GeoSearcher: Location-Based Ranking of Search Engine Results

Carolyn Watters and Ghada Amoudi
Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia. Canada. E-mail: watters@cs.dal.ca

Many web queries have geospatial dimensions. While online shopping is built on the premise that distance and location are irrelevant (with the possible exception of shipping charges), tourism and onsite inspection of goods have a geospatial dimension and distance and location are relevant factors. Current search engines build indices based on keyword occurrence and frequency for query negotiation using these indices. This approach is fast, robust, and generic but when queries are related to physical locations and distances rather than cyberdistances this approach leaves the user to sort through pages of results. In this paper, we describe an algorithm that assigns location coordinates dynamically to web sites based on the URL. A prototype search system was built using this algorithm that uses this information to re-rank the results of search engines for queries with a geospatial dimension. We found that over 80% of the URLs tested could be assigned correct location coordinates. This work makes a contribution to retrieval on the web by providing an alternative ranking order for search engine results so that users with queries with a geospatial dimension can more readily use the results of general search engines rather than special purpose applications.

1. Introduction

The World Wide Web (WWW or Web) is a domain of documents estimated to be currently at one billion documents (Notess, 2001) with as many as one million new pages added each day. This count does not include the "hidden web" (Mardis, 2001), documents hidden in intranets or documents created dynamically from databases for one time use. In this "visible web," the majority of searches are accomplished by using a search engine, such as: Google, Altavista, Yahoo, NorthernLight, or HotBot. Each such engine covers some portion of the available documents to generate its own indices of keyword occurrence and frequency and uses this information along with link and usage analysis to rank the results of query negotiation. Users are presented with sets of results that are ordered by general purpose-ranking schemes imposed by the search engine.

Not surprisingly, a survey for Realnames (Search engine, 2000) reports that 44% of the web users are frustrated by navigation and search engine use. Current search engines build indices based on keyword occurrence and frequency for query negotiation using these indices. This approach is fast, robust, and generic but when queries are related to physical locations and distances rather than cyberdistances this approach leaves the user to sort through pages of results. Online shopping, for example, is built on the premise that distance and location are irrelevant (with the possible exception of shipping charges) while tourism and onsite inspection of goods have a geospatial dimension and do depend on distance and location. In this paper, we evaluate a prototype that provides ranking of search engine results for queries with a geospatial dimension.

Many queries have geospatial dimensions, when physical interaction is ultimately anticipated, such as driving, touring, or shipping. Current search engines provide results to these queries but in a different context than may be most useful. For example, the query, *Maine skiing,* ranks sites containing all of the keywords highest but the remaining results are not ranked in any particularly helpful order. If, for example, the user is interested in skiing within a day's drive, then distance, irrespective of the occurrence of the term *Maine,* could be useful in the ranking the results and would include sites in Quebec, New Brunswick, and New Hampshire.

2. Background

Geospatial information is information that refers to the position of an entity on the Earth, and includes information such as street, city, and national borders as well as longitude and latitude coordinates. *Geoparsing* is the process of recognizing geographic and geospatial contexts, such as Niagara Falls or Japan, while *geocoding* is the process of assigning geographic coordinates to such contexts (Larson,

Received April 2, 2002; revised July 22, 2002; accepted August 5, 2002

1996). Geospatial querying refers to queries about spatial relationships, such as intersection, containment, adjacency, distance between entities based on geocoded information (Egenhofer, 1994; De Floriani et al., 1993). Geospatial or location-based queries involve both geometric (co-ordinates) and topological (names) features and typically fall into two general categories (Frew et al., 1995), *what is here* and *where is this*. The *what is here* query is looking for information about a given location, such as Paris, while the *where is this* query is trying to find what location(s) are relevant to a given feature, such as wine tasting. Location-based queries may, of course, have both components.

Location-based queries can, of course, often be satisfied by carefully crafted Boolean queries when the user has appropriate geographic knowledge. For example, in our earlier query on skiing, a query of the form *skiing AND (Maine OR Quebec OR New Hampshire OR New Brunswick)* would likely provide suitable results. According to Baeza-Yates and Ribeiro-Neto (1999) there are two types of problems in retrieval: problems with the data and problems with the users. Many new systems, such as GenieKnows (2001), try to improve search results by automatically expanding the query for the user based on thesauri and other vocabulary tools. Presumably, queries could be expanded to include appropriate location-based terms as well. Yokaji et al. (2001), for example, extracted addresses from the content of web pages and found a 25% increase in search hits for region-based queries when the location plus keywords were used.

Our interest lies more in the data, where the feature set for retrieval is not complete enough to satisfy the requirements of location-based information queries. These include queries about what is in a region or where something is located, i.e., queries that are not well served by traditional keyword-based queries.

A great deal of work has been done to improve the results of search engines, both by search engine companies and researchers. From early systems based on Boolean combinations of keywords, search engines now use sophisticated algorithms using metadata, keyword frequency, document length, and access frequency to rank results. Google (2002) and CLEVER (Chakrabarti, 1998) search engines, for example, exploit the link structure in combination with traditional keyword search to provide ranking to the results and to improve the quality of high ranking results. Search engines often use post search filtering to improve the results for users. AltaVista (2002), for example, filters results by language, region, date, domain, or URL.

Geographic Information Systems (GIS) have been an area of interest from the 1960's and deal with the indexing, searching and retrieval of geo-referenced information, i.e., information with associated geographic coordinates (Larson, 1996). GIS often combine databases with high-end spatial information capabilities to enable map retrieval, identification of locations or routes based on geospatial criteria (McCurley, 2001). GIS resources are available and should be applicable to helping resolve location-based queries on the web.

In addition to traditional GIS systems, researchers have built retrieval systems to exploit geographical features to improve search results. For example, Buyukkokten et al (1999) report on a prototype that they built to use geographic information to improve search engine results. Their goal was to impose a *globality* dimension to the results of a search based on the geographical distribution of web pages that linked to a given site using the Google search engine. They accomplished this using three databases; *Whois*, area codes, and zip codes. When the user enters a URL, Google returns pages in the .edu domain that have a link to that URL. These sites are given geographic locations and mapped to a digital map. Further work by Ding et al., (2000) uses geospatial information to determine the geographic scope of interest for web resources, such as online newspapers.

McCurley (2001) built a navigation tool to map web resources to a digital map so that users can select documents based on geographic clues. McCurley uses the *Whois* database, area codes, phone numbers, as well as geographic feature names. McCurley reports that only 4.5% had recognizable ZIP codes, 8.5% had a recognizable phone number, and 9.5% had at least one of these.

GIPSY (Geo-referenced Information Processing System) is a system that extracts geographical index terms from documents (Woodruff and Plaunt, 1994) using a geographic thesaurus to map vocabulary to geographic areas as well as a place name database to determine the geographical points of interest. For example, a document may refer to a dam and eagle habitats and using the vocabulary resource geographic areas of intersection could be determined. This level of sophisticated natural language processing may, however, be too complex for the on-the-fly requirements of processing search engine results.

Search engines are aware of the need for providing location-based searches. Altavista.com, for example, provides a very coarse-grained location qualifier for advanced searches that filters results by large geographic areas, like Asia and North America, or by country code. GeoTags (2002) uses longitude and latitude coordinates to rank results by distance from some reference point and to position results on a digital map. This system, however, depends on metatagging of individual web sites with the required information. Northern Light search engine (Northern Light, 2002), has a geosearch option which selects local web sites with addresses for information about professional services, reviews, local businesses, publishers and products anywhere in the US or Canada.

