# Geospatial Mapping and Navigation of the Web

Kevin S. McCurley
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
U.S.A.

## ABSTRACT

Web pages may be organized, indexed, searched, and navigated along several different feature dimensions. We investigate different approaches to discovering geographic context for web pages, and describe a navigational tool for browsing web resources by geographic proximity.

*Keywords:.* Geospatial information retrieval, geographic information systems, browsers, navigation.

## 1. INTRODUCTION

Since the early days of the world wide web, many different efforts have been made for organizing, navigating, and searching documents. Simple inverted keyword searches (perhaps not so simple now that the web has grown so large!) continue to provide one of the most useful techniques for users to find information on a given topic, but they also form a starting point for a number of higher-level semantic queries that can be used in an information search. For example, a reader may only be interested in finding documents that that are written in Japanese, or that were created or modified after a certain date. A reader may also only be interested in pages that are written for a fifth-grade reading skill. Such information retrieval tasks are not well served by a simple keyword search, and systems organized to answer such queries generally require higher-level semantic information about the content.

Navigational tools may also rely on information that has been extracted during a crawl and stored in a database. One novel feature for navigation is the "What's Related" service provided by Alexa corporation[14], and incorporated into both Netscape Communicator and Microsoft Internet Explorer. A further example was provided by the "backlinks browser" presented by the author at this conference two years ago[5], allowing a user to traverse hyperlinks in the reverse direction. One might also imagine various ways of creating a temporal browsing mechanism, offering the reader the chance to view the change history of a resource referenced by a URL as time progresses.

This paper addresses the problem of indexing and navigation of web resources by geospatial criteria. Information retrieval based on geographic criteria is a fairly common task. Examples include travelers who wish to learn what opportunities exist in a destination, students who are studying another part of the world, intelligence analysts preparing a report on a given area, business planners investigating an area for expansion, or government planners who wish to focus on environmental or economic impact in a specific region. For a good introduction to the field of information retrieval using geographic criteria, see [15].

In addition to geographic information retrieval and navigation of Internet content, geography plays an important role in other aspects of the Internet. Advertising is much more effective for both consumers and advertisers when it offers products and services that are readily available in the area of the consumer. Content that is tailored to a region is potentially more useful to users from that region. For example, local customs and laws influence the expectations of users when they interact with the network. Note that a user or server need not have a static location, since mobile usage of the Internet is expected to grow rapidly in the future. We call such geographic contexts *entity-based*, since they represent a feature of the user or computing system itself.

A more interesting and subtle source of geographic context is provided by the content itself (we refer to these as *content-based* contexts). For example, the web page of a company may contain the address or phone number of the company headquarters, conveying a geographic context within the page. A page reporting on a news event should be assigned a geographic context of the location where the event took place. Note that a content-based context for a web page may be inferred from the entity-based context of the server that hosts the page, since the URL itself may be regarded as part of the content.

Geospatial contexts may be point locations, or arbitrarily shaped regions. The regions need not be contiguous (e.g., the country of Japan consists of many islands). They need not even be planar; one interesting feature of our project is that we will eventually be able to answer queries about all pages with geographic contexts for land locations below sea level. A web resource will very often have multiple geographic contexts, and it is impossible to assign just one (e.g., a personal page of a person who lives in one place, but comments about a site they visited on vacation).

Many pages on the web have no strongly determined ge-


*WWW10,* May 1-5, 2001, Hong Kong.
ACM 1-58113-348-0/01/0005.

ographic context, so it would be misleading to expect such for every page on the net. Other pages have a very obvious geographic context (e.g., those containing a postal address for a physical location). The significance of a geographic context is difficult to determine, but multiple indicators of proximate geographic context can be used to improve the assurance of correlation to a geographic location. Moreover, some indicators give higher confidence than others. Consider for example the two cases of recognizing the phrase "Caspian Sea" within a web page vs. recognizing a phone number that is located on the shore of the Caspian Sea. The phone number may be included only as a way of reaching a person who has further information about the actual topic of the page, and their residence on the Caspian Sea may in fact be irrelevant. By contrast, unless the term "Caspian Sea" has another recognizable meaning, there is little reason to doubt the geographic significance of this reference. While it is always possible that a geographic indicator is only incidental, it still provides a way to rank documents according to their geographic relevance. Those that contain a geographically significant indicator are clearly more relevant to a geographically constrained information retrieval task than those that do not.

The field of Geographic Information Systems (GIS) deals primarily with data sets of geographic significance, such as demographic data, weather data, or geologic data. In this field the process of recognizing geographic context is referred to as *geoparsing*, and the process of assigning geographic coordinates is referred to as *geocoding*. In this paper we report on efforts to geoparse and geocode web pages, and some of the opportunities that this offers for navigation, indexing, and organizing.

It is natural to wonder what fraction of web pages contain a recognizable geographic context. Based on experiments with a fairly large partial web crawl, we found that approximately 4.5% of all web pages contain a recognizable US zip code, 8.5% contain a recognizable phone number, and 9.5% contain at least one of these. The actual fractions may be somewhat higher - these are ones that are actually recognized by our conservative parser. This does *not* say that the use of geographic context is limited to only 10% of the web, because there are other indicators that may be used, and there also exist opportunities for inference of implicit geographic context for every web resource from the entity-based context of the server (see section 2).

