ITR/IM+SOC+HCI GEOWEB
PROJECT DESCRIPTION

1. PROLOG

The Web is like the jungle: lively, dense, evolutionary, encroaching, disorienting. Entering it to reach a specific destination, retrieve (harvest or mine) information, or for sheer intellectual stimulation, it is easy to get distracted, diverted or lost. In normal exploratory activity, our primary guideposts are physical location. Much of our knowledge of the world is already tied to place and time. We therefore propose to investigate geo-indexing the Web and the extent to which this can increase its usefulness and appeal.

Political and sports events exhibit regional affiliations. National, provincial and municipal government offices have clear geographic jurisdictions and constituencies. Forestry conservation has a sylvan focus, hydrology constrains canoe sales, and railroad schedules address distinct linear and point features. We intend to make these geographic ties more explicit using text-analysis techniques in conjunction with gazetteers, directories, and other searchable repositories of spatial information. For our specific purpose, natural language understanding will be complemented by geographic analysis.

Every real object and event - as opposed to abstractions like philosophy, chemistry and mathematics - has a well defined physical location and extent. Even abstractions, emotions and creative works are usually associated with some person who has a "residence" or "office". Creative works are often tied directly to a place ("The Dubliners", "Lilies, Giverny", "Prague Symphony"). We believe that most pages on the Web can be tied to geographic coordinates and mapped like any other artifact. Even though the Web purports to erase all geographic boundaries, it must perforce remain tied to activities taking place on the surface of the earth.

Other essential geographic concepts are scale and scope. There is already evidence indicating that most Web interactions, like telephone calls, are local. Aside from obvious items like server location, scope may also be indicated by the spatial pattern of links and accesses. We expect, however, that most of the necessary indexing information must be located by geo-specialized text-analysis of Web pages and browser/server logs.

Current approaches attempt to map the Web into conceptual or cognitive spaces. Such spaces are, however, intrinsically subjective. We believe that an objective scheme based on geographic attributes would augment ease of conceptualization, navigation, and retrieval of useful information. Much of the information about the current state of the world is becoming available first on the Web. Therefore any research that can benefit from geo-statistical data (demography, socio-geography) requires geo-indexed access. Prompt and comprehensive automated indexing will contribute to the emerging disciplines of web-anthropology, web-criminology, web-art, as well as electronic commerce. Because of the encompassing nature of the problem, the proposed research provides an exciting opportunity to interest students in some of the ethical and epistemological issues raised by the Web paradigm.

Geo-indexing augments rather than replaces traditional subject indexing. Both geographical and conceptual indexing are of a hierarchical nature: the intersection of orthogonal search planes should greatly improve the precision of retrieval for many queries. Because we expect that geospatial indexing will eventually become commonplace, we propose to conduct fundamental research on tracking the amount and nature of incidental geographic information on the Web, on methods of distilling it, and on combining it with traditional taxonomies for improving scientific, educational, business, and personal access.

Although we are targeting only spatial information, temporal context is also important. Some of the proposed techniques for spatial structure extraction, codification and utilization should apply to chronological focus and duration.

2. SPECIFIC RESEARCH GOALS

We need to accomplish there are five related tasks: collection of sample Web pages, design of a geographically-oriented taxonomy for spatial information, automation of geographic context extraction, development of a geographic index, and construction of a prototype server/browser to combine subject and location-constrained queries.

2.1 Geographically-unbiased sampling.

Most Web sampling methods are intrinsically biased geographically. The pages viewed by the members of any community are likely be biased by the local concentration of interest. Topic-oriented sampling brings into play exactly the type of correlation with geography that we wish to measure independently. Collecting samples by expanding arbitrary links enriches certain geographic contexts at the expense of others. To ensure a sample with representative geographic characteristics, we must explore other sampling methods.

The accepted way to collect a "representative" sample is through pseudo-random probes into an index of the population of interest. In our case, we need access to a large collection of Web sites such as those maintained by the Registries. Then we must randomly extract one or a few pages from various levels of each file directory system (only the top level is registered).

In preliminary work, Nagy and a student are generating Universal Resource Locators in .com from a lexicon of English words. Although many URLs are based on acronyms, proper nouns, or phrases, considering that about 10 of the 15 million registered domains are in *.com*, it is hardly surprising that almost all of the URLs generated from lexicon words exist. Some show up in an unexpected form (e.g., the PhoneMe site from "phoneme").

We are using this sampling scheme only as an unbiased stop-gap until we gain access to the Registries. Currently we download each page that we reach, and visually categorize any geographic information that we find. The method is too slow to yield a statistically reliable cross-section, but it does provide some insight into the relative frequencies of different types of

geographical reference. We consider research on sound, content-oriented, Web census tools of intrinsic interest.

2.2 A geographic web taxonomy

In classifying geo-spatial connotations, it is helpful to distinguish between items buried in narrative descriptions or tables, possibly included only incidentally or serendipitously in material prepared for other purposes, and large formatted files of structured geographic data.

Structured data includes maps, airborne and satellite images, and tabulated observations (weather, agricultural practices, stream flow, demographics). Such data is usually posted by organizations with an avowed geographical responsibility. They may be branches of national or local governments, such as USGS, NIMA, NASA, the Department of Commerce, and land-grant universities with conservation and survey charters, or private corporations such as ESRI, ERDAS or Space Imaging. Such data is usually well indexed and can be searched and retrieved with standard software or with proprietary GIS packages. Furthermore, both access information (the URL of the provider) and the contents of the database would be known to systematic users of geospatial data. We are interested in this type of repository only to the extent that it can help to locate, map or display *incidental* data, or provide enrichment in the form of maps, geophotos, and other "local color".