Sources of Geospatial Information

Although the great appeal of the Web for users is that much of the information is independent of geographic location, many web sites do contain information that is of more particular relevance to a specific city or region, such as apartments, theatres, or schools and many users have location-based information needs. If we imagined having a database that related each URL on the Internet with its geographical coordinates then we could imagine exploiting this information in web searches. For example, looking for business schools within driving distance from Halifax would return results ranked by distance from downtown Halifax or, potentially, the user's current coordinates. This result would be decidedly different from a search ranking the results on the frequency of occurrences of the terms *business*, *school*, and *Halifax*.

Handling location-based web queries depends on the successful geocoding of sources and on the Web this depends on accurate metatagging of pages or accurate geoparsing of either the web page host or the actual page content. Geocoding is generally based on longitude and latitude rather than on place names for several reasons. Place names are not unique, have spelling variations, often change over time, and may be only temporary. The use of exact coordinates facilitates the calculations needed to determine areas, distance, and shortest paths. In addition, the use of longitude and latitude values for geocoding provide a useful link to mapping systems for visualization of results on digital maps (Larson, 1996).

The metatagging of web pages to include geographic information is not considered in this paper as a viable option for location-based searching on the "visible" web because of the scale of the document base, the reliance on self-reporting, and the inherent difficulties of self-maintenance of the information. Accurate metatagging of this information on a large proportion of web sites would, of course, provide the basis for fast ranking of sites based on location.

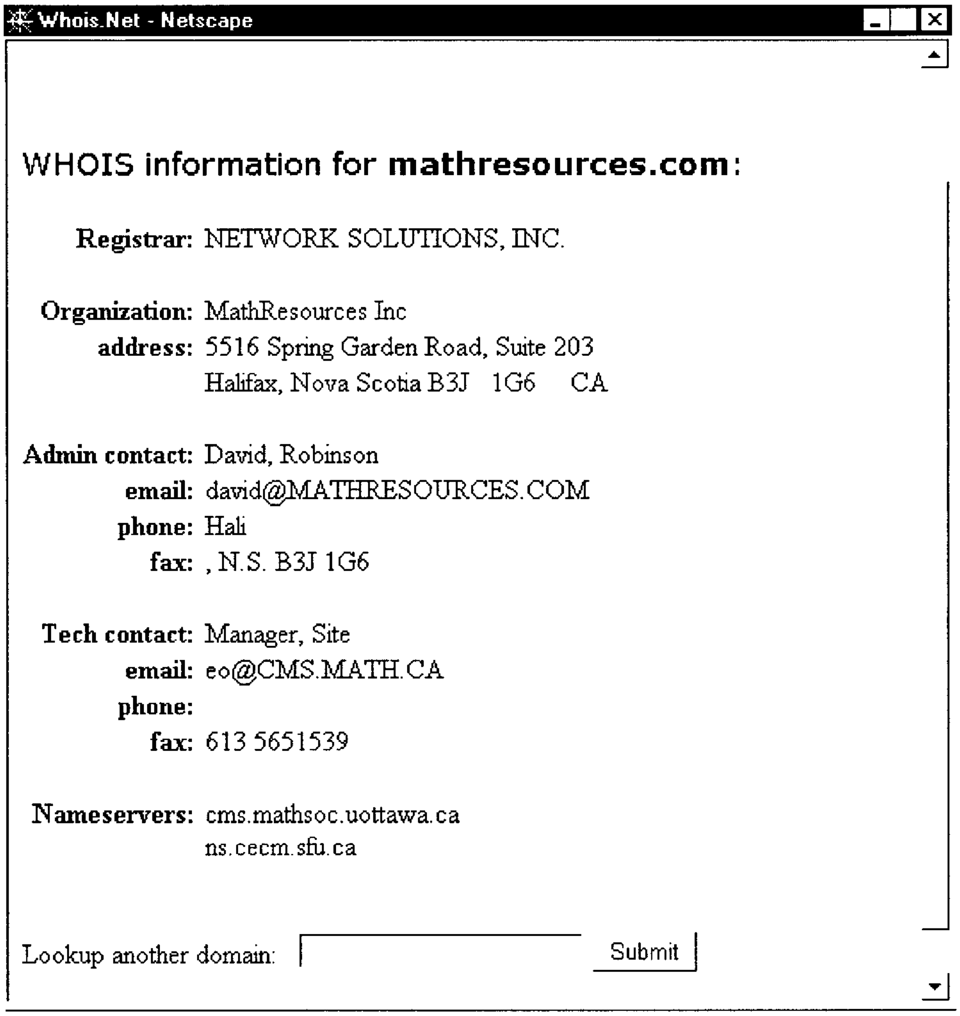
Web pages are a rich source for all sorts of geospatial information, which can be determined directly from page contents or indirectly through the URL host. Deriving this information from page content is, however, problematic. For example, phone numbers, addresses, geographic references in the text, or identifying the language used may provide clues. While this remains reasonably easy for humans, it is not yet trivial to do in a manner that is both accurate and efficient enough to be accomplished on the fly. Phone numbers are written in a wide variety of formats, slashes, dashes, brackets, with and without area codes, and it remains difficult to parse phone numbers accurately keeping both false negatives and false positives to an acceptable level (McCurley, 2001). Geographic feature names found in web pages also provide location information that can be looked up in a database such as the Geographic Names Information System (GNIS) to get longitude and latitude coordinate information. In an earlier study (Carrick and Watters, 1997) 90% of geographic names found in news articles were matched in a standard geographic database. In addition, personal, organization, and corporate names represent a potential source of geographic reference. Names are reasonably easy to identify, both personal (Carrick and Watters, 1997) and corporate (Rau, 1991), and could be used to identify locations by using as entries to directories of people or yellow pages that provide or confirm geographic locator information.

Feature extraction from web pages presents two types of difficulties at this time: semantic problems and efficiency problems. Using phone numbers or geographic names, for example, found in web sites to identify a single geographic location is problematic. First, care is needed to resolve conflict when more than one name or area code are found, either in descriptive narrative or in addresses. Second, care is needed to resolve ambiguity. For example, *Dartmouth*, could refer to Dartmouth College, Dartmouth in Nova Scotia, Dartmouth Street in Boston, or Ms. Dartmouth of Inuvik. The actual and accurate extraction of these features from full text remains computationally expensive and as such likely not applicable to a document store the size of the web at this time. The scale of search engines coupled with the speed expected of users makes these approaches most attractive as secondary or confirming features for a small number of sites rather than the primary feature extraction method.

Geographic information can also be derived directly from the host URL using resources available on the Web, such as Domain Name System (DNS) country codes, Whois, or IPtoLL. The country code top-level domain (ccTLD) of the DNS (DNS, 2002) uses the ISO 3166 standard two-letter abbreviations for most of the countries of the world, such as *.ca* or *.jp*. Each country is responsible for the structure of domain names within that domain name space, and this structure may be used to provide finer-grained geographic information. For example, in the United States, we find URLs, like *washington.or.us,* that use state and city references in the domain name. The *Whois* directory service (Whois, 2002) is a searchable directory service maintained by the InterNIC, a service of the US Department of Commerce (InterNic, 2002), with information about domains, sites, and contacts for *net*, *com*, and *org* domains. It currently has registration information for over 35 million host URLs with another six million on hold. *Whois* is a TCP based query/response server running on a few specific servers to support records for people, records for hosts, and records for domains. These records, such as the one shown in Figure 1 for MathResources.com, contain postal addresses and telephone numbers that are useful for extracting geographic indicators. The example *Whois* record shown in Figure 1 is interesting. Notice that the technical contact and the administrative contact are, in fact, in different parts of the country, British Columbia and Nova Scotia, and that the phone field is missing or incorrect. IPtoLL (Olson, 2002) is a tool that uses the *Whois* database to map host names to longitude and latitude values. IPtoLL resolves US sites to the city, Canadian sites to the province, and other sites to the capital city of the country.