Our primary focus in this paper is on information retrieval and navigation. In section 5 we describe the implementation of a navigational tool that allows the user to browse web pages by their geographic proximity rather than their hyperlink proximity. Geospatial information for web pages has many other uses as well. For example, when combined with an inverted keyword index, we can perform queries such as "find me pages that talk about biking trails in a geographic context within 50 kilometers of my brother's home". Such queries may not easily be answered with a simple inverted keyword index, for it is often not clear how to formulate a keyword search that will exclude pages from outside of a given region.

We are not the first to recognize and consider the use of geospatial information in web sites. Buyukokkten et. al. [2] studied the use of several geographic keys for the purpose of assigning site-level geographic context. By analyzing "whois" records, they built a database that relates IP addresses and hostnames to approximate physical locations. By combining this information with the hyperlink structure of the web, they were able to make inferences about geography of the web at the granularity of a site. For example, they were able to confirm from their study that the New York Times has a more global reputation than the San Francisco Chronicle, based on the fact that is linked to from a more geographically diverse set of sites.

Among commercial sites, Northern Light has very recently started offering a "GeoSearch" capability, where keyword searches can be constrained by proximity to a specified address or phone number within the USA and Canada. They apparently recognize addresses within web pages, and use an address database to locate web pages containing nearby addresses. Yahoo offers a "Yellow Pages" service for businesses in the USA, with search and ontologies organized along geographic as well as business lines. Their site appears to have been built from commercial databases of businesses, and makes no attempt to organize non-business web pages. In the future the author expects to see a flood of development in this direction, partly in order to address the needs of mobile IP users.

## 2. SOURCES OF GEOSPATIAL CONTEXT FOR HOSTS

Context for servers or sites may be derived from any information that is correlated to that machine. One useful source (used in [2]) is the Whois database for domain registrations. This database was originally developed by the NIC for the purposes of problem solving and accountability, and contains phone numbers, addresses, and email addresses for administrative and technical points of contact for each domain. After the commercialization of the Internet, the responsiblity for this database has been dispersed to several organizations and usage is now constrained by proprietary interests. Note that the IP2LL package[13] uses this approach to assign latitude and longitude to hostnames. We constructed a similar package in C++ for determining entity-based geographic context of servers.

While the whois database is quite useful, it is also possible to often derive site-level information of this degree of accuracy by analyzing the predominant geographic context of individual pages on that site. Commercial sites often contain "contact" or "employment" pages that are linked from the top level or from many places on the site. Additional heuristics are provided by the use of hostname aliases and hosts that share the same IP address.

An additional source of geographic information is related to the way that traffic is routed on the Internet. IP packets are typically forwarded through a sequence of hops in routers. The route is dynamic, but most sites are connected to the backbone by only one or two individual connections. Because of the way communication services are sold, these backbone connection points are typically located a relatively short physical distance from the server itself. The traceroute utility provides a mechanism of discovering the last hop before reaching the server, and a truncated example of traceroute output is shown in figure 1.

From a database of physical locations of these connection point routers, it is possible to determine the approximate location of any server connected to the Internet with high probability. It is possible to determine the identity of these

```
-> traceroute www.ieee.org
traceroute to sphinx5.ieee.org (199.172.136.40), 30 hops max, 40 byte packets
 1  tenkuu (198.59.115.1)
 2  198.59.183.1 (198.59.183.1)
 3  Serial5-1-0.GW1.PHX1.ALTER.NET (157.130.227.209)
 4  143.ATM3-0.XR2.LAX2.ALTER.NET (146.188.249.134)
 5  152.63.112.162 (152.63.112.162)
 6  131.at-5-0-0.TR2.NYC8.ALTER.NET (152.63.5.142)
 7  184.ATM6-0.XR2.EWR1.ALTER.NET (152.63.20.233)
 8  192.ATM1-0-0.HR2.NYC2.ALTER.NET (146.188.177.37)
 9  Hssi11-0.New-Brunswick4.NJ.ALTER.NET (137.39.100.61)
10  IEEE-gw.customer.ALTER.NET (137.39.228.50)
11  sphinx.ieee.org (199.172.136.8)
```

**Figure 1: Traceroute from swcp.com to www.ieee.org. Note that ALTER.NET uses airport codes and city names in their hostnames, so it is easy to determine the city through which a packet travels. In this case the last hop before entering the connection to www.ieee.org is located in New Brunswick, New Jersey, and the IEEE operations center is actually located in Piscataway, New Jersey.**

backbone access routers the traceroute protocol and utility. This is the method used by the Gtrace tool[17] to display an approximate location of servers connected to the Internet. The data collected in this way is of very limited accuracy however.

Parts of the Domain Name System (DNS) provide a very rough determination of geographic context for a host. A study by the Internet Software Consortium[12] revealed that approximately 25% of all second-level domains belong to a top level domain identified with a country (e.g., `.cn`), and an even higher percentage of hosts belonged to one of these domains. Some of these domains are further partitioned by geography. Most of the `.us` domain has a second level derived from state abbreviations, and a third level based on city or county. SRI has recently proposed [19] to ICANN that the DNS system incorporate a new `.geo` top level domain that will further be broken down by region specified in latitude and longitude. Such a system would be very useful for deriving host-based geographic context.

RFC 1876[7] defined a LOC resource record format for location records in the Domain Name System (DNS), but few sites maintain these records. In addition, the DNS provides a few other records that provide indirect hints on location of servers. These include the GPOS resource record[9] (obsoleted by RFC 1876), and TXT records, which may contain arbitrary text describing the server. For mobile applications, RFC 2009[11] defines a way of encoding GPS information into DNS. In practice these are currently a rare source of geospatial information about hosts, since few sites utilize them.

