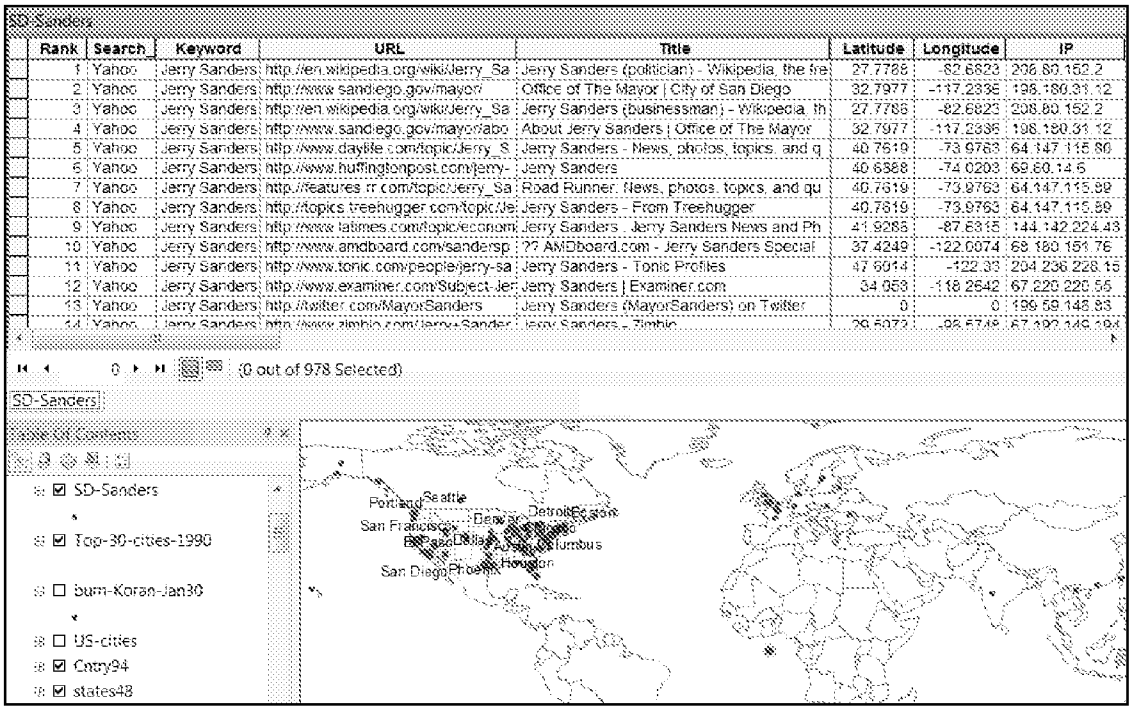


Figure 4

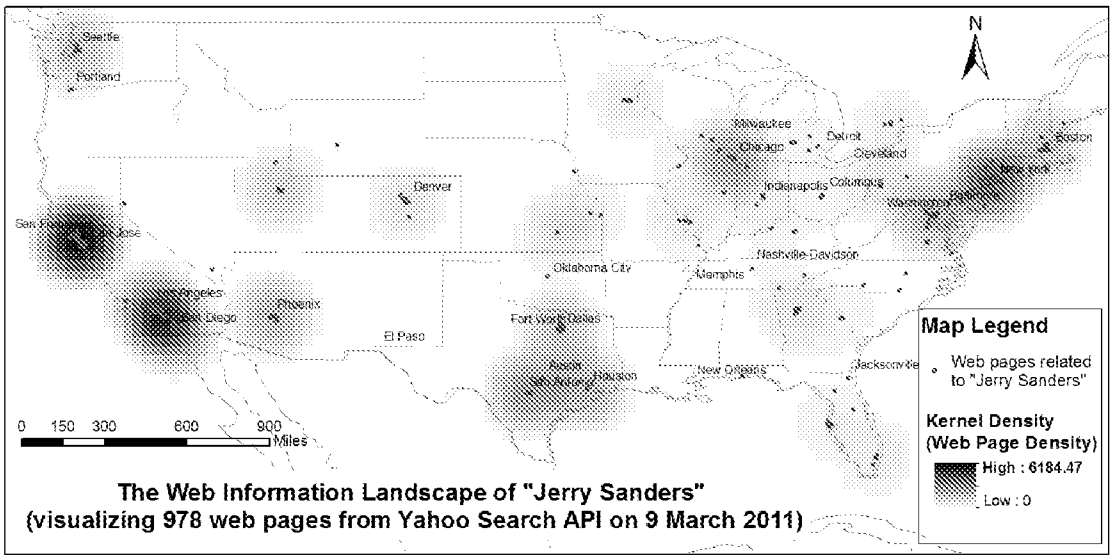


Figure 5

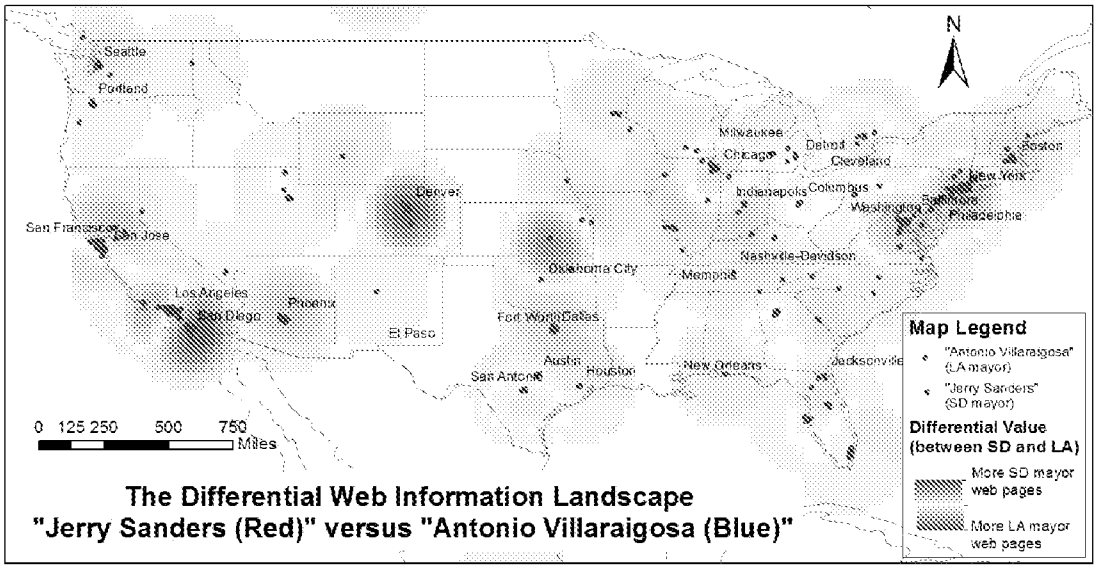

N
Seattle
Portland
San Francisco
San Jose
Los Angeles
San Diego
Phoenix
Denver
El Paso
Oklahoma City
Fort Worth
Dallas
Austin
San Antonio
Houston
Memphis
Nashville-Davidson
New Orleans
Jacksonville
Milwaukee
Chicago
Detroit
Cleveland
Indianapolis
Columbus
Washington
Philadelphia
New York
Boston
0 125 250 500 750
Miles
The Differential Web Information Landscape
"Jerry Sanders (Red)" versus "Antonio Villaraigosa (Blue)"
Map Legend
"Antonio Villaraigosa" (LA mayor)
"Jerry Sanders" (SD mayor)
Differential Value (between SD and LA)
More SD mayor web pages
More LA mayor web pages

Figure 6

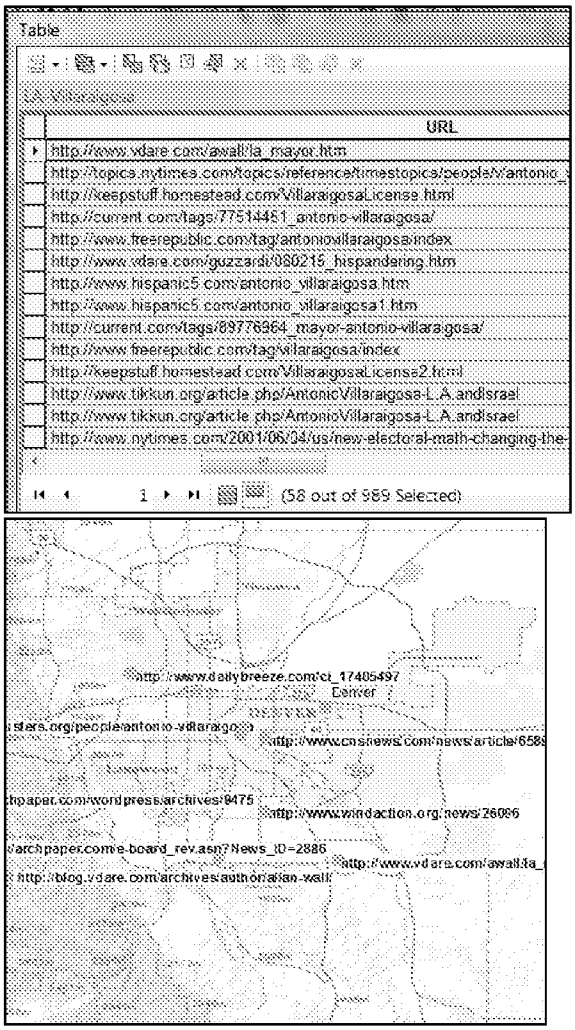
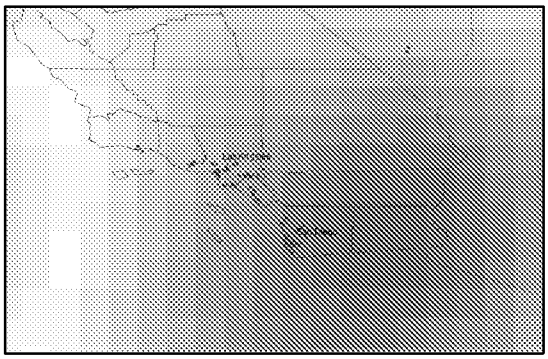
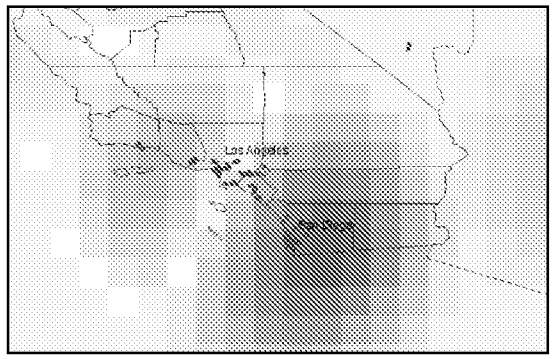