3. GeoSearcher

Clearly the success of any location-based ranking system depends on both the ability to identify a reference location



WHOIS information for **mathresources.com**:

Registrar: NETWORK SOLUTIONS, INC.

Organization: MathResources Inc
address: 5516 Spring Garden Road, Suite 203
Halifax, Nova Scotia B3J 1G6 CA

Admin contact: David, Robinson
email: david@MATHRESOURCES.COM
phone: Hali
fax: , N.S. B3J 1G6

Tech contact: Manager, Site
email: eo@CMS.MATH.CA
phone:
fax: 613 5651539

Nameservers: cms.mathsoc.uottawa.ca
ns.cecm.sfu.ca

Lookup another domain: □

FIG. 1. Sample Whois record.

and on the ability to map this location to geographic coordinates. In this work we decided to concentrate on analysis of the URL to determine the location of the site to avoid the cost and uncertainty of content analysis of individual web pages. We chose the Getty Thesaurus of Geographic Names (Getty, 2002) to map this onto the geographical coordinates. The Getty Thesaurus contains more than one million names along with additional information of places, including nations of both the modern political world and historical worlds. The Getty Thesaurus provides geographic coordinates for each such place calculated as the mid-point of the place, political entity, or geographic feature. In the case of linear entities, such as rivers, the coordinates of the source are used. In addition we created three other tables for fast lookup: a Country Code table of top level domain country codes (ccTLD), a USA and Canada state and province table, and a USA and Canada area code table. We used the country names in English and two-letter abbreviations provided by ISO 3166 (ISO, 2001) and the corresponding geographic coordinates came from Gtrace (Periakruppan and Nemeth, 1999). The table contains 236 country entries with geographic coordinates. The USA and Canada table contains the full state and province names, with two-letter abbreviations and the Getty Thesaurus to geographic coordinates each. Three hundred area codes for USA and Canada were downloaded from SuperPhone (2001). From these we removed 800 and 888 and any unused codes leaving 258 area codes related to states and provinces.

Process

The process is relatively straightforward. The user presents a keyword query and a reference point. The system sends the query and reference point to the search engine and accepts the results. The algorithm then tries to identify the geographic coordinates for the first two hundred sites and calculates the distance, as the crow flies, from the geographic coordinates for each of these sites to the reference point. These sites are then ranked for the user in ascending order by distance. Figure 2 shows the general architecture of the system.

Geocoding

Geocoding is the process of assigning latitude and longitude coordinates to the host for each site. During each

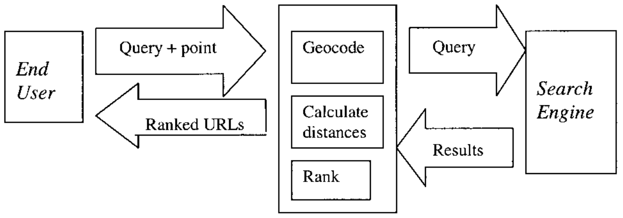


FIG. 2. Overall architecture.

session a table of hosts with their coordinates is kept for faster access. The geocoding process has several steps. First we check in the host table for the current session to see if we already have coordinates for that host. Second, we send the complete domain name to *Whois* using a perl module extension of the Net::ParseWhois (Dajoba, 2000) client for *Whois* access. This returns a two-letter country code and area code for each host, where a match was made. If the country code returned by *Whois* is *ca* or *us* (Canada or United States) and an area code is in the record, we perform a lookup in the area code table for Canada and the United States to get the corresponding province or state name. In all other cases we take the two-letter country code in the *Whois* record and lookup the country in the country table. If no record is found in the *Whois* database, the domain name is stripped down one level and resent to *Whois*. This is often successful. For example, the name *thyroid.about.com* is unsuccessful while *about.com* is successful.

Neither Country Code (such as .ca or .jp) nor the US military (.mil) top-level domains, however, are supported by *Whois*. On a test set of one hundred randomly chosen sites, we found only thirteen with a Country Code top-level domain. If no record is found in *Whois* then we examine the top-level domain (i.e., .com or .ca) and check if it is a country code. If it is a country code we first check if it matches a country code in our Country table or in the USA domain table. Finally, any unmatched names are sent to IPtoLL. Both Net::ParseWhois and IPtoLL use the *Whois* database for finding geographical location information. One would expect that results would be similar. In the test using one hundred randomly chosen sites, less the 13 country code sites, we found that while Net::ParseWhois found 68 of the 87 sites (78%) IPtoLL found only 36 of the URLs (43%). The difference seems to lie in the method of extracting the area code of the phone number. IPtoLL uses the administrative contact, which includes 800 and 888 numbers although these have no geographic coordinates. The results only overlap somewhat and so some URLs that could not be found using Net::ParseWhois are picked up by IPtoLL. All of the results are then stored in the Hosts table.

Distance Calculation

The ranking process involves distance calculations for each URL in the Hosts table from the reference location. The calculation, which depends on the spherical shape of the earth, requires both spherical geometry and trigonometry to calculate the "straight line" or "as the crow flies" distance between two points. These distances are included in the session Host table. The distance calculation, in kilometers, follows, where P_1 is the reference location and P_2 is the geographic location of the URL:

$$P_1 = \cos(\mathrm{lat}_1/57.2958) * \cos(\mathrm{lat}_2/57.2958) * \cos(\mathrm{long}_2/57.2958 - \mathrm{long}_1/57.2958)$$

$$P_2 = \sin(\mathrm{lat}_1/57.2958) * \sin(\mathrm{lat}_2/57.2958)$$

$$\mathrm{EarthRadius} = 6378$$

$$\mathrm{Distance} = \mathrm{acos}(P_1 + P_2) * \mathrm{EarthRadius}$$

Ranking

A straightforward insertion sort is used to sort the URLs in the results from the search engine according to their distance from the reference point. The current algorithm is coarse-grained using provinces and states as the smallest geographic unit. Furthermore, the coordinates used represent the midpoint of the unit and so distance calculations are from midpoint to midpoint. Refinement of geocoding depends on the development of efficient geoparsing algorithms to improve the extraction of finer-grained geographic information from a broad spectrum of web pages.