# 3. SOURCES OF GEOSPATIAL CONTEXT FROM CONTENT

In previous work, attention was paid to associating geospatial information with hosts. In practice, the *content* of a URL may have a geospatial context quite different from the host that serves the URL. For example, a personal home page may provide information about a vacation experience, or a company site may be hosted by a third party whose physical location is uncorrelated to the company they are serving. Companies may have several offices, but have their web servers hosted in one centralized location.

The geospatial context may be recognized from the content by a large number of clues, with varying degrees of precision and reliability. For example, one obvious source of a coarse geographic context arises from the choice of language used for content. The use of Chinese in a web page does not in itself imply that the page has a geographic context in China, but the correlation of geography and language is still fairly high. The recognition of language from a resource is facilitated by a number of factors, including `meta` tags in HTML pages, HTTP response headers, and direct linguistic analysis.

In order to assign more accurate geographic context to web resources, we have devised several strategies for analyzing content to recognize more specific indicators. The rest of this section is devoted to discussion of some of these.

## 3.1 Addresses and Postal Codes

One obvious source of geospatial information is from postal addresses, which have evolved over centuries to facilitate the delivery of physical mail to a specific location around the world. Recognition of mail addresses is a fairly well studied problem, but is complicated by the fact that standards for formatting vary considerably from one country to another.

Postal addresses are typically segmented into a variety of fields, separated by commas or line breaks. They may be relative, assuming for example that the reader already knows that the address is within the United States, and omitting the country. They often contain misspellings or variations on format. Software for performing this parsing is more complicated than one might first expect, due to the wide variation in abbreviations, punctuation, line breaks, and other features that are used.

Once a postal address has been recognized and parsed, it must be geocoded into a coordinate system such as latitude and longitude. This is not always easy, since a mailing address within a large organization may not pinpoint the location with high accuracy. For example, the Stanford University campus spreads over several miles, and individual point locations within the campus are handled by the local mail delivery system.

The postal delivery services of more advanced countries typically maintain databases of all deliverable addresses, and attempt to canonicalize the address written on a piece of

mail to associate it with a unique identifier for local delivery. Some countries also maintain databases that specify coordinate information, and some are available for purchase. In the United States, the Postal Service distributes a product called Tiger/Zip+4 containing approximately 35 million records, and specifying a street address range for each deliverable 9-digit postal code. Combining this with the Census department's TIGER (Topologically Integrated and Geographic Encoding and Reference System) dataset, it is possible to recover a latitude and longitude for a point associated with each unique deliverable address in the United States. These data sets are repackaged and sold by several commercial data vendors. Unfortunately, nine-digit ZIP+4 codes are not used universally in the United States, and the older five digit ZIP codes are much more common.

Note that the previously mentioned data sets provide only point locations. The US Postal Service has development underway on a database of polygonal boundaries called ZIP Code Tabulation Areas (ZCTAs). This will allow even better determination of the region associated with a postal code.

In the United Kingdom, the Ordnance Survey offers licensing on an Address Point database that references 25 million point locations. In addition, they offer a product called Code-Point that gives coordinates for the 1.6 million different postal codes within England, Scotland, and Wales. Each postal code contains an average of fifteen adjoining addresses. The resolution of this data is 1 meter.

No doubt such data sets are available for other countries as well, but there is no single source of such data. One major difficulty in using these data sources is the fact that each one has a non-standard format, and they must be reprocessed into a consistent format in order to be combined into an overall web mining system. Moreover, each database is dynamic, and a mechanism must be built to ingest updates. Just injesting updates is not enough however, because the web content may remain static, and make a reference to whatever was at the address at the time that the URL was created. If one takes these factors into account, then the task of maintaining a reliable source of point location information from addresses can be daunting.

In our prototype system, we used a small database of US postal codes and associated latitude/longitude values for the centroid of each region. US postal codes are either five digit numeric strings or nine digit numeric strings. An average five-digit zip code covers an area of about 120 square miles, but the median size is only about 40 square miles. The largest is about 27,000 square miles (zip code 99723 in northern Alaska). Not all of the US is covered by zip codes, and their boundaries are not well defined. We used a flex parser to recognize zip codes in pages, but in order to avoid too many false positives we required that the postal code be immediately preceded by a reference to the associated state. In other words, we did not accept 95120 as a legitimate zip code unless it was preceded by white space and one of the many representations for the state of California, including "Calif.", "Cal.", or "CA".

## 3.2 Telephone Numbers

Most telephone numbers have traditionally been segmented according to geography, in order to provide for efficient routing of traffic. There are numerous exceptions to this rule, such as 800 numbers in the US, which need not have any geographic significance. International telephone numbers typically begin with a country code, which allows a very coarse assignment of geographic context. More information about the structure of the number assignments within a region can provide more precise geographic context.

Recognizing telephone numbers within web pages is not completely foolproof, as they may be confused for various other numbers such as serial numbers or part numbers. Moreover, there are several standard ways of encoding a telephone number, depending on whether it is regarded as a local call or an international call. International telephone numbers are most often written according to the standard outlined by CCITT[3, 4], and begin with a +, followed by a country code, followed by a city code (in the US this corresponds to what is called an "area code"). These country codes and city codes are variable length, but since they are used for routing of calls, they are prefix-free and unambiguous. Thus by identifying a number within a page, we are able to determine the country and city of the number without difficulty.