Table
URL
http://www.vdare.com/awall/la_mayor.htm
http://topics.nytimes.com/topics/reference/timestopics/people/v/antonio_
http://keepstuff.homestead.com/VillaraigosaLicense.html
http://current.com/tags/77514451_antonio-villaraigosa/
http://www.freerepublic.com/tag/antoniovillaraigosa/index
http://www.vdare.com/guzzardi/080215_hispandering.htm
http://www.hispanic5.com/antonio_villaraigosa.htm
http://www.hispanic5.com/antonio_villaraigosa1.htm
http://current.com/tags/89776364_mayor-antonio-villaraigosa/
http://www.freerepublic.com/tag/villaraigosa/index
http://keepstuff.homestead.com/VillaraigosaLicense2.html
http://www.tikkun.org/article.php/AntonioVillaraigosa-L.A.andIsrael
http://www.tikkun.org/article.php/AntonioVillaraigosa-L.A.andIsrael
http://www.nytimes.com/2001/06/24/us/new-electoral-math-changing-the-
1
(58 out of 989 Selected)
Denver
http://www.dailybreeze.com/ci_17405497
http://www.cnsnews.com/news/article/658
http://www.vdare.com/awall/la_

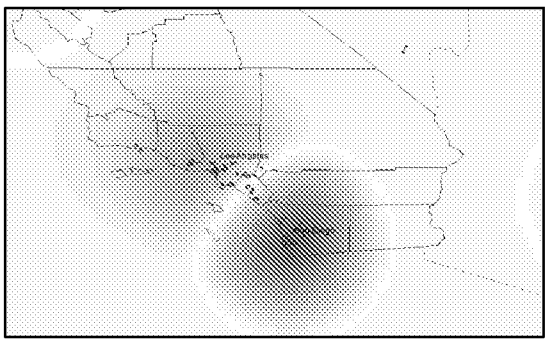
Figure 7



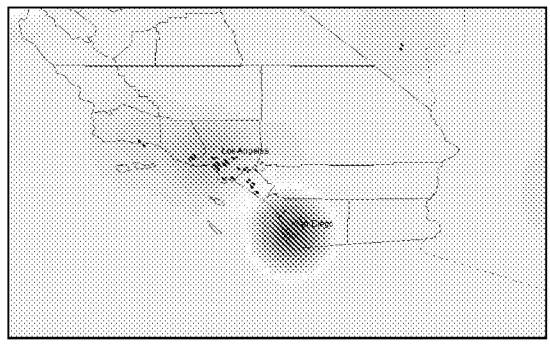
(a) Radius: 6, Output Grid: 0.5 (State level)



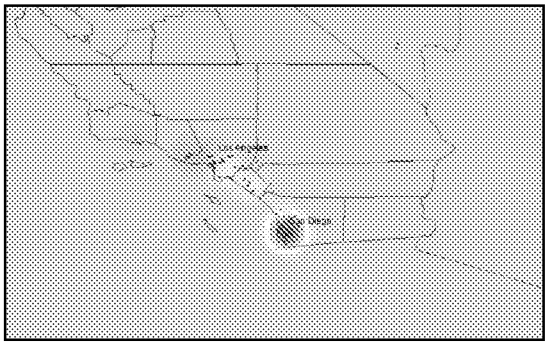
(b) Radius: 3, Output Grid: 0.5

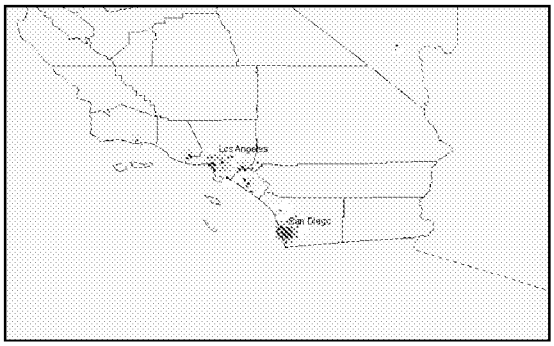


(c) Radius: 2, Output Grid: 0.1 (County level)



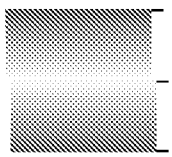
(d) Radius: 1, Output Grid: 0.1





(e) Radius: 0.5, Output Grid: 0.1 (City level)

(f) Radius: 0.2, Output Grid: 0.1

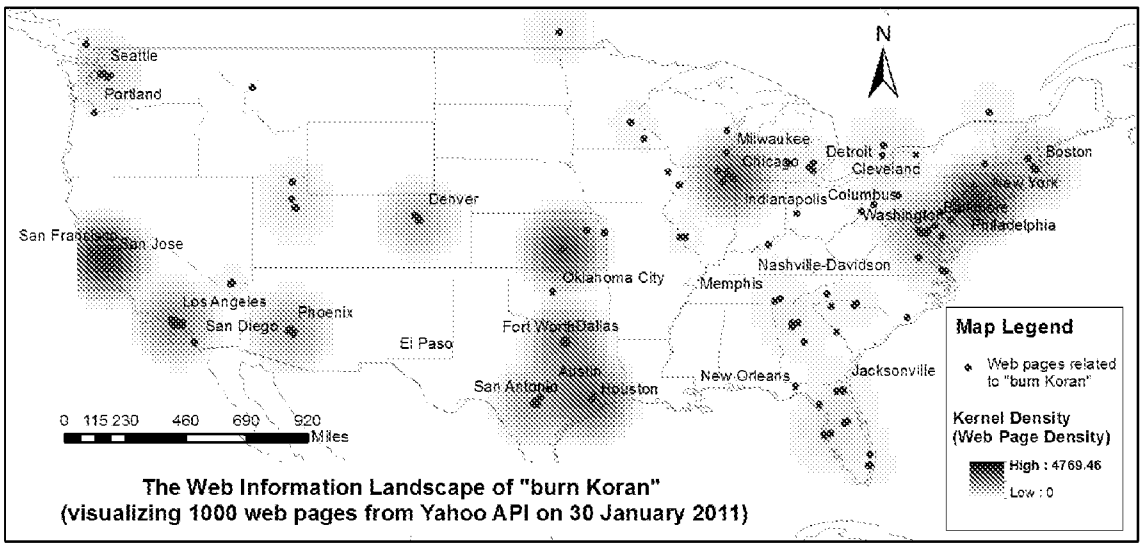


Red: The "Jerry Sanders" web page density is higher.

Blue: The "Antonio Villaraigosa" web page density is higher.

Figure 8

a).


N
Seattle
Portland
Milwaukee
Detroit
Chicago
Cleveland
Boston
New York
Indianapolis
Columbus
Washington
Philadelphia
Denver
San Francisco
San Jose
Oklahoma City
Memphis
Nashville-Davidson
Los Angeles
San Diego
Phoenix
El Paso
Fort Worth
Dallas
Austin
San Antonio
Houston
New Orleans
Jacksonville
0 115 230 460 690 920
Miles
Map Legend
Web pages related to "burn Koran"
Kernel Density (Web Page Density)
High : 4769.46
Low : 0
The Web Information Landscape of "burn Koran"
(visualizing 1000 web pages from Yahoo API on 30 January 2011)

b).