4. Prototype

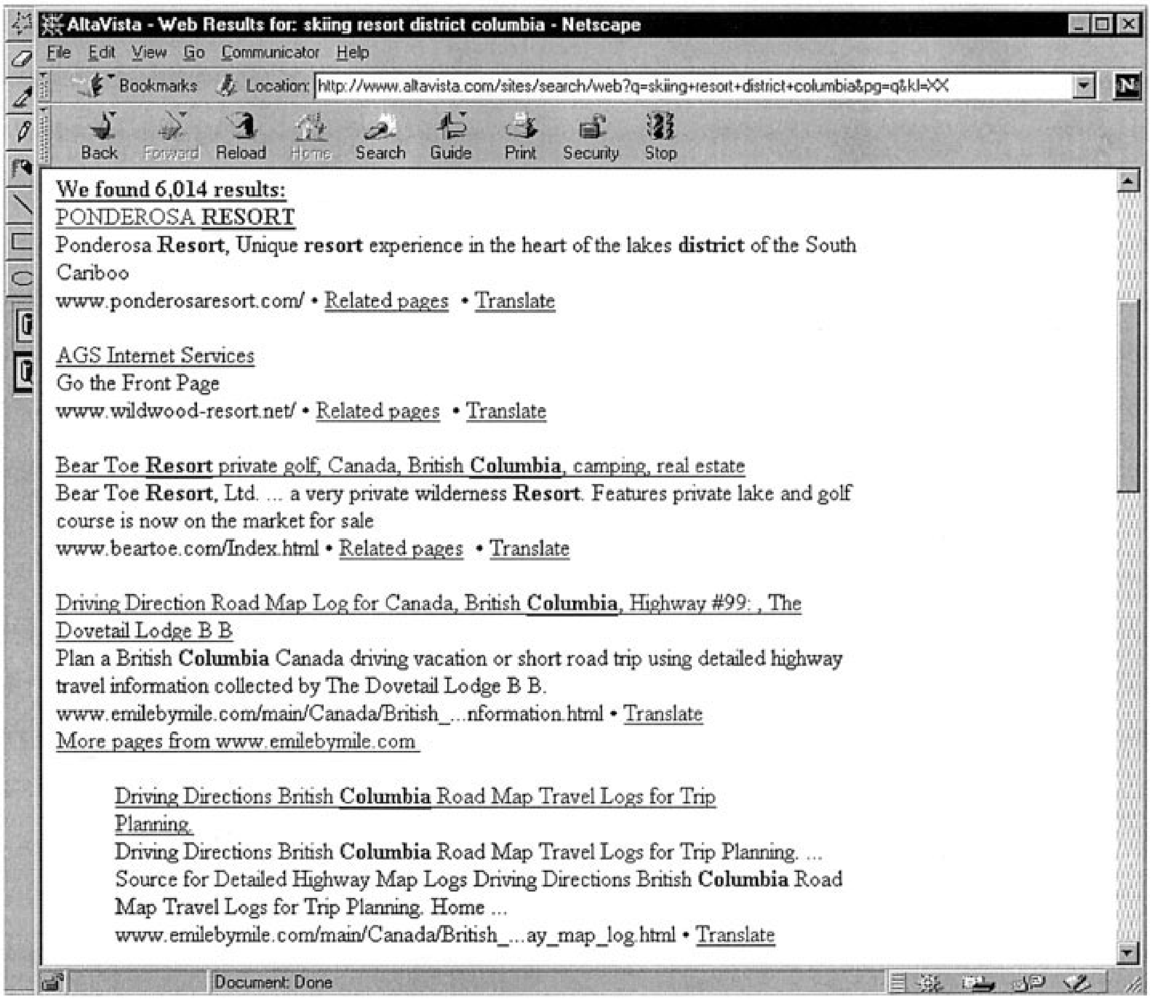
A prototype, GeoSearcher, was built to show the capabilities of this approach using searches in real time. The prototype, written in Java with servlets and perl, uses AltaVista as the search engine and the top 100 URLs returned by the search engine for which geographic coordinates can be found are ranked and presented to the user. The user enters a typical keyword query plus a location indicator. Currently the location choices are worldwide at the country and at the province and state level in North America. The user also has access to the sample searches and sets of random URLs that were used in the evaluation. As an example, we show the results for the query, *skiing resort*, and reference point, *District of Columbia*.

Figure 3, below, shows the results of the query *skiing resort district columbia* in Altavista directly. Interestingly, the query *skiing resort district columbia* ranked sites with skiing in British Columbia highly, which is some distance from Washington, DC, there being virtually no skiing resorts in DC.

Figure 4 shows the same query using the GeoSearcher prototype with the results in the order provided by AltaVista.

Figure 5 shows the same query with the results ranked using District of Columbia as the reference point.

Finally, we notice a collateral value of distance-based ranking. Distance-based ranking provides an alternative clustering dimension for the user. For example, in Figure 5, we see the sites are grouped by region. This can be used both by users looking for other activities in the same area,


AltaVista - Web Results for: skiing resort district columbia - Netscape
File Edit View Go Communicator Help
Bookmarks
Location:
Back
Forward
Reload
Home
Search
Guide
Print
Security
Stop
Document: Done

Location: http://www.altavista.com/sites/search/web?q=skiing+resort+district+columbia&pg=q&kl=XX

We found 6,014 results:

PONDEROSA **RESORT**
Ponderosa **Resort**, Unique **resort** experience in the heart of the lakes **district** of the South Cariboo
www.ponderosaresort.com/ • Related pages • Translate

AGS Internet Services
Go the Front Page
www.wildwood-resort.net/ • Related pages • Translate

Bear Toe **Resort** private golf, Canada, British **Columbia**, camping, real estate
Bear Toe **Resort**, Ltd. ... a very private wilderness **Resort**. Features private lake and golf course is now on the market for sale
www.beartoe.com/Index.html • Related pages • Translate

Driving Direction Road Map Log for Canada, British **Columbia**, Highway #99: , The Dovetail Lodge B B
Plan a British **Columbia** Canada driving vacation or short road trip using detailed highway travel information collected by The Dovetail Lodge B B.
www.emilebymile.com/main/Canada/British_...nformation.html • Translate
More pages from www.emilebymile.com

Driving Directions British **Columbia** Road Map Travel Logs for Trip Planning.
Driving Directions British **Columbia** Road Map Travel Logs for Trip Planning. ... Source for Detailed Highway Map Logs Driving Directions British **Columbia** Road Map Travel Logs for Trip Planning. Home ...
www.emilebymile.com/main/Canada/British_...ay_map_log.html • Translate

FIG. 3. AltaVista results for *skiing resort* query.

confluence of resources, or by a datamining algorithm looking for location-based connections.