Further complication for international calling arises from the fact that in some countries you need to dial an international access code (called an International Direct Dial code or IDD) in order to make an international call (e.g., in the U.S. it is 011). This is generally prepended to the international number, and some people will write this even though it is not part of the standard. We also encountered pages that contain phone dialing instructions from inside Centrex systems, where you need to dial an outside access code (e.g., a 9 or an 8) before dialing the full number.

Most telephone traffic is local rather than international, and therefore many numbers are written according to local customs rather than an international standard. Unless a geographic context can be inferred from some other source to expand the number into its international form, the construction of the full international number will be impossible to construct. In distinguishing a local vs. international number, we have some hints to help us. In some countries you need to dial a national direct dial code (NDD) in order to make a call from one city to another city still in the same country. In this case the country code is omitted, and a local access number is sometimes prepended, so that the local form is not necessarily a suffix of the international number. Thus an international number from the U.K. that would ordinarily be written as +44 12 345 6789 might be written as (0)12 345 6789 when referenced inside the UK. If the call is originated to this number from inside the same city, then you would only need to dial 12 345 6789. In order to represent both the international and intranational forms, some people have adopted a form like +44 (0)12 345 6789. Further complication arises from the nonstandard use of parentheses, spaces, slashes, and dashes within the number. All of these rules were incorporated into a flex parser.

In addition to recognizing a country and city code for a phone number, it is possible to infer even finer detail in some cases. Telephone numbers in the US are segmented by 3-digit "area code" (also called a Number Plan Area or NPA), and blocks of 10,000 numbers (called an exchange or NXX) within an NPA are then assigned to individual carriers. This assignment falls under the control of the North American Numbering Plan (NANP). Individual exchanges (NPA/NXX combinations) are generally confined to a rather small geographic region, and are thus a good

source of geographic context. We downloaded a database from the Internet[8] that provides latitude and longitude for over 100,000 NPA/NXX telephone prefixes in the USA. A more authoritative database is available under a license and subscription basis from the Traffic Routing Administration[20]. Their database provides vertical and horizontal coordinates (V&H) in a unique coordinate system that is used for calculating billing rates between different exchanges (see http://www.trainfo.com/products_services/tra/vhpage.html). With a little effort, conversion from V&H coordinates to latitude and longitude is possible.

Canada and the Carribean are included in the NANP, but Dykstra's database did not include the Caribbean, so we downloaded a separate list from `areacode.com`. For international numbers, we used the list of city and country codes from the Telecom archives [18]. This list is somewhat out of date, but gave us a list of approximately 24,000 city codes and country codes. We then attempted to match this list of cities against the NIMA database (see section [16]) of geographic place names, which returned the latitude and longitude of 16,500 of the cities. In this way we were able to construct a rudimentary database to prove the concept of phone number lookups, although clearly more thorough coverage would be possible with more effort and more authoritative data.

Going down to the level of individual numbers, databases exist that provide a physical address for most telephone numbers. Such databases are generally proprietary and expensive due to their commercial value and the level of maintenance required to keep them up to date. By combining such databases with others such as the Tiger data set, it is possible to determine the geographic context from telephone numbers with fairly high precision. For non-US telephone numbers, such databases are probably also available, though finding them for individual jurisdictions would be tedious.

## 3.3 Geographic Feature Names

One rich source of geographic context is the recognition of proper names of geographic entities. We identified two very useful databases for this purpose. The first of these is the Geographic Names Information System (GNIS), available from the United States Geologic Survey. This data set provides a listing of approximately 2.5 million geographic entities along with their latitude and longitude. The listings cover entities in many categories, including lakes, schools, churches, parks, buildings, valleys, airports, populated places, mines, etc.

The second data set that we found is the Geographic Names System from the United States National Imagery and Mapping Agency's (NIMA, formerly the Defense Mapping Agency) database of international geographic feature names. This comprises a database of locations for approximately 3.5 million features worldwide whose names are approved by the U.S. Board on Geographic Names. The purpose of this database is to standardize the spelling of place names within US government publications, but as a result it provides a very up-to-date and accurate database of geographic place names and locations. Each month, approximately 20,000 of the items are updated.

Combining these two databases, we have a list of approximately six million worldwide geographic feature names along with their latitude and longitude. Recognizing such a large number of patterns within a web page requires a little attention to algorithms and software, since it is unacceptable to perform a separate check for each pattern. For small phrase lists, flex provides a mechanism for recognizing such patterns using a finite state machine. At this scale however, flex is inappropriate. Another approach is to use fgrep, which uses the Aho-Corasick algorithm[1]. The Aho-Corasick algorithm is perfectly suited to this application. If the combined length of the patterns is $M$, then once a precomputation of time $M$ has been performed, the time to recognize these patterns is $O(n)$, where $n$ is the length of the document to be searched.

When the name "Chicago" occurs in a web page, we might naturally assume that the web page has something to do with the city in Illinois, and make the assignment of geographic context. Such an assignment is clearly going to have some errors associated with it however, since the word "Chicago" has multiple meanings, including the name of a city in the US, a pop music group, and the internal project name used for Microsoft Windows'95 during its development. Thus when we recognize the word "Chicago", it is not always correct to assume a geographic context, and we should probably use corroboration of another factor before making this association. Even if it was a proper geographic assignment, it is interesting to note that the word "Chicago" appears in the name of over 243 entities outside the state of Illinois, spread across 27 U.S. states. No doubt many of these are obscure, but it illustrates the danger in using common words as indicators of geographic context.