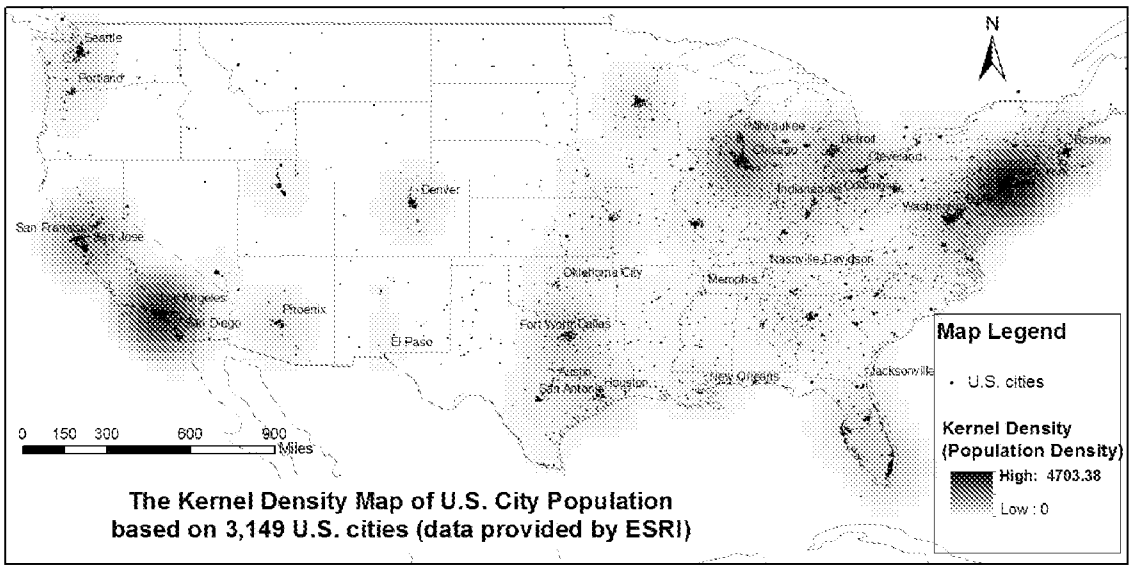
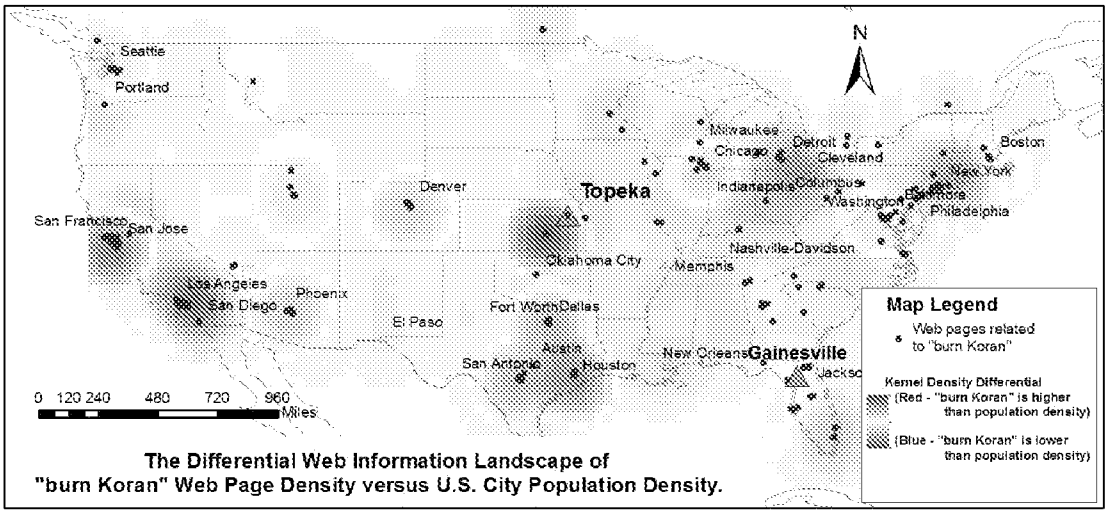
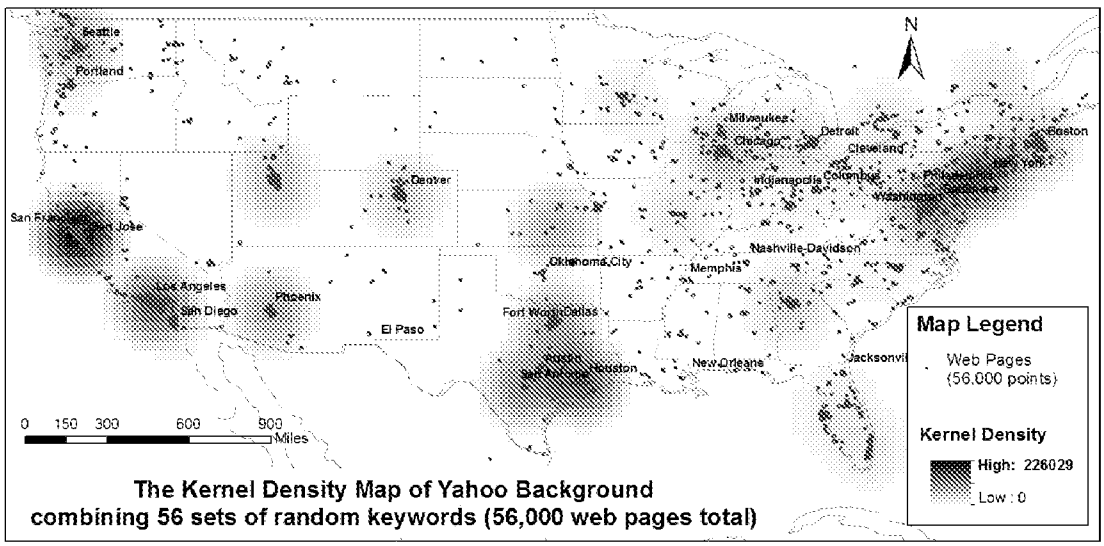

N
Seattle
Portland
Denver
San Francisco
San Jose
Oklahoma City
Memphis
Nashville-Davidson
Los Angeles
San Diego
Phoenix
El Paso
Fort Worth
Dallas
Houston
San Antonio
New Orleans
Jacksonville
0 150 300 600 900
Miles
Map Legend
U.S. cities
Kernel Density (Population Density)
High: 4703.38
Low : 0
The Kernel Density Map of U.S. City Population
based on 3,149 U.S. cities (data provided by ESRI)

Figure 9



The Differential Web Information Landscape of "burn Koran" Web Page Density versus U.S. City Population Density.

Figure 10



The Kernel Density Map of Yahoo Background combining 56 sets of random keywords (56,000 web pages total)