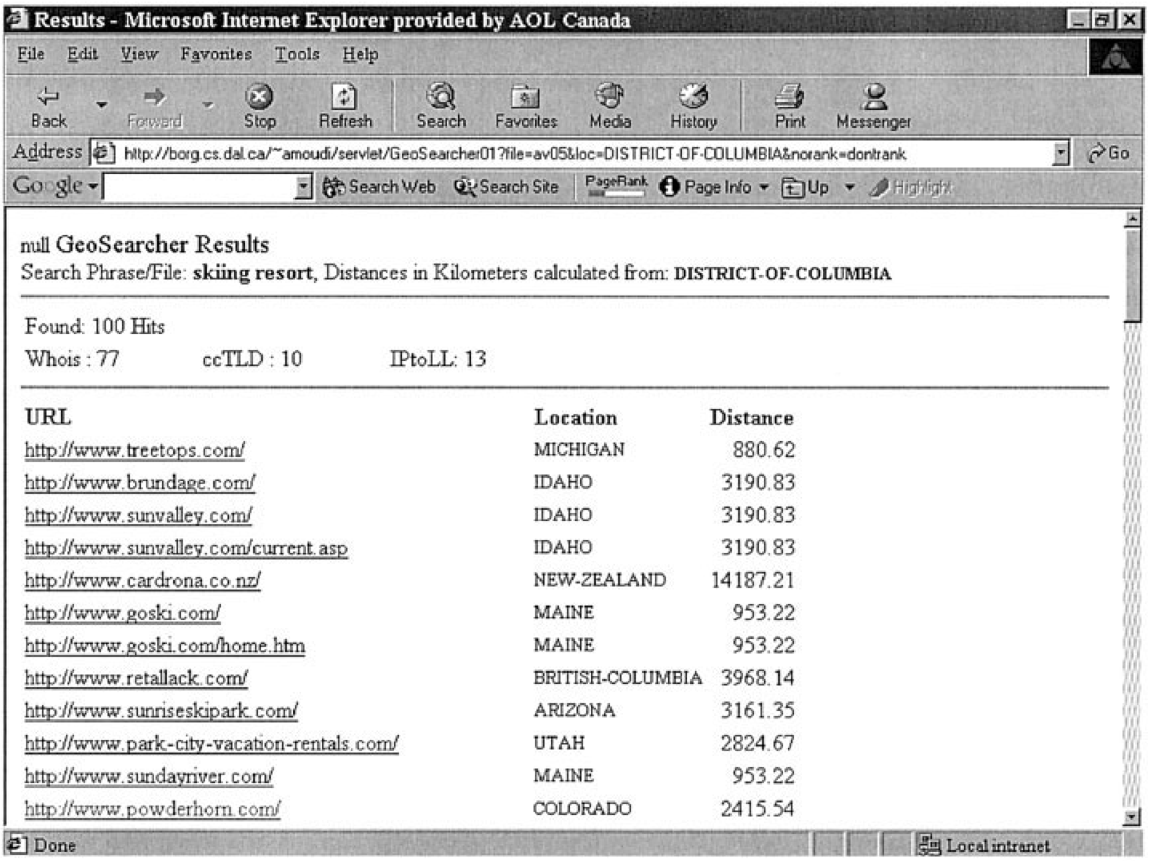
5. Evaluation of Results

We first tested the algorithm to validate its accuracy in assigning relevant location information to web sites. Second we tested the effectiveness of the algorithm in terms of the coverage of web sites that it was able to assign geographic information to, at the state, province and country level.

Validation of Accuracy

First we examined the results for a set of 100 URLs to validate that the geocoding was accurate. That is, we were concerned that when the algorithm was able to assign a geographic location to a web site that this geographic location actually reflected a useful feature of the site. The 100 URLs used for this test were generated using Yahoo random Link Generator (Yahoo, 2001). We ran the algorithm for each of these URLs and then examined each of the 100 sites manually looking for address, contact links, phone numbers, or geographic clues in the text with which we could confirm the results of the algorithm. Where country codes (e.g., .ca or .jp) were used in the URL we assumed these to be correct as the registered authority for these comes from the public and/or private sector within that territory.

In the test set of 100 URLs, the algorithm was able to assign geographic locations to 90 of the sites. The overall results of this test are given in Table 1. Of the 90 URLs that GeoSearcher succeeded in assigning geographic coordinates, we were able to check only 83 out of 90, because the actual pages of seven sites could not be accessed, either because the web page was no longer accessible or because the URL was incorrect. Furthermore, fifteen of these pages provided no verifying information that we could use for confirming the geocoding. Overall 68 (81.93%) of the 83 possible pages had enough information to be assigned locations by hand. Of those 68 sites that we were able to find information to verify the correctness of the geographic location assigned by the algorithm, 65 (95.58%) had been assigned reasonably correct geographic coordinates. The results of this test are shown in Table 1, below. This represents 78.31% of the possible 83 URLs, including those



Address: http://borg.cs.dal.ca/~amoudi/servlet/GeoSearcher01?file=av05&loc=DISTRICT-OF-COLUMBIA&norank=dontrank

null GeoSearcher Results

Search Phrase/File: **skiing resort**, Distances in Kilometers calculated from: **DISTRICT-OF-COLUMBIA**

Found: 100 Hits

Whois : 77 ccTLD : 10 IPtoLL: 13

URL	Location	Distance
http://www.treetops.com/	MICHIGAN	880.62
http://www.brundage.com/	IDAHO	3190.83
http://www.sunvalley.com/	IDAHO	3190.83
http://www.sunvalley.com/current.asp	IDAHO	3190.83
http://www.cardrona.co.nz/	NEW-ZEALAND	14187.21
http://www.goski.com/	MAINE	953.22
http://www.goski.com/home.htm	MAINE	953.22
http://www.retallack.com/	BRITISH-COLUMBIA	3968.14
http://www.sunriseskipark.com/	ARIZONA	3161.35
http://www.park-city-vacation-rentals.com/	UTAH	2824.67
http://www.sundayriver.com/	MAINE	953.22
http://www.powderhorn.com/	COLORADO	2415.54

FIG. 4. Unranked Geosearcher results for *skiing resort*.

pages without enough information to verify the geocoding. Interestingly, 15 of the 83 pages, or 18.07%, had no geographical clues and one could conclude that they may be either misrepresented by geographic coordinates based on the host URL or that if a site does not have enough information to verify a geographic location that it is not likely to be relevant to a location-based search.

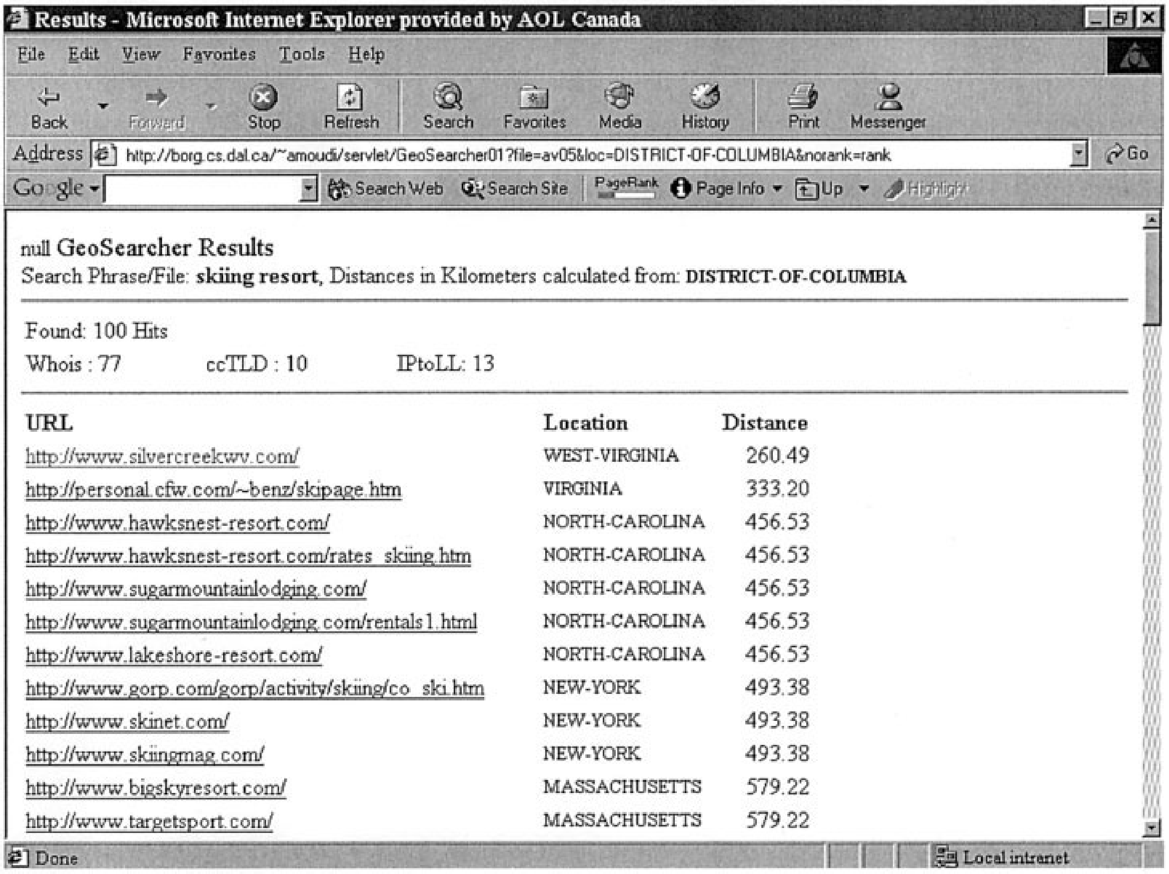
The manual inspection of the 68 pages with geographic clues provided some insight into which components of a web page would benefit most from more sophisticated analysis. We used country codes, addresses, phone numbers, and other content clues such as geographic reference in descriptive text. Table 2 shows the breakdown of these determinations.