## 3.4 Context derived from hyperlinks

Many pages will not themselves contain geographic context, yet it can be implicit from the context in which the page is ordinarily encountered. The other pages that are linked to or from a page may themselves have an explicit geospatial context associated with them. In this case we can make an implicit association to the target page, using an aggregation of contexts from linked pages. Such an assignment has much lower reliability than an explicit reference, but can still be valuable for indexing and navigation. Given that we are able to make an explicit determination of geospatial context for about 10% of all pages on the web using simple geocodes in the content, and since the average number of outlinks from a page is approximately ten, one might expect that there is a high probability that most pages are linked from a page containing an explicitly recognizable geospatial context. Many pages will have multiple geographic contexts inferred this way, and multiple contexts that are in agreement provide increased confidence in the inference.

## 3.5 Other sources

In addition to the methods described in the previous sections for assigning geographic context, there are a few that we did not attempt to exploit. For example, we could use a whitepages directory of people to identify names of people in webpages, and assign their address as a geographic context for the page. The difficulties in distinguishing people with the same name makes this approach fairly error-prone. The use of business directory databases and yellow pages directories to correlate businesses against their web presence also obviously offers some promise, although this is restricted to commerce related information discovery.

## 3.6 Geocoding of HTML by Authors

This project is an attempt to automatically geocode web pages after geoparsing them, and to explore ways that browsing and indexing can exploit this geographic context. If one assumes that the author of a document has the best sense of the geographic context of a document, it seems natural to ask that authors insert metadata into their documents that conveys this context directly.

Such a mechanism was proposed in a recent Internet Draft[6][1], in which a simple `META` tag was proposed to encode positional information. The proposed tags allow specification of a latitude, longitude, altitude, region (e.g., ISO country code), and place name. The existence of such a tag offers some advantage, since the author is likely to best understand the correct geospatial context for the content. Unfortunately, authors seldom have the correct tools to determine precise coordinates for their content. The proposal also fails to address several other needs, including mobile locations and the ability to encode polygonal boundaries (e.g., for a river or a county). Moreover, all metadata initiatives suffer from the chicken and egg problem of wishing that the existing web was retrofitted with metadata. We believe that our work demonstrates that it is already possible to construct fairly reliable geospatial metadata directly from the content without the need for rewriting the web, although clearly it would be better if the authors were to author documents in a way that made their geographic context explicit and readily parsable.

Note that an experimental service for such geocoding is available online at http://geotags.com. They also offer a rudimentary search engine with the ability to restrict search by spatial boundaries, but they only index pages that contain the specified META tag. The number of pages containing these tags is very small at present.

There exist several other efforts to specify data formats for geospatial metadata. One is contained in the Dublin Core[22] "Coverage" tag, which comprises an XML tag that specifies either a temporal converage, a spatial location, or jurisdiction for the resource. The spatial location may then be specified either as a place name, a point, a polygonal region, or a globally unique geocode (such as a nine-digit zipcode in the US). The recommendation recognizes the need for a controlled vocabulary for place names, and gives the example of the Getty Thesaurus of Geographic Names[10]. It also describes ways to encode polygonal regions in a coordinate system. Another attempt to define geographic metadata has been developed by the United States Federal Geographic Data Committee. Their focus is almost entirely on encoding geographic data sets, and there is little attention to formatting of metadata that would allow for easy parsing and resource discovery in web pages.

Interoperability and exchange of data between GIS systems has always proved difficult, and there are few standards to help in the area. The Open GIS Consortium promotes interoperability and standards for data format and exchange. They have proposed a rather elaborate mechanism for encoding of geographic information via the Geographic Markup Language (GML). Their proposal uses XML to encode geographic place names, and geometric information for geographic features. Due to the highly technical nature of the format, the author believe that it is ill suited to a document collection as diverse as the web.

[1]Internet Drafts are considered works in progress, and are not for reference. This draft expires in July, 2001.

## 4. DATABASE OPERATIONS

Once web pages are geoparsed and geocoded, we require an organizational structure that will allow us to exploit the data. As with any database, the design of the database is dictated by the kinds of queries that the database needs to answer. The three types of queries that are desired are:

1. given a URL, find the geospatial contexts associated with that URL.
2. given a point location and an integer $k$, return a prioritized list of $k$ URLs, ranked by the distance of their geographic context from the point.
3. given the specification of a geographic region, return a list of URLs that have a geographic context that either overlap or lie completely within the region.

The first query is very simple to perform - we simply require a suitable database indexed by URL. The second query is somewhat more complicated, and requires the ability to find the $k$ nearest neighbors in the database. Algorithms for this problem typically require $O(k + \log n)$ and space $O(n \log n)$ for data sets of size $n$. The third query requires a range query, for which fairly efficient algorithms are known, particularly if the region is confined to rectangular regions with sides aliged along lines of latitude and longitude. In this case there are simply coded and efficient algorithms that take time $O(A + \log^2 n)$ time and $O(n \log n)$ space, where $A$ is the size of the output. Luckily for graphical display in a spatial browser this is the most natural query.