Figure 11

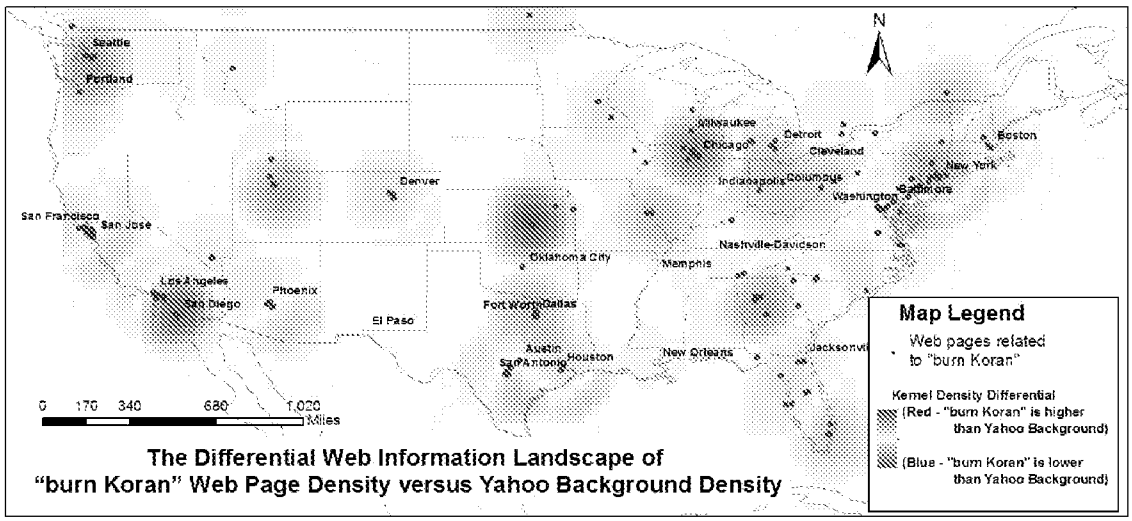


Figure 12

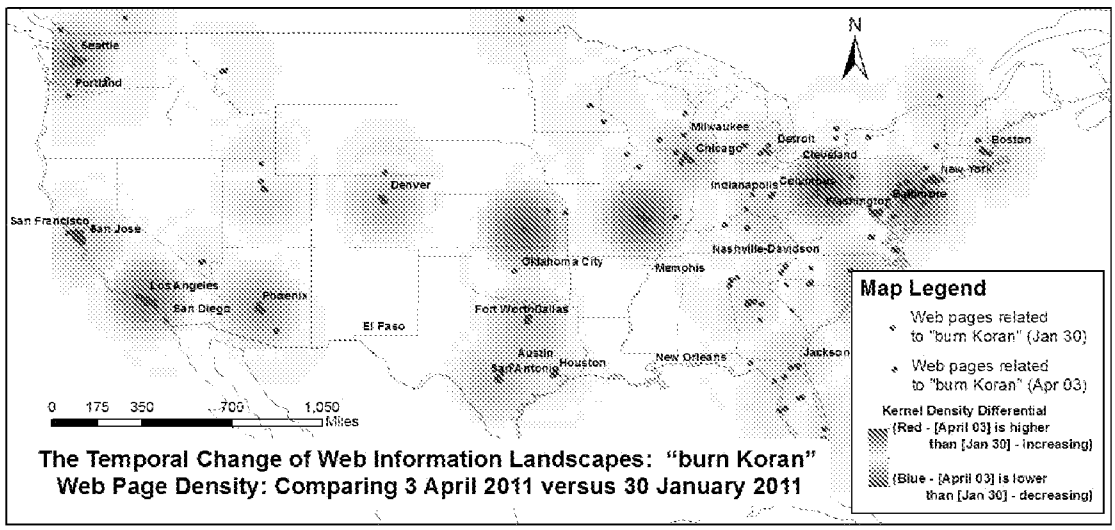
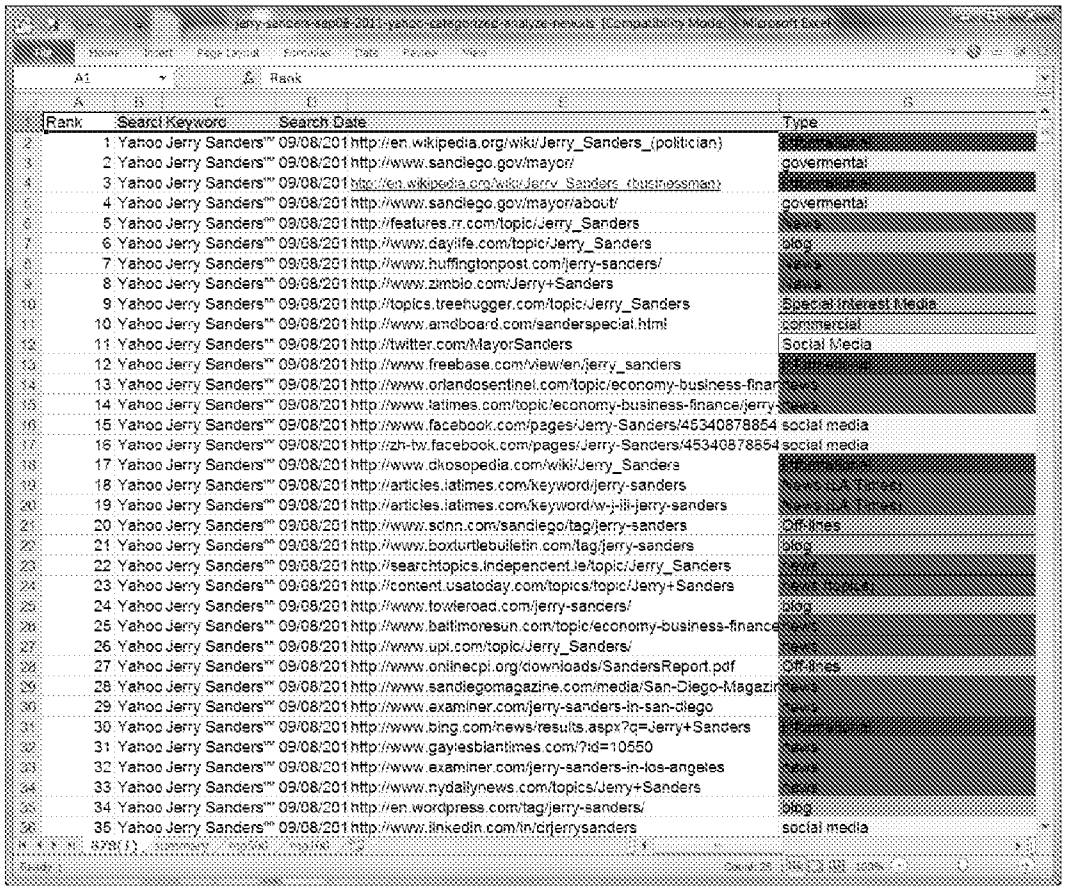


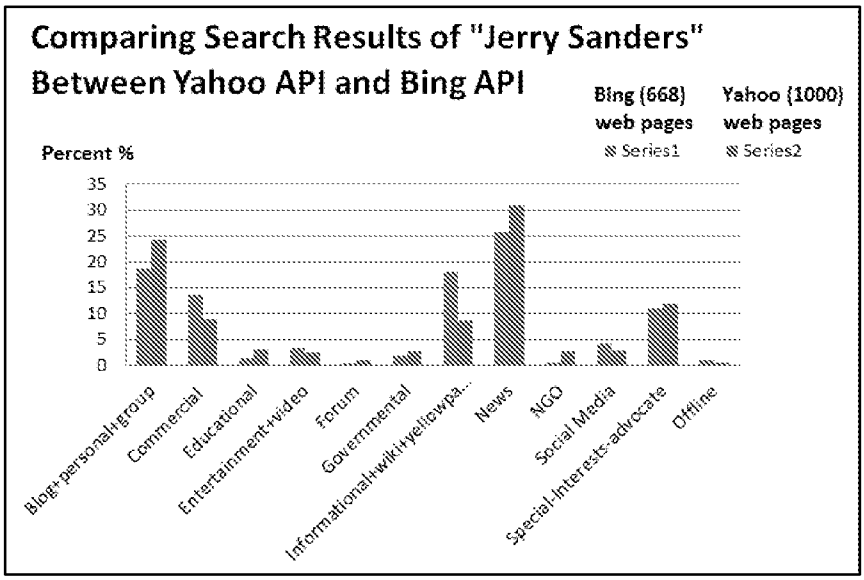
Figure 13



Rank	Search Keyword	Search Date	Type
1	Yahoo Jerry Sanders™	09/08/201	http://en.wikipedia.org/wiki/Jerry_Sanders_(politician)
2	Yahoo Jerry Sanders™	09/08/201	http://www.sandiego.gov/mayor/	governmental
3	Yahoo Jerry Sanders™	09/08/201	http://en.wikipedia.org/wiki/Jerry_Sanders_(businessman)
4	Yahoo Jerry Sanders™	09/08/201	http://www.sandiego.gov/mayor/about/	governmental
5	Yahoo Jerry Sanders™	09/08/201	http://features.rr.com/topic/Jerry_Sanders
6	Yahoo Jerry Sanders™	09/08/201	http://www.daylife.com/topic/Jerry_Sanders	blog
7	Yahoo Jerry Sanders™	09/08/201	http://www.huffingtonpost.com/jerry-sanders/
8	Yahoo Jerry Sanders™	09/08/201	http://www.zimbio.com/Jerry+Sanders
9	Yahoo Jerry Sanders™	09/08/201	http://topics.treehugger.com/topic/Jerry_Sanders	Special Interest Media
10	Yahoo Jerry Sanders™	09/08/201	http://www.amdboard.com/sanderspecial.html	commercial
11	Yahoo Jerry Sanders™	09/08/201	http://twitter.com/MayorSanders	Social Media
12	Yahoo Jerry Sanders™	09/08/201	http://www.freebase.com/view/en/jerry_sanders
13	Yahoo Jerry Sanders™	09/08/201	http://www.orlandosentinel.com/topic/economy-business-finan
14	Yahoo Jerry Sanders™	09/08/201	http://www.latimes.com/topic/economy-business-finance/jerry-
15	Yahoo Jerry Sanders™	09/08/201	http://www.facebook.com/pages/Jerry-Sanders/45340878854	social media
16	Yahoo Jerry Sanders™	09/08/201	http://zh-tw.facebook.com/pages/Jerry-Sanders/45340878854	social media
17	Yahoo Jerry Sanders™	09/08/201	http://www.dbpedia.com/wiki/Jerry_Sanders
18	Yahoo Jerry Sanders™	09/08/201	http://articles.latimes.com/keyword/jerry-sanders
19	Yahoo Jerry Sanders™	09/08/201	http://articles.latimes.com/keyword/w-j-iii-jerry-sanders
20	Yahoo Jerry Sanders™	09/08/201	http://www.sdnn.com/sandiego/tag/jerry-sanders	Off-lines
21	Yahoo Jerry Sanders™	09/08/201	http://www.boxturtlebulletin.com/tag/jerry-sanders	blog
22	Yahoo Jerry Sanders™	09/08/201	http://searchtopics.independent.ie/topic/Jerry_Sanders
23	Yahoo Jerry Sanders™	09/08/201	http://content.usatoday.com/topics/topic/Jerry+Sanders
24	Yahoo Jerry Sanders™	09/08/201	http://www.towleroad.com/jerry-sanders/	blog
25	Yahoo Jerry Sanders™	09/08/201	http://www.baltimoresun.com/topic/economy-business-finance
26	Yahoo Jerry Sanders™	09/08/201	http://www.upi.com/topic/Jerry_Sanders/
27	Yahoo Jerry Sanders™	09/08/201	http://www.onlinecpi.org/downloads/SandersReport.pdf	Off-lines
28	Yahoo Jerry Sanders™	09/08/201	http://www.sandiegomagazine.com/media/San-Diego-Magazin
29	Yahoo Jerry Sanders™	09/08/201	http://www.examiner.com/jerry-sanders-in-san-diego
30	Yahoo Jerry Sanders™	09/08/201	http://www.bing.com/news/results.aspx?q=Jerry+Sanders
31	Yahoo Jerry Sanders™	09/08/201	http://www.gaylesbiantimes.com/?id=10550
32	Yahoo Jerry Sanders™	09/08/201	http://www.examiner.com/jerry-sanders-in-los-angeles
33	Yahoo Jerry Sanders™	09/08/201	http://www.nydailynews.com/topics/Jerry+Sanders
34	Yahoo Jerry Sanders™	09/08/201	http://en.wordpress.com/tag/jerry-sanders/	blog
35	Yahoo Jerry Sanders™	09/08/201	http://www.linkedin.com/in/drjerrysanders	social media

Figure 14

a) Comparison of search results from Yahoo APIs and Bing APIs.



b) Comparison of "Texans" and "Jerry Sanders" results in Yahoo APIs.