The Country Code in the URL was used to verify non North American locations; United Kingdom, Netherlands, Belgium, Spain, Brazil, South Africa, and Australia. The address information, often found in the *about* page was used to confirm an additional 27% of the locations. Information in the page content was used in 20% of the cases if direct address information was not readily available. As an example of content clues, the site *www.successlink.org* had a main header *Linking Educational Innovators Across Missouri* plus a sponsorship logo, *SuccessLink is sponsored by the Missouri Department of Elementary and Secondary Education*, from which we determined a geospatial interest in Missouri, which confirmed the result of the algorithm.

The three sites, assigned coordinates by the algorithm but for which we found a conflict between the location assigned by the algorithm and that assigned manually based on the page contents are interesting. One was *broadcast.com*, a *yahoo* site, for which we could not actually find either an address or specific content in the document. The other two were news providers for non-North American audiences, *Africanews.org* and *Arabnews.com*, which were using North American ISPs. In the case of the *Africanews.org* site the URL indicated a North Carolina location while the content of the page indicated a Washington DC site, which we attributed to physical location for the office and ISP differences. In both these cases, one could make a case that these sites are intended for North American readers but we determined that this would not likely be relevant in a location-based query.

Coverage Effectiveness

After validating that the algorithm produced appropriate geographic location information when it could be applied, we needed to determine the coverage or proportion of web sites for which it could be applied. We use coverage effectiveness to measure this, where coverage is the proportion of web sites on which the algorithm is used that the algorithm was successful in assigning geographic coordinates. We ran


Results - Microsoft Internet Explorer provided by AOL Canada
File Edit View Favorites Tools Help
Back Forward Stop Refresh Search Favorites Media History Print Messenger
Address
Go
Google
Search Web
Search Site
PageRank
Page Info
Up
Highlight
Done
Local intranet

Address: http://borg.cs.dal.ca/~amoudi/servlet/GeoSearcher01?file=av05&loc=DISTRICT-OF-COLUMBIA&norank=rank

null GeoSearcher Results

Search Phrase/File: **skiing resort**, Distances in Kilometers calculated from: **DISTRICT-OF-COLUMBIA**

Found: 100 Hits

Whois : 77 ccTLD : 10 IPtoLL: 13

URL	Location	Distance
http://www.silvercreekwv.com/	WEST-VIRGINIA	260.49
http://personal.cfw.com/~benz/skipage.htm	VIRGINIA	333.20
http://www.hawksnest-resort.com/	NORTH-CAROLINA	456.53
http://www.hawksnest-resort.com/rates_skiing.htm	NORTH-CAROLINA	456.53
http://www.sugarmountainlodging.com/	NORTH-CAROLINA	456.53
http://www.sugarmountainlodging.com/rentals1.html	NORTH-CAROLINA	456.53
http://www.lakeshore-resort.com/	NORTH-CAROLINA	456.53
http://www.gorp.com/gorp/activity/skiing/co_ski.htm	NEW-YORK	493.38
http://www.skinet.com/	NEW-YORK	493.38
http://www.skiingmag.com/	NEW-YORK	493.38
http://www.bigskyresort.com/	MASSACHUSETTS	579.22
http://www.targetsport.com/	MASSACHUSETTS	579.22

FIG. 5. Ranked GeoSearcher results for *skiing resort*.

two tests to evaluate the coverage effectiveness of the algorithm. The first test examined URLs returned as sample query results and the second examined URLs generated by Yahoo as large unconnected somewhat random URLs.

Search results

To test the effectiveness of the algorithm in determining geographic coordinates to the URLs of actual query results, 60 queries were taken from MetaSpy (Metaspy, 2001), ten at a time over a week. MetaSpy is a web tool provided by MetaCrawler, a metasearch engine that uses several search engines (AltaVista, DirectHit, Excite, FindWhat, LookSmart, and Overture) simultaneously and combines and ranks the results for the user. MetaSpy is a dynamic tool that shows ten searches currently being executed. This site was chosen to generate queries for our test as the queries represent a wide variety of real searches and produce reasonably small answer sets. Queries with serious misspellings were not used. The queries and success rate for the algorithm are shown in Table 3. The average result set size was 28.95 URLs with standard deviation of 6.28. In this test sample of queries, the overall success rate of the GeoSearcher algorithm using the combined methods was 83.27% with standard deviation of 10.91.

The algorithm uses a combination of three methods: country code (ccTLD), Net::ParseWhois, and IPtoLL. The order of use is first Net::ParseWhois, ccTLD, and then IPtoLL. These methods have iterative combined success rates. The success rate for the individual methods, used in this sequence, shown in Table 4 is interesting. Recall that the algorithm sends the URL first through Net::ParseWhois, then ccTLD, and finally IPtoLL so that each successive comparison is made only on the remaining URLs. In this case Net::ParseWhois returns geocode information on average for 43.39% of the URLs, then country code (ccTLD)

TABLE 1. Validation results.

Total pages	Location determined by Geosearcher	Pages available online	Location determined manually	Correct location
100	90	83	68	65

TABLE 2. Breakdown of validation information.

Total	Country code	Address	Content clue	No clue
83	28	23	19	15
%	33.73	27.71	20.48	18.07

TABLE 3. Results for MetaSpy queries.

MetaSpy query	URLs in result	Algorithm success	Algorithm fail	Percent success
anthrax	31	23	8	74.19
IP address lookup	32	31	1	96.88
friendship poems	33	32	1	96.97
interior decorating	29	22	7	75.86
outdoor pool images	37	35	2	94.59
christmas piano music	38	32	6	84.21
easy recipes	32	27	5	84.38
moon phases	37	26	11	70.27
internet security	30	26	4	86.67
science olympiad	36	33	3	91.67
testing blood sugar	34	31	3	91.18
armour thyroid medication	34	21	13	61.76
cincinnati enquirer	24	20	4	83.33
cisco hubs	26	24	2	92.31
rome map	19	17	2	89.47
dutch oven	32	25	7	78.13
hurricane tracking	39	26	13	66.67
hierarchical+methods	17	17	0	100.00
university of missouri columbus	32	29	3	90.63
math questions	32	27	5	84.38
government of canada	16	15	1	93.75
weight lifting equipment	31	27	4	87.10
tim berners-lee	33	29	4	87.88
body piercing	27	21	6	77.78
italian dictionary	22	18	4	81.82
hotels	32	24	8	75.00
online pet store	29	24	5	82.76
notre dame	30	26	4	86.67
qualitative research method	34	29	5	85.29
harry potter figures	20	16	4	80.00
skiing	33	26	7	78.79
real estate qld	50	23	27	46.00
oil & gas services companies	22	19	3	86.36
nintendo game cube	34	27	7	79.41
london england	30	26	4	86.67
thomas cook	23	19	4	82.61
rolex watches	32	17	15	53.13
university of the phillipines	23	20	3	86.96
marriott	24	23	1	95.83
chat rooms	31	21	10	67.74
interest rates	24	18	6	75.00
best buy	23	21	2	91.30
gardening supplies	19	19	0	100.00
baby products	32	28	4	87.50
singer sewing machine parts	22	15	7	68.18
time zones	26	21	5	80.77
wireless phone companies	31	31	0	100.00
toshiba	28	20	8	71.43
christmas decorations	24	21	3	87.50
fertility specialists	23	20	3	86.96
smoking public	21	18	3	85.71
san francisco times	27	25	2	92.59
ancient chinese culture	35	32	3	91.43
tax refund	29	22	7	75.86
discounted toy outlets	37	34	3	91.89
public relations magazines	28	26	2	92.86
statue of liberty	24	18	6	75.00
jewelry stores	26	21	5	80.77

TABLE 4. Individual average success rates.