In our prototype system we made several choices that greatly simplify the design of the database. One of these was to confine ourselves to a relatively small list of locations (approximately 150,000). By collapsing all geographic contexts to this small number of actual locations, we are able to greatly reduce the complexity of the database. Another choice we made was to use simple hashtable lookups, and build query logic on top of this. Thus we designed the database using only two lists. The first list stores the location IDs associated with a given URL. The second list stores the URLs associated with a given location ID. On top of this we built software to support range queries and $k$ nearest neighbor searches.

## 5. SPATIAL WEB BROWSING

The major motivation for this project was to explore the opportunities for navigation of web resources by their spatial context. Hypertext has the unique property that information need not be processed linearly, but individual thoughts within a document may be followed by traversing a hyperlink. In some information discovery tasks, the thought that needs to be followed is spatial in nature. For example, when making travel plans for a conference, many people look for unusual activities that are available in close proximity to their destination. Traditional travel guides offer the opportunity to find such opportunities, but they are often many clicks away via traditional hyperlink navigation. Moreover, travel portals are often out of date or incomplete due to the need for human editing. We hypothesize that when a user wishes to see web resources that have similar geospatial context, the natural thing to do is to design an environment in

which the user can browse by geographic proximity rather than by link traversal. We designed a system that maps web resources to a map, and allowed the user to select documents for browsing according to their geographic location. Web resources appear as icons overlayed on this map. In this way a user can use mouse selections to refine their navigation, selecting portions of the map to concentrate on, or by clicking on the icons of the map to display the pages that discuss the location they are displayed at.

In the design of graphic user interface tools, screen real estate is very precious. We do not seek to replace the traditional hypertext mechanism, but rather to augment it with a spatial element. We do not see spatial browsing as a primary means of following ideas, but rather as a complement to standard hypertext browsing and keyword searches. In order to minimize the number of actions that a user must perform in order to switch into spatial browsing mode, we conceived of the idea of a "what's nearby" button that functions much the same way that the "what's related" button functions in the Netscape and Microsoft Internet Explorer browsers that use the Alexa service[14]. By clicking on this button, the user is presented with a view of web resources whose geographic context is in close proximity to the page that is currently being displayed. The process of geoparsing and geocoding web pages and depositing them into a database makes this an easy task.

There are several obvious choices for the design of a navigational tool based on geospatially related web resources. One is to simply present a separate window listing hyperlinks. This has the advantage of being very simple and using little screen real estate, but provides no sense of the geographic relation to the original page. A more compelling design is to provide a truly geographic navigational system, using a map or sattelite imagery. We implemented a prototype system based on these principles. The system uses a Java applet to trigger display of spatial information associated with the currently viewed page. The applet displays a single button marked "Where" in a separate window. When this button is clicked, the applet reads the URL of the currently displayed window, and opens a browser window to request a page from a server equipped with the aforementioned database. The URL loaded from the server causes a search on the database for the locations associated with the URL, as well as locations of any nearby points. The coordinates of these locations are used to create a rectangular region containing these points. The server then responds to the browser with a page containing a reference to an image map referencing a map from a map server, and overlaying menu information over the map that allows the user to inspect and navigate to URLs at the locations displayed on the map. Locations corresponding to the URL being inspected are color coded as blue dots, and nearby locations with other URLs associated with them are color coded as green. Each dot represents a "hot spot" so that when a user clicks on the dot, a popup menu appears showing a truncated list of URLs associated with that site, and an option to display them all in a list. In addition, the user may zoom in or out. Each zoom operation involves an HTTP request to the database server.

The map server that we use for our demonstration project is run by the US Department of Census, and only provides coverage for the US. The maps are generated from public domain data in the TIGER database, and are sent as GIF raster images. Other map servers could easily be substituted for this, but the only web map server with worldwide coverage known to the author is the Xerox PARC server, which provides very low detail in maps.

A screen shot of the system is shown in Figures 2 and 3.

## 5.1 Precision and Reliability

Our databases of zipcodes and phone numbers allowed us to segment all web pages into about 150,000 point locations, and the navigational tool would clearly benefit from better spatial resolution on our geocodes (e.g., using nine digit zip codes, full addresses, and individual phone numbers). If we apply our current database to the entire web, the approximately 100 million geocoded web pages would fall into 150,000 different buckets, which would place an average of about 700 web pages per location. In order to present a usable interface, we would require a more precise resolution of coordinates. Alternatively, we can provide a second level mechanism for navigating such a large number of documents. One approach is to automatically categorize documents into a coarse ontology, and another is to provide an inverted keyword search on that location. In future work we plan to extend the database in this direction. Even with this limitation, we feel that the existing system is good enough to provide valuable insight into the geospatial separation of topics in web pages.

It is clear from the discussion in section 3 that there will be false positives and negatives in recognizing geographic context within web pages. We have not pursued methodology to measure the accuracy rate, but experience suggests that our few simple heuristics can had a huge impact on the reliability of geocoded web data.

## 5.2 Future developments

We were able to demonstrate proof of concept for the utility of spatial browsing of web data, but we also discovered numerous roadblocks to the development of a globally available system of this type. Map data of sufficient quality is freely available for only a small fraction of the earth's surface. Most map data is available only in image format, which consumes a great deal of bandwidth in the transmission of maps to the browser. The Scalable Vector Graphics offers great hope for the transmission and storage of cartographic data in the future, since zooming in or out would no longer require the retransmission of an image for the new map. Unfortunately, almost no map data is available in this relatively new format.

Full scale deployment of such a system for a large number of users appears to be constrained only by the ability to generate maps. The database operations are relatively simple, and the total amount of data is relatively small compared to that required for an inverted keyword index.