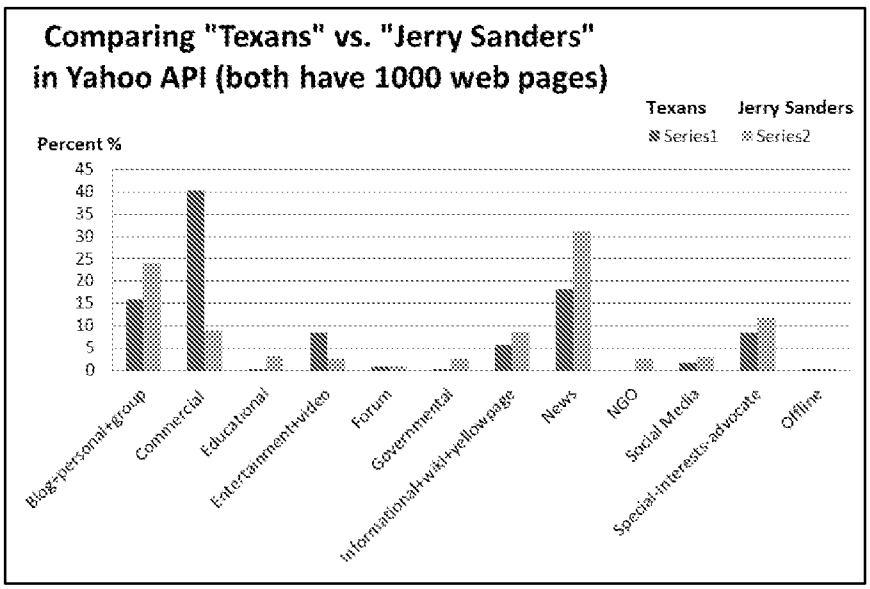


Figure 15

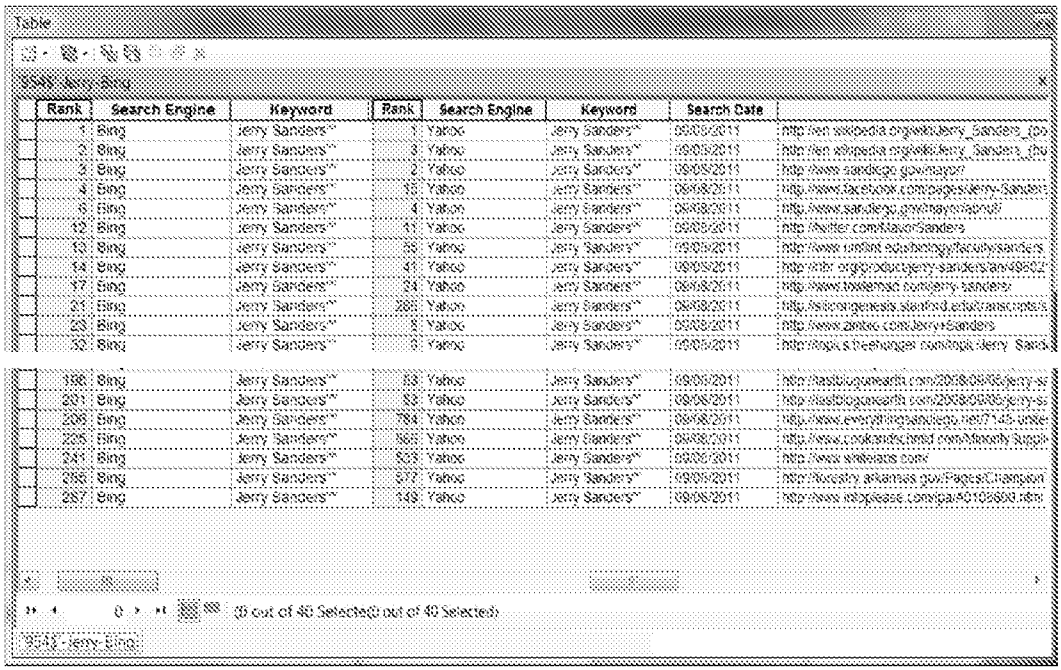
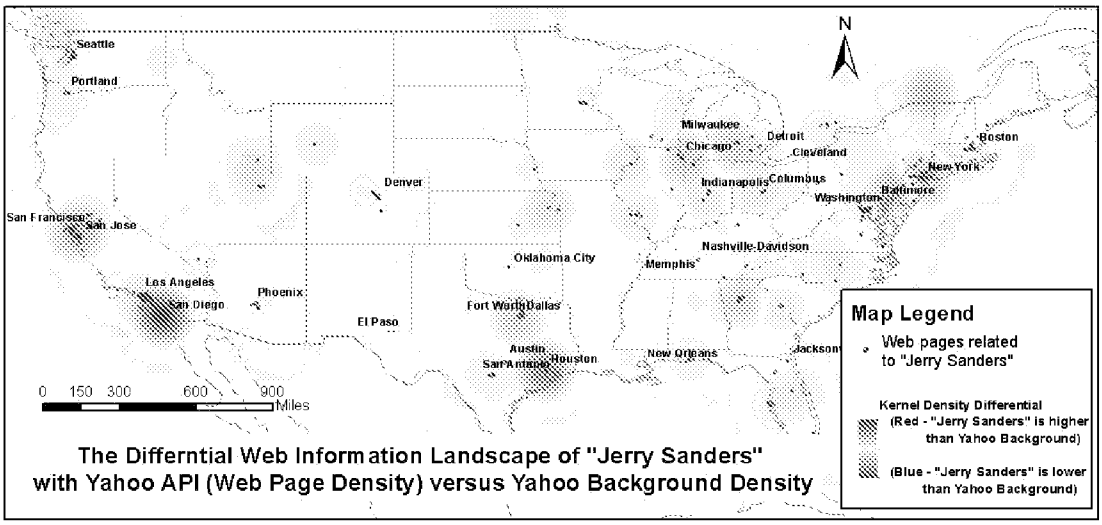


Table
Rank
Search Engine
Keyword
Search Date
Bing
Yahoo
Jerry Sanders™
09/08/2011
(0 out of 40 Selected)(0 out of 40 Selected)

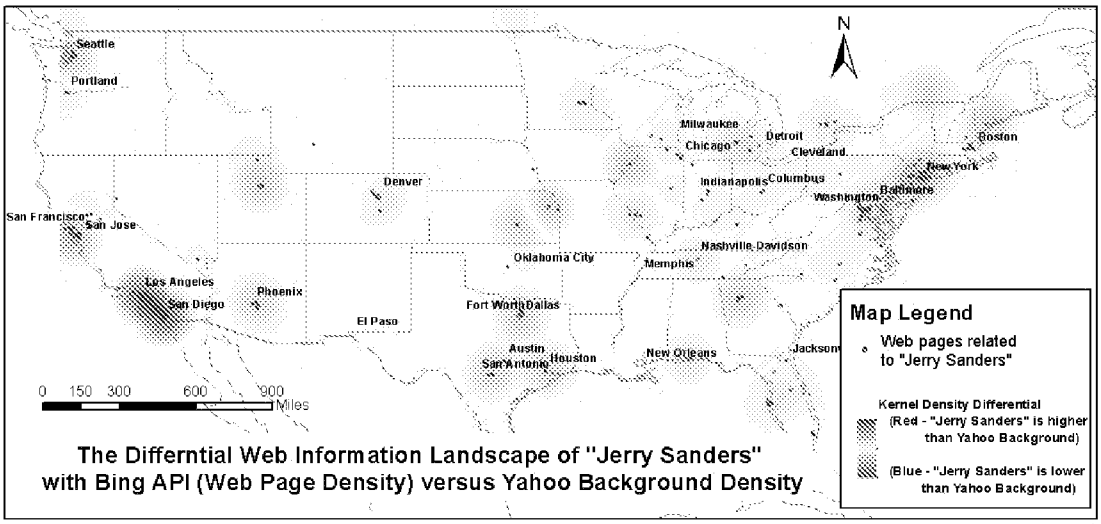
Figure 16

A).



The Differntial Web Information Landscape of "Jerry Sanders" with Yahoo API (Web Page Density) versus Yahoo Background Density

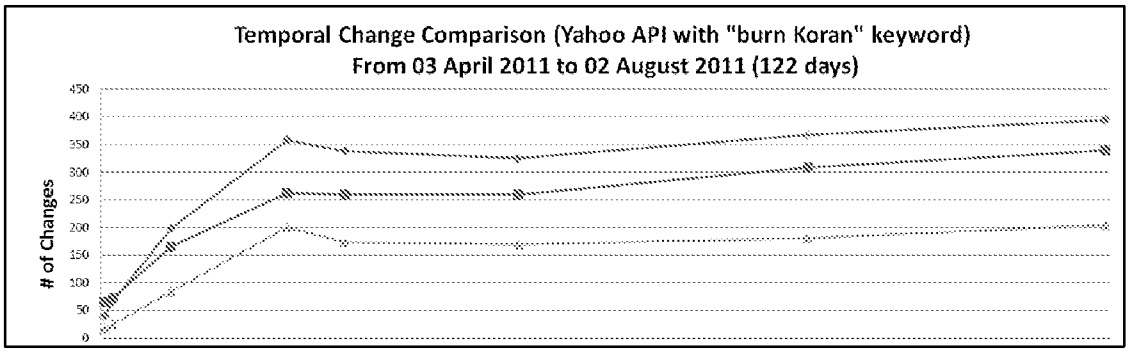
B).