Method	Average percent success	Standard deviation
Net::ParseWhois	43.39	13.36
ccTLD	5.17	7.49
IPtoLL	34.22	14.87

matches 5.17%, and finally IPtoLL matches 34.22% of the URLs.

From the large standard deviations we can see that no single method can handle the variety of URLs we can expect on the web consistently but that the three together cover the majority of cases, i.e., where one does badly another will step up. For example, the simple country code lookup was effective with the queries *thomas cook* and *real estate*, while IPtoLL was effective with *science olympiad* and *qualitative research methods* and Net::ParseWhois for *weight lifting equipment* and *online pet store*. If one could detect the pattern then the most appropriate locator algorithm could be used first to improve efficiency.

Following the initial testing with MetaCrawler, the sixty queries were used to test the prototype first with Google and then with AltaVista as the search engines.

Random Sites

To test the algorithm effectiveness on general web sites, i.e., not ones in response to specific user queries, we generated ten sets of one hundred URLs using the Yahoo Random Link generator (Yahoo, 2001). Although the URLs in these sets are likely not strictly speaking random, they represent a fair cross section of web sites. We processed each set of one hundred URLs using the GeoSearcher algorithm and calculated the following metrics: percentage of successful host mapping, and the overlap between individual methods of mapping (i.e., ccTLD, IPtoLL, and Net::ParseWhois).

Table 5 shows the overall results of the algorithm in geocoding the URLs of the test sets. We see that on average the algorithm was successful in identifying a geographic location for 80% of the cases, standard deviation of 5.16, with low of 74% and a high of 90%.

TABLE 5. Geocoding effectiveness.

Set number	Percent success	Percent failure
1	90	10
2	81	19
3	78	22
4	77	23
5	76	24
6	81	19
7	80	20
8	88	12
9	74	26
10	77	23
Average	80.2%	19.8%

TABLE 6. Sample overlap results.

URL (www.)	ccTLD	IPtoLL	Net::ParseWhois
Railpace.com	Not found	Not found	New Jersey
Ucc.co.kr	Korea	Korea	Not found
Shoefits.com	Not found	Not found	New Jersey
Members.tripod.com	Not found	Massachusetts	Massachusetts
Geocities.com	Not found	California	California
Alfplan.com	Not found	Not found	Not found
Superfinedesign.com	Not found	Not found	California
Theunionleader.com	Not found	Not found	California
Milknhoney.co.il	Israel	Israel	Not found
Ckshoefty.com.hk	Hong Kong	Hong Kong	Not found

The URLs for which the algorithm could not get a geographic location were tested manually to see if there was a pattern in the failures. To do this AllWhois (AllWhois, 2002) was used. AllWhois is an integrated interface to most of the Whois servers worldwide that provides online domain name searches. Each of the failures was sent to this service individually and we discovered three general patterns in the failures. The first group of sites was registered by registrars not supported by Whois. The second group contained those sites with domain names that are currently for sale and not registered. Finally, the third group had sites with errors in the URL.

Overlap

Although the algorithm uses three sources in combination, Net::ParseWhois, IPtoLL, and the country code top-level domain table, we were interested in measuring the overlap in using these sources. A special version of GeoSearch was written to process three sets of one hundred URLs used in the testing. For this test, each URL was sent to each component algorithm independently. Table 6 below shows a sample of the results. We can see that generally one of the sources will be helpful and often only one.

The findings from this test were that (as suspected) ccTLD and Net::ParseWhois have no overlap, Net::ParseWhois and IPtoLL have incomplete overlap and that at least IPtoLL and Net::ParseWhois would both need to be used to gain good coverage. For example, *www.India-today.com* was correctly located by Net::ParseWhois but not found by IPtoLL.

The overlap of country code for non-North American URLs is nearly complete. For example, in the first test set, 28 URLs had an identifiable non-North American country code, and 27 of these where identified correctly by both ccTLD and IPtoLL. The advantage of keeping a local ccTLD country code table is that it is small and very easy to access at run time. The overall results for the 300 URLs are given below in Table 7.

While there is considerable overlap at about one third in the coverage provided by the use of Net::ParseWhois and IPtoLL it is only with the combination that we achieve over 80% coverage.

6. Discussion

For those queries with a geospatial dimension, combining keyword queries with geographic location and or distance features provides an alternate ranking for users. In this paper we explored the integration of web facilities to provide geographic coding from the URL to permit a ranking of search engine results simply and dynamically.

The combined results of using country code (ccTLD), *Net::Whois* and *IPtoLL* to identify a geographic location associated with a given URL provides a success rate of over 80% in several tests. We tested to verify that we needed a combination of methods and found that *Net::ParseWhois*, at 66% coverage, and *IPtoLL*, at 50% coverage, did provide overlapping coverage, but that neither one alone provided adequate effectiveness. Evaluation tests showed that the algorithm was accurate and effective. The reliance on real-time queries to *Whois* or through *IPtoLL* is not, however, efficient enough to be used in real time for very large search results. A database or special resource would be require to reduce the number of requests in real time to the online

TABLE 7. Overlap.

Set number	URLs	Whois + ccTLD	IPtoLL	Overlap	All methods
1	100	73	59	42	90
2	100	58	48	25	81
3	100	67	43	32	78
Average %		66%	50%	33%	83%
Standard deviation		7.55	8.19	8.54	6.24

Whois database in order to increase the efficiency enough to handle quantities of queries and large sets of search engine results.

An advantage of using tools such as *Net::ParseWhois* and *IPtoLL*, which also uses *Whois*, lies in removing concerns of maintenance and optimization from the application. These advantages apply, as well, to working with the output of existing search engines rather than creating a new search application. Building intelligent front ends to modify user queries for search engines would also be useful for location-based queries but requires more sophisticated analysis of the query terms, context, and intent.

We used this geographic coordinate information to reorder the results of search engine queries, in a prototype system, by calculating their distance to a geographic reference point, chosen by the user. The prototype system, GeoSearcher, illustrates the use of the algorithm on the web using Google and AltaVista search engines. In addition to providing distance-related ranking of the sites for the user the ranking also provides a clustering by location.

While this work makes a contribution to retrieval on the web by providing an effective algorithm for providing alternative ranking order for search engine results based on geographic location, it represents a baseline for further work. In particular, the granularity of the geocoding needs to be refined while maintaining the same high level of accuracy and effectiveness. Second, the potential usefulness of geospatial information related to web sites is much broader than simply improving the order of search engine results.