The Open GIS Consortium has made a proposal for open access to web-based map servers, allowing the browser to request an overlay of map data. Since a web browser is unlikely to have all of the spatial information on their desktop to facilitate spatial browsing, such a network protocol will be helpful in the further development of the field.

We have chosen to display positional representation of web resources over a map, but another option is to use rasterized imagery. We avoided this due to the limited availability of such imagery, and the difficulty in determining coordinates atop such imagery.

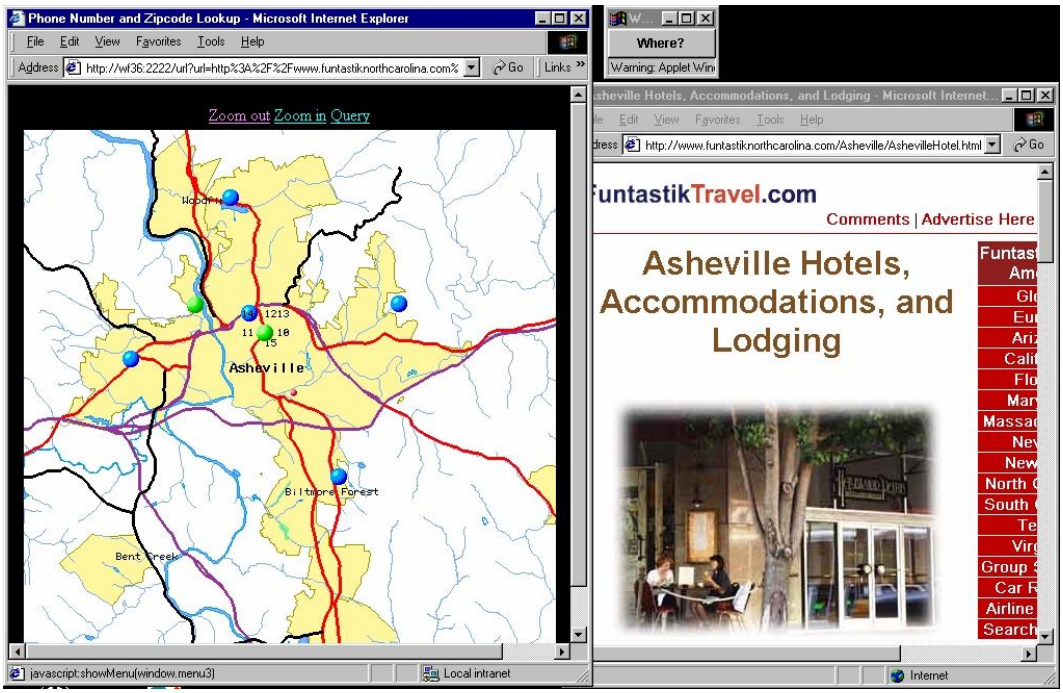


**Figure 2: A screen-shot of the spatial browser. The browser is showing a page about hotels in Asheville, North Carolina. By clicking on the "where" button, the browser opens a new window displaying a map of the geographic contexts represented within the page (in this case, the city of Asheville). Each green or blue dot represents a potential web site on the map, with blue used for showing locations associated with the page itself, and green used for other pages.**

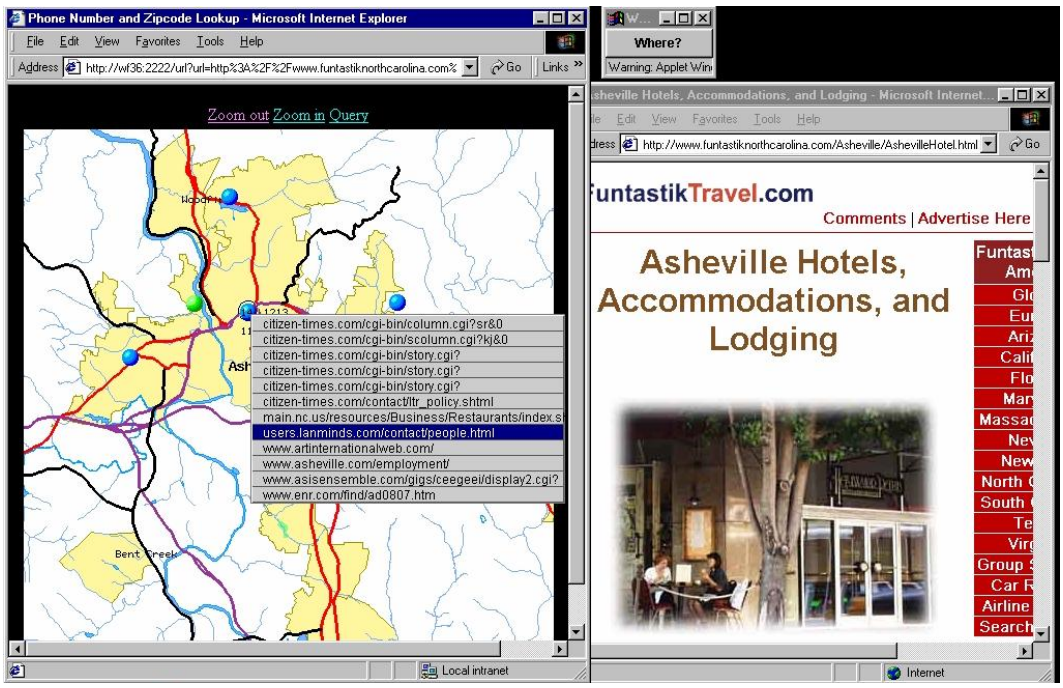


**Figure 3: By clicking on one of the dots displayed on the map, a menu is presented of the list of URLs associated with that location. Selection of one of the menu items will navigate the original window to this location. In the event that there are too many URLs to be displayed, a truncated list is given, along with an option to display the entire list in a new window.**

There are clearly many opportunites for a more sophisticated user interface, allowing selection of a portion of the displayed map, control over the resolution, the use of imagery, overlays with other spatially correlated data, and better integration into one of the existing browsers (eliminating the need for the free-floating button).