The Differntial Web Information Landscape of "Jerry Sanders" with Bing API (Web Page Density) versus Yahoo Background Density

Figure 17

a).



	12	9	22	30	51	86		122
Days	1	2	9	22	30	51	86	122
URL (diff)	38	67	198	358	359	526	367	594
IP (diff)	65	73	365	262	260	268	309	339
Host (diff)	17	24	84	XXX	374	169	581	205

b).

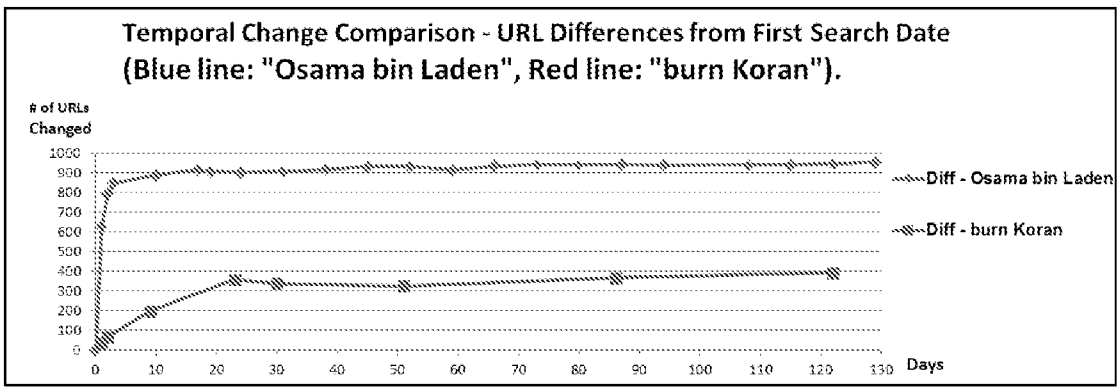
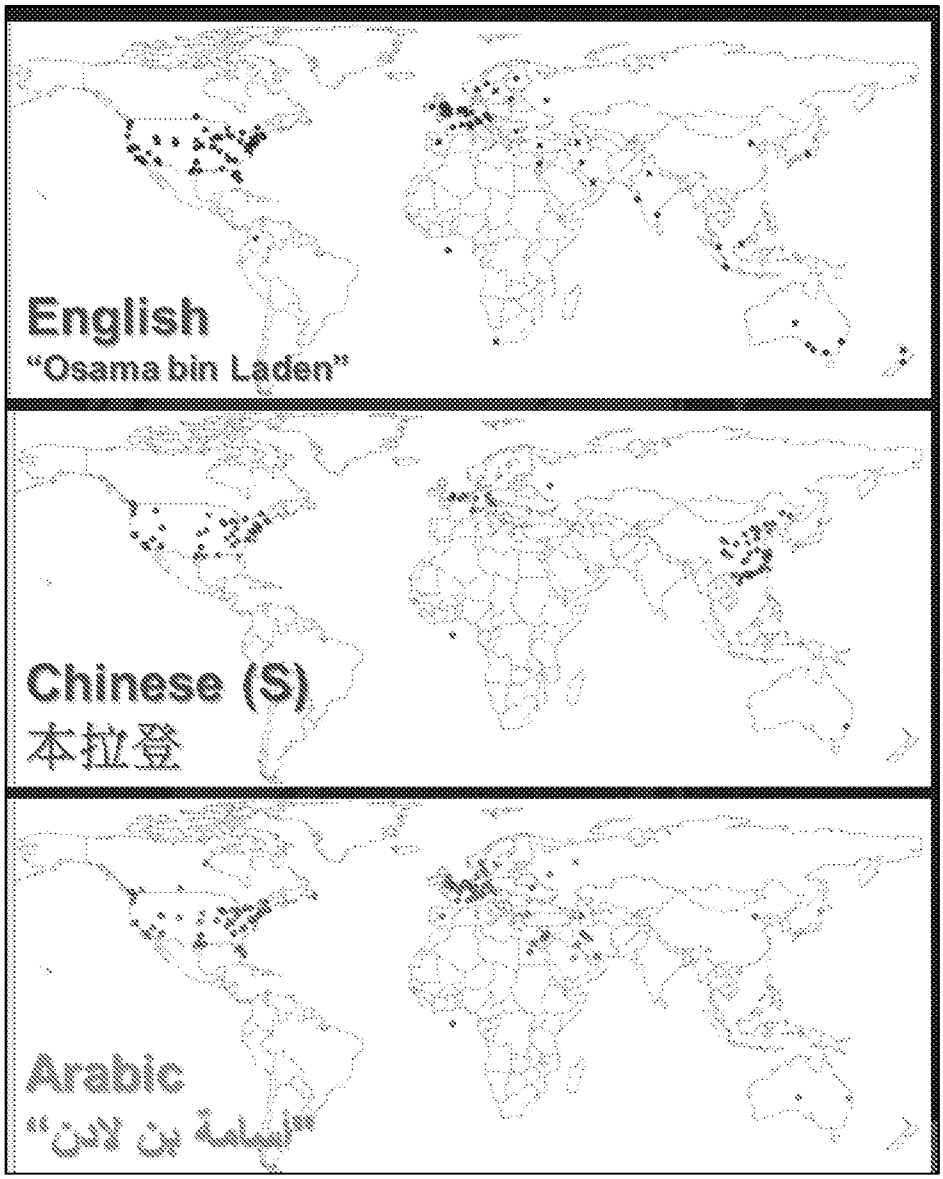


Figure 18



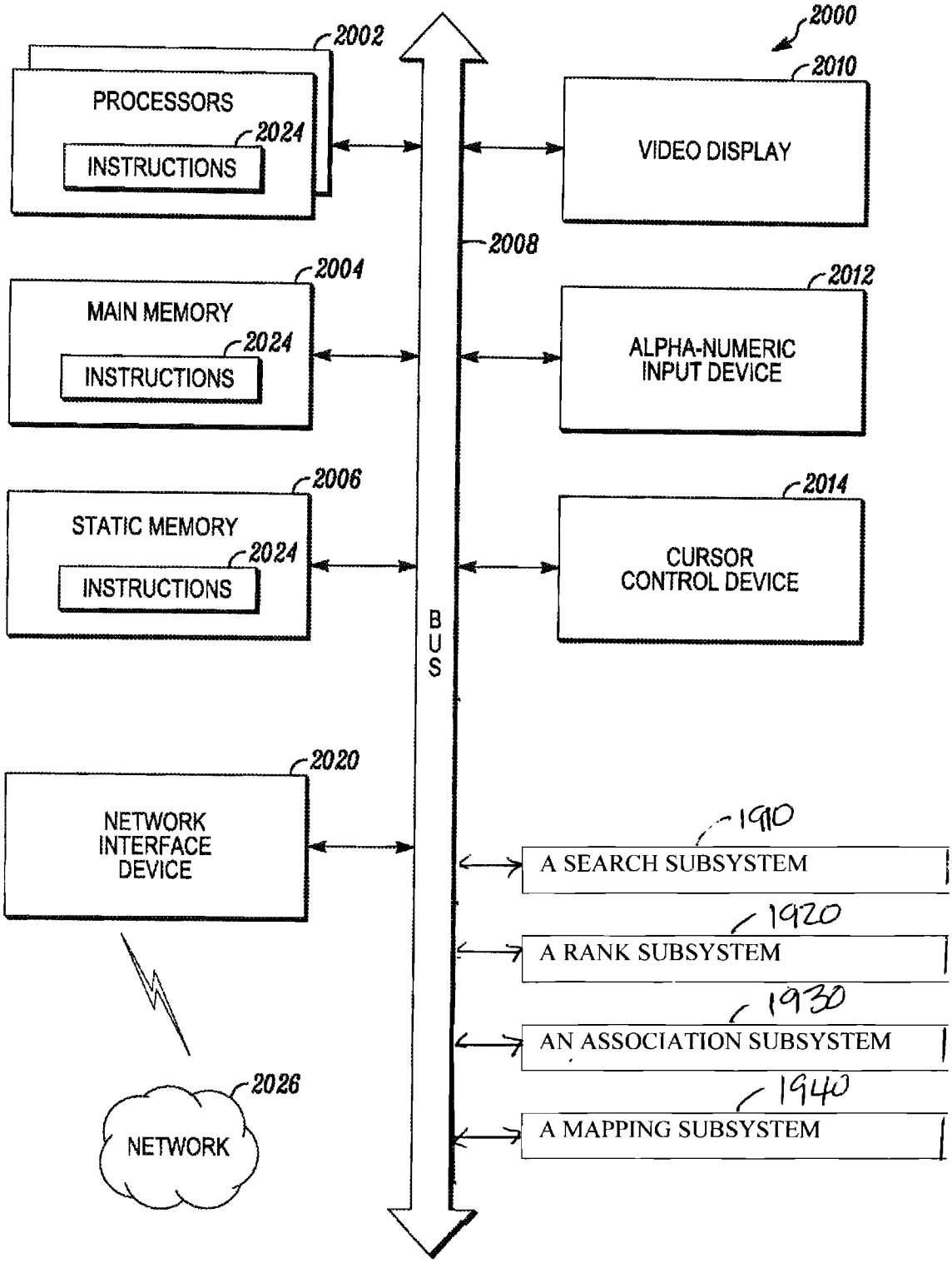

2000
2002
PROCESSORS
2024
INSTRUCTIONS
2010
VIDEO DISPLAY
2008
2004
MAIN MEMORY
2024
INSTRUCTIONS
2012
ALPHA-NUMERIC INPUT DEVICE
2006
STATIC MEMORY
2024
INSTRUCTIONS
2014
CURSOR CONTROL DEVICE
BUS
2020
NETWORK INTERFACE DEVICE
1910
A SEARCH SUBSYSTEM
1920
A RANK SUBSYSTEM
1930
AN ASSOCIATION SUBSYSTEM
1940
A MAPPING SUBSYSTEM
2026
NETWORK