The algorithm described in this paper provides coarse-grained geospatial ranking. That is, for the USA and Canada geocodes are provided at the state or province level while for other countries the geocodes are at the country level only. While this is helpful to some level, clearly one would want to extend the algorithm to capture finer-grained geospatial information. Currently there is a tradeoff between consistency and granularity. For those sites with identifiable area codes, cities can be determined. For those sites with postal codes, streets or regions within cities or rural areas can be identified. However, currently, only about 10% of sites provide useful area codes or postal codes. Furthermore, the process of extracting these with high precision from the content of web sites is complex and not yet well done.

An area of use of geocoding that we did not follow but that became apparent during our testing was in feature extraction for web datamining particularly for e-commerce related explorations. For example, extensions of Buyukkokten et al's work (1999), which used the geocoding of sites to examine the geographic distribution of links pointing to given sites, could be used to examine concentrations of services in regions or regions of overlap.

Both query formulation and result presentation used in the prototype were quite simple but the latitude and longitude information could be used to support more sophisticated systems. For example, query formulations could include features such as near, within, further west, or map based non-textual input. At the same time, alternate presentation of the results, such as visual mapping systems, may be effective for a variety of tasks.

Geocoding also has potential in the dynamic typing of hypertext links on the web. This use would be two-part: First, as an additional feature of the embedded hypertext link that the user could request as needed. Second, to arrange or cluster links from an entire site for use at the overview level. Recognizing that not all web information-seeking behavior has a geospatial basis geocoding that can be done dynamically on an as needed basis has appeal.

Finally, the increasing popularity of mobile devices makes the integration of search and geospatial factors more important. Finer-grained geospatial coordinates than were used in this research will be needed, however, to find sites within city boundaries or blocks. Furthermore, the mapping of site coordinates to digital maps, which has been done elsewhere, needs to be integrated along with suggested routes to get to the destination from the current position of the user.

This work makes a contribution to those web users who can benefit from an alternative search engine result ranking based on geographic location. This means that users with queries with a geospatial concern can more easily manage the results of general search engine results. Ranking of search engine results based on task oriented features is becoming more important as more searching is done on small screens, on which only a few hits can be expected to be viewed from any given search. The use of geospatial information for ordering is just one of many reranking possibilities. Others include ranking by date of last update, domain of interest, language, or even the presence of images, music, or videos.

Acknowledgement

This work has been supported by the National Science and Engineering Research Council of Canada.

References

AllWhois. (2002). OnLine at: [www.allwhois.com] Available: March 15, 2002.

AltaVista. (2002). AltaVista Search Company. Online at: [www.altavista.com] Available: March 11, 2002.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: Addison-Wesley (ACM Press series).

Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., & Shivakumar, N. (1999). Exploiting geographical location information of web pages. Proc. of the ACM SIGMOD Workshop on the Web and Databases (WebDB'99).

Carrick, C., & Watters, C. 1997. Automatic association of news items. Information Processing & Management, 33(5), 615-632.

Chakrabarti, C., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., & Rajagopalan S. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. Proceedings of the 7th World-Wide Web Conference. Brisbane, Australia. 1998.

Dajoba. (2000). Dajoba Project: Net::ParseWhois. Online at: [www.dajoba.com/projects/netparsewhois] Available: March 11, 2002.

De Floriani, L., Marzano, P., & Puppo, E. (1993). Spatial queries and data models. In Frank, A.& Campari, I. (Eds.) Spatial information theory: a theoretical basis for GIS. Springer Verlag Lecture Notes in Computer Science, 716.

Ding, J., Gravano, L., & Shivakumar, N. (2000). Computing geographical scopes of web resources (pp. 113–138). Proc. of the 26th International Conference on Very Large Data Bases (VLDB'00).

DNS. (2002). Domain name system. OnLine at:[www.dns.net] Available: Feb.11, 2002.

Egenhofer, M.J. (1994). Spatial SQl: a query and presentation language. IEEE Trans. On Knowledge and Data Engineering, 6(1), 86-95.

Frew, J., Carver, L., Fischer, C., Goodchild, M., Larsgaard, M., Smith, T., & Zheng, Q. (1995). The Alexandria Rapid Prototype: building a digital library for spatial information. In 1995 ESRI User Conference Proceedings. OnLine: [www.esri.com/library/userconf/proc95/to300/p255.html] Available: Feb.11, 2002.

GenieKnows. (2002). GenieKnows MetaCrawler. Online at: [www.genieknows.com] Available: March 11, 2002.

GeoTags. (2002). GeoTags Search Engine. Online at: [www.geotags.com] Available: March 11, 2002.

Getty Thesaurus of Geographic Names. (2002). OnLine at: [shiva.pub.getty.edu/research/tools/vocabulary] Available: Feb 11, 2002.

Google. (2002). Google Search. Online at: [www.google.com] Available: March 11, 2002.

InterNIC (2002).Online at: [www.internic.org] Available: Feb.11, 2002.

ISO (2001). ISO 3166 Maintenance Agency. Online at: [www.din.de/gremien/nas/nabd/iso3166ma] Available: November 10, 2001.

Larson, R.R. (1996). Geographic Information Retrieval and Spatial Browsing in GIS and Libraries: Patrons, Maps and Spatial Information, edited by Linda Smith and Myke Gluck, Urbana-Champaign : University of Illinois. p. 81-124.

McCurley, K. (2001). Geospatial mapping and navigation of the web**.** Proc. of the WWW10 Conf. May 1-5, 2001, Hong Kong. p. 221 –229.

Mardis, M. (2001). Uncovering the hidden web, Part I: Finding what the search engines don't . ERIC Digest. October, 2001. OnLine at: [www.ericit.org/digests/EDO-IR-2001-02.shtml]. Available: Feb 11, 2002.

MetaSpy. (2002). OnLine at: [www.metaspy.com] Available: Feb.11, 2002.

Northern Light. (2002). Geosearch. Online at:[www.northernlight.com/geosearch.html]. Available: Feb.11, 2002.

Notess, G. (2001). Search Engine Statistics OnLine at: [www.searchengineshowdown.com/stats/size.shtml] Available: Feb.11, 2002.

Olson, R. (2002). Host Name to Latitude/Longitude. Online at: [cello.cs.uiuc.edu/cgi-bin/slamm/IPtoLL/] Available: Feb.11, 2002.

Periakaruppan, R., & Nemeth, E. 1999. Gtrace—A graphical traceroute tool. Online at: [www.caida.org/tools/visualization/gtrace/paper/Gtrace.doc] Available: September 15, 2001.

Rau, L.F. (1991). Extraction company names from text, Seventh IEEE Conference on Artificial Intelligence Applications p. 29-32.

Search Engine. (2000). Survey reveals search habits. The Search Engine Report. June 2. Online At: [searchenginewatch.com/sereport/00/06-realnames.html] Available: Feb.11, 2002.

SuperPhone. (2001). Online at: [www.superphone.net] Available: November 10, 2001.

Whois. (2002). Whois.net. Online at: [www.whois.org] Available: Feb.11, 2002.

Woodruff, A.G., & Plaunt C. 1994. GIPSY: Geo-referenced Information Processing System. Journal of American Society for Information Science. 45, 645-655.

Yahoo. (2002). Random Yahoo Link OnLine at: [random.yahoo.com/bin/ryl] Available: Feb.11, 2002.

Yokoji, S., Takahashi, K., & Miura, N. (2001). Kokono search: a location based search engine. Proceedings of the Tenth International World Wide Web Conference. Hong Kong. p. 1146. OnLine at:[www10.org/cdrom/posters].