## 6. CONCLUSIONS

We have demonstrated that discovery and exploitation of geographic information in web pages is quite feasible, and exploitation of such information provides a useful new paradigm for the navigation and retrieval of web information. In spite of the many limitations for our prototype system, we have found it to be a useful adjunct to the existing methods of navigation and indexing, and we expect that the future development of roaming web users connected through wireless wide area networks will increase the demand for spatial information about the web.

## Acknowledgement

The author would like to thank Linda Hill of the University of California Santa Barbara for helpful discussions on geographic name databases, and Sridhar Rajagopalan for help with flex.

## 7. REFERENCES

[1] A. V. Aho and M. Corasick, "Efficient String Matching: An Aid to Bibliographic Search," Communications of the ACM, 18, No. 6 (June 1975), 333-340.

[2] Orkut Buyukokkten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, Narayanan Shivakumar, "Exploiting geographical location information of web pages." In Proceedings of Workshop on Web Databases (WebDB'99) held in conjunction with ACM SIGMOD'99, June 1999.

[3] CCITT/ITU-T Recommendation E.123: Telephone Network and ISDN Operation, Numbering, Routing and Mobile Service: Notation for National and International Telephone Numbers. 1993.

[4] CCITT/ITU-T Recommendation E.164/I.331 (05/97): The International Public Telecommunication Numbering Plan. 1997.

[5] Soumen Chakrabarti, David A. Gibson, and Kevin S. McCurley, "Surfing the Web Backwards", Proceedings of the 8th International World Wide Web Conference, Elsevier, (1999) pp. 601-615.

[6] Andrew Daviel, "Geographic registration of HTML documents", Internet Draft draft-daviel-html-geo-tag-04.txt, work in progress (December 2000, Expires July 2001). Available online at http://geotags.com/geo/draft-daviel-html-geo-tag-pre04.html.

[7] C. Davis, P. Vixie, T. Goodwin, I. Dickinson, A Means for Expressing Location Information in the Domain Name System, RFC 1876, The Internet Society, January, 1996. Available online at http://www.ietf.org/rfc/rfc1876.txt.

[8] P. Dykstra, NPA/NXX database. Available online at http://sd.wareonearth.com/~phil/npanxx/.

[9] C. Farrell, M. Schulze, S. Pleitner, and D. Baldoni, "DNS Encoding of Geographical Location", RFC 1712, The Internet Society, November, 1994. Available online at http://www.ietf.org/rfc/rfc1712.txt

[10] Getty Thesaurus of Geographic Names, Available at http://shiva.pub.getty.edu/tgn_browser

[11] T. Imielinski and J. Navas, "GPS-Based Addressing and Routing", The Internet Society, November 1996. Available online

[12] Internet Domain Survey, Internet Software Consortium. Available online at http://www.isc.org/ds/.

[13] The IP2LL Perl package. Available online at http://www-pablo.cs.uiuc.edu/Project/VR/ip2ll/faq.htm.

[14] . B. Kahle and B. Gilliat. Alexa - navigate the Web smarter, faster, easier. Technical report, Alexa Internet, Presidio of San Francisco, CA., 1998. See also http://www.alexa.com/

[15] Ray R. Larson, "Geographic Information Retrieval and Spatial Browsing" In: GIS and Libraries: Patrons, Maps and Spatial Information, Linda Smith and Myke Gluck, Eds., University of Illinois, (1996), 81-124. Available online at http://sherlock.berkeley.edu/geo_ir/PART1.html

[16] National Imagery and Mapping Agency GeoNet names server, available online at http://164.214.2.59/gns/html/.

[17] Ram Periakaruppan and Evi Nemeth, "GTrace - A Graphical Traceroute Tool," 13th Systems Administration Conference - LISA '99, November 7-12, 1999, Seattle, Washington, USA.

[18] The Telecom archives. Available online at http://hyperarchive.lcs.mit.edu/telecom-archives/archives/areacodes/.

[19] SRI Proposal for a `.geo` top level domain. Available online at http://www.dotgeo.org

[20] Traffic Routing Administration, Available online at http://www.trainfo.com/tra/catalog.htm.

[21] A. Vaha-Sipila, "URLs for Telephone Calls", RFC 2806, The Internet Society, April 2000. Available online at http://www.ietf.org/rfc/rfc2806.txt.

[22] S. Weibel, J. Kunze, C. Lagoze, M. Wolf, "Dublin Core Metadata for Resource Discovery". RFC 2413, The Internet Society, September 1998. Available online at http://www.ietf.org/rfc/rfc2413.txt.

Kevin S. McCurley joined IBM Almaden Research Center in 1997. He received a Ph.D. in Mathematics from the University of Illinois, and has previously held positions at Michigan State University, University of Southern California, and Sandia National Laboratories. His current research interests include information security and web technologies.