FIG. 19

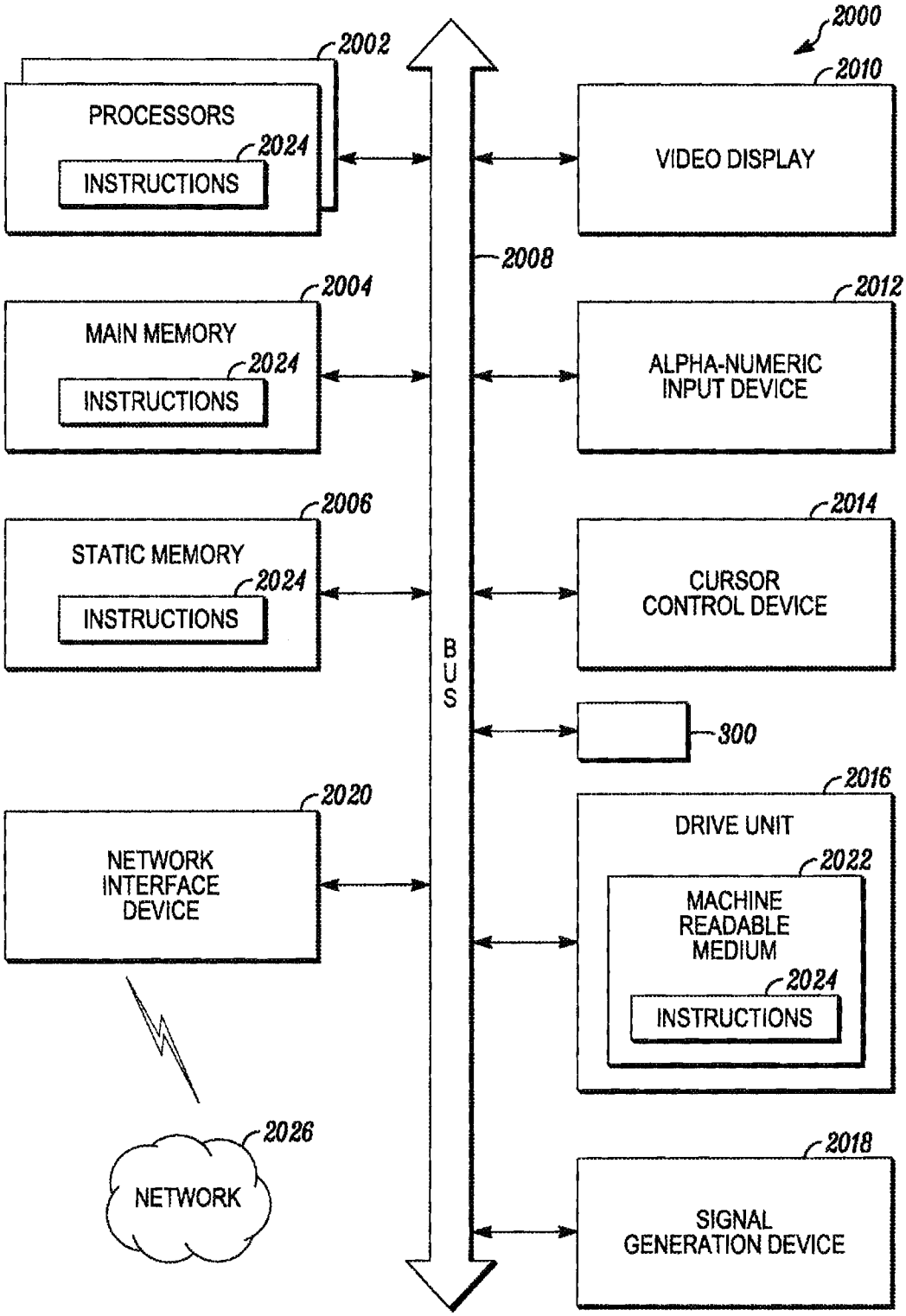

2000
2002
PROCESSORS
2024
INSTRUCTIONS
2010
VIDEO DISPLAY
2008
2004
MAIN MEMORY
2024
INSTRUCTIONS
2012
ALPHA-NUMERIC INPUT DEVICE
2006
STATIC MEMORY
2024
INSTRUCTIONS
2014
CURSOR CONTROL DEVICE
BUS
300
2016
DRIVE UNIT
2022
MACHINE READABLE MEDIUM
2024
INSTRUCTIONS
2020
NETWORK INTERFACE DEVICE
2026
NETWORK
2018
SIGNAL GENERATION DEVICE

FIG. 20

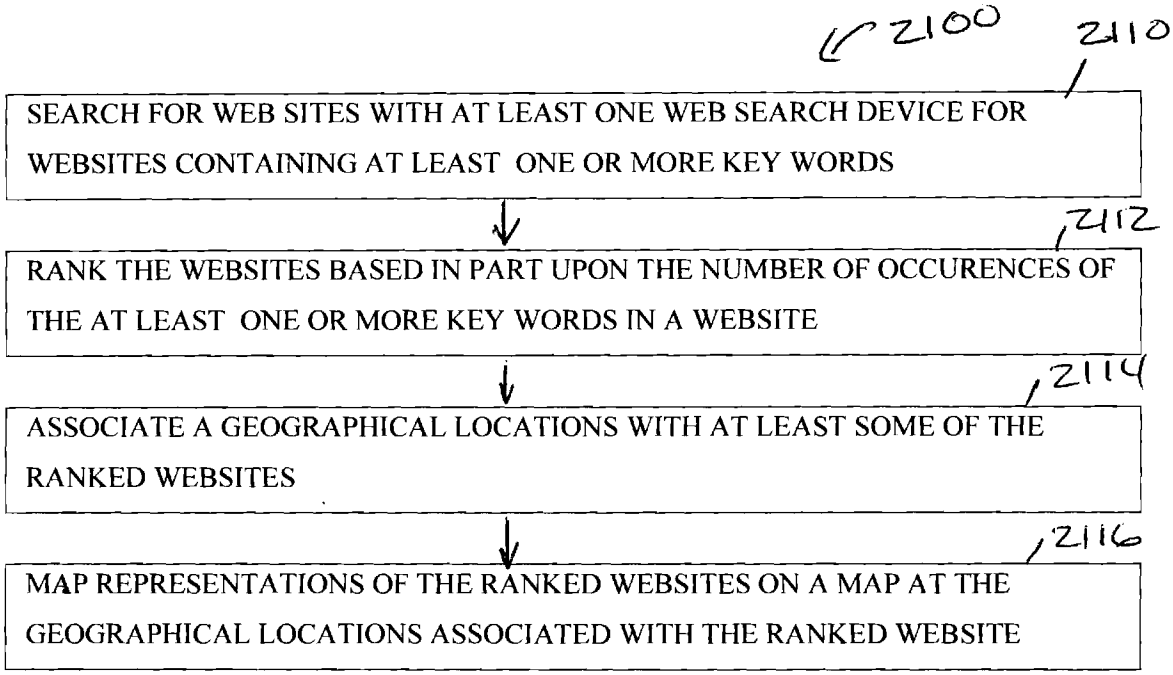

2100
2110
SEARCH FOR WEB SITES WITH AT LEAST ONE WEB SEARCH DEVICE FOR WEBSITES CONTAINING AT LEAST ONE OR MORE KEY WORDS
2112
RANK THE WEBSITES BASED IN PART UPON THE NUMBER OF OCCURENCES OF THE AT LEAST ONE OR MORE KEY WORDS IN A WEBSITE
2114
ASSOCIATE A GEOGRAPHICAL LOCATIONS WITH AT LEAST SOME OF THE RANKED WEBSITES
2116
MAP REPRESENTATIONS OF THE RANKED WEBSITES ON A MAP AT THE GEOGRAPHICAL LOCATIONS ASSOCIATED WITH THE RANKED WEBSITE

FIG. 21

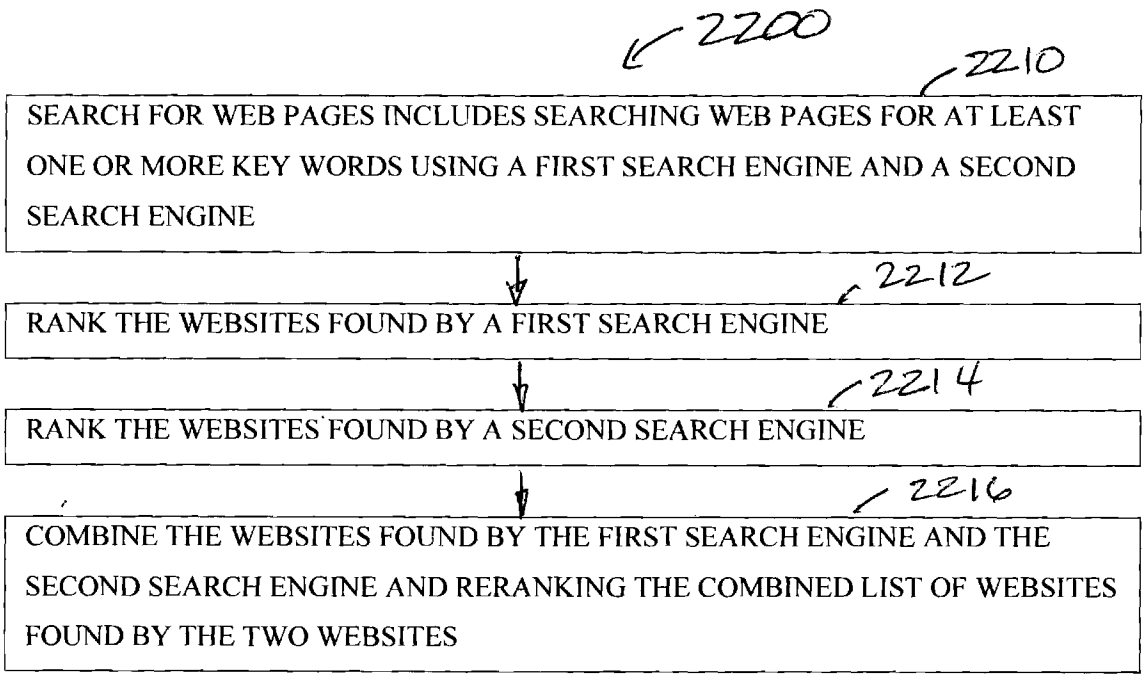

2200
2210
SEARCH FOR WEB PAGES INCLUDES SEARCHING WEB PAGES FOR AT LEAST ONE OR MORE KEY WORDS USING A FIRST SEARCH ENGINE AND A SECOND SEARCH ENGINE
2212
RANK THE WEBSITES FOUND BY A FIRST SEARCH ENGINE
2214
RANK THE WEBSITES FOUND BY A SECOND SEARCH ENGINE
2216
COMBINE THE WEBSITES FOUND BY THE FIRST SEARCH ENGINE AND THE SECOND SEARCH ENGINE AND RERANKING THE COMBINED LIST OF WEBSITES FOUND BY THE TWO WEBSITES

FIG. 22

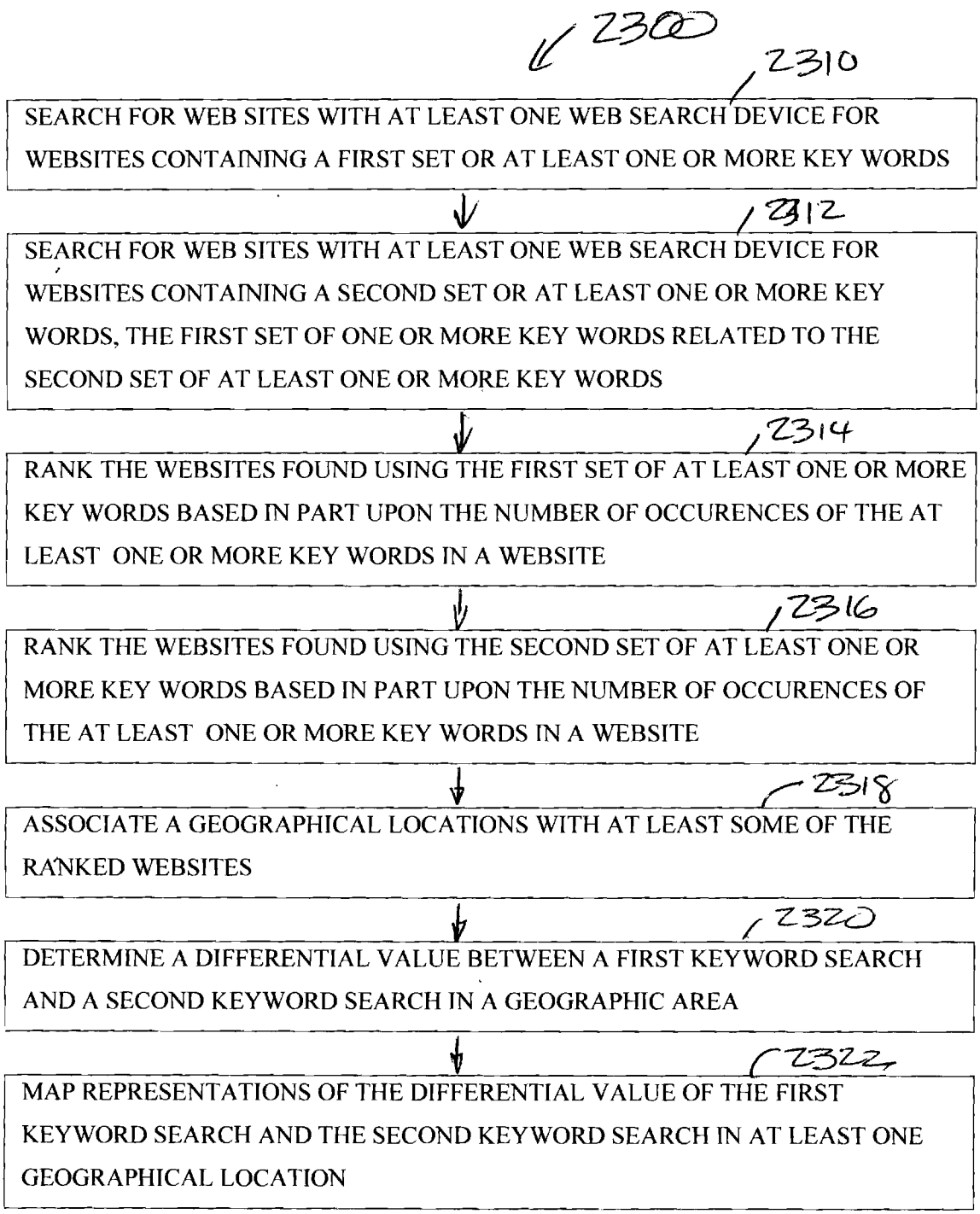

2300
2310
SEARCH FOR WEB SITES WITH AT LEAST ONE WEB SEARCH DEVICE FOR WEBSITES CONTAINING A FIRST SET OR AT LEAST ONE OR MORE KEY WORDS
2312
SEARCH FOR WEB SITES WITH AT LEAST ONE WEB SEARCH DEVICE FOR WEBSITES CONTAINING A SECOND SET OR AT LEAST ONE OR MORE KEY WORDS, THE FIRST SET OF ONE OR MORE KEY WORDS RELATED TO THE SECOND SET OF AT LEAST ONE OR MORE KEY WORDS
2314
RANK THE WEBSITES FOUND USING THE FIRST SET OF AT LEAST ONE OR MORE KEY WORDS BASED IN PART UPON THE NUMBER OF OCCURENCES OF THE AT LEAST ONE OR MORE KEY WORDS IN A WEBSITE
2316
RANK THE WEBSITES FOUND USING THE SECOND SET OF AT LEAST ONE OR MORE KEY WORDS BASED IN PART UPON THE NUMBER OF OCCURENCES OF THE AT LEAST ONE OR MORE KEY WORDS IN A WEBSITE
2318
ASSOCIATE A GEOGRAPHICAL LOCATIONS WITH AT LEAST SOME OF THE RANKED WEBSITES
2320
DETERMINE A DIFFERENTIAL VALUE BETWEEN A FIRST KEYWORD SEARCH AND A SECOND KEYWORD SEARCH IN A GEOGRAPHIC AREA
2322
MAP REPRESENTATIONS OF THE DIFFERENTIAL VALUE OF THE FIRST KEYWORD SEARCH AND THE SECOND KEYWORD SEARCH IN AT LEAST ONE GEOGRAPHICAL LOCATION

FIG. 23

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 2013/026433

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 *(2006.01)*

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 7/00, 17/00-17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatSearch (RUPTO internal), USPTO, PAJ, Esp@cenet, Information Retrieval System of FIPS (http://www.fips.ru)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TSOU M. H. Mapping cyberspace: tracking the spread of ideas on the internet. Proceedings of the 25th International Cartographic Conference held in Paris, July 2011 [online] [retrieved on 2013-04-25]. Retrieved from the Internet: <URL: http://http://icaci.org/files/documents/ICC_proceedings/ICC2011/Oral%20Presentations%20PDF/D3-Internet,%20web%20services%20and%20web%20mapping/CO-354.pdf>, p. 5	1-21
A	US 2006/0200490 A1 (ROGER OWEN ABBISS) 07.09.2006	1-21

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 April 2013 (26.04.2013)	23 May 2013 (23.05.2013)
Name and mailing address of the ISA/ FIPS Russia, 123995, Moscow, G-59, GSP-5, Berezhkovskaya nab., 30-1 Facsimile No. +7 (499) 243-33-37	Authorized officer I. Nikolaeva Telephone No. (499) 240-25-91

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 2013/026433

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
☐ No protest accompanied the payment of additional search fees.