Incidental geo-references appear at millions of Web sites all over the world. Individuals, representing either themselves or their organizations, post information expected to interest various constituencies. The amount of duplication is staggering both in terms of actual files and links. So is the variety of content. We intend to develop systematic means to gain access to this wealth of amorphous location information, and organize it into a useful geographic framework.

Examples of the type of information to be found are the following:

> *The postal address of the owner of the page.* Many pages list a street address or at least a postbox - post-office - zip-code combination. For some operations (hotels, resorts, retail outlets, maintenance facilities, museums), the location is of prime concern to prospective customers. Currently, search engines and chambers of commerce do provide some listings by area, but these are not obtained automatically and are seldom up to date or complete. For example, a motel just outside the town line may well be missed.
>
> *The postal addresses associated with in and out links.* Such links can be traced recursively to find suitable associations. We will determine the distribution of content-based physical distance associated with in and out links of various degree. If the association is significant, then it can be used to index pages without explicit geographic reference.
>
> *The location of individuals who access the target page.* Although likely to exhibit some proximity effects, this information is more difficult to determine, because the postal address from where a hit is generated may not be available. Some Internet providers, such

as universities and town libraries, serve primarily a local clientele, but others, like AOL, are global. We shall develop methods for classifying providers before we attempt to use access information.

Direct references to geographic locations. Examples of geographic or pseudo-geographic place names are Ithaca, West Point, Pentagon. Their physical location is not as easy to map as zip codes, but can often be determined from gazetteers. Unless the reference is very specific (*Ithaca, NY*), the list of candidates may have to be pared down with the help of information from other types of geo-context. If many of the links point to Cornell, or many of the accessors have a Cornell email address, then the subject is probably Ithaca, NY. But if there are pointers to pages in Greek, or to the ferry schedule to Cephalonia, then it must be the island in the Ionian Sea. Many place names are ambiguous: there are at least six towns named Paris in the United States. Paris could also refer to the abductor of Helen of Troy, or to the Personal Augmented Reality Immersive System at the University of Illinois.

Indirect references. A reference to a Knicks' home game could be traced to New York City by associating *home game* with competitive team sports, then consulting a list of team names and team affiliations. Similarly, if there is enough information to determine that *Desert Storm* appears in a military rather than a meteorological context (caps do help), then perhaps it could be traced to Iraq. Although these inferences may appear speculative, the restriction of the search to geographic attributes will greatly facilitate the analysis compared to wider-ranging natural language understanding.

The physical location of the server. In some instances, the location of the server may be an important clue to the geographic focus of the content of the page. Large corporations often use computers at headquarters or at an important branch as Web servers, and much of their discourse is geocentric. Universities also often have on-site servers, and many of their pages bear on a local context. None of these clues are perfectly reliable, but in combination they may provide dependable geographic context.

2.3 Automated geo-content extraction

As indicated above, the geographic focus and extent of a Web page must be inferred from information distributed throughout the page. Different types of information must be fused to obtain a comprehensive and consistent geographic view. The information may be complementary or contradictory. Furthermore, since there is no universal quality control on posted contents, there is also an accuracy issue: typographic errors, and accidental or deliberate misinformation. We can, at best, check only consistency.

Extraction of already-coded geographic metadata.

Documents posted on the Web will become increasingly self-describing with content-tagging conventions like XML (Extensible Markup Language). XML describes the class of data objects to guide the behavior of programs that process them. While we cannot depend on every author to

provide accurate geographical descriptors (especially when the spatial information is considered incidental), we intend to make use of XML-indexed pages to improve the extraction of geographic references from pages without such metadata. Specifically, Scott will use tagged data to train machine-learning algorithms to recognize geographic identifiers.

Extraction of explicit postal addresses.

Embley extracted and organized atomic attribute-value pairs from data-rich documents, such as advertisements, movie reviews, weather reports, travel information, sports summaries, financial statements, and obituaries. He applied a conceptual-modeling approach to extract and structure the data. The approach was based on an ontology - a conceptual model instance - that describes the data of interest, including relationships, lexical appearance, and context keywords.

By parsing the ontology, he was able to automatically produce a database scheme and recognizers for constants and keywords, and then invoke routines to recognize and extract data from unstructured documents and structure it according to the generated database scheme. His experiments showed that it is possible to achieve good recall and precision ratios for documents that are rich in recognizable constants and narrow in ontological breadth. He will extend these techniques to the atomic geo-attribute-value pairs, e.g., (Street, <street name>), (City, <city name>), (PostalCode, <postal code>. Tools are available for converting these to geographic coordinates. Excellent validation methods, based on agreement between address components, are available from the postal OCR community [Srihari 92].

Translation of geographic place names to lat-long form.

In many cases, the translation of place names to lat-long form involve a simple database lookup. The coordinates of geographic entities can be also recovered using Geographic Information Systems such as ArcInfo. Approximate values can be obtained from the grid coordinates of atlases and gazetteers. There are also several Web sites that offer such a facility. However, two sources of potential ambiguity exist. First, a single name can represent multiple locations. Resolution of these ambiguities is discussed in "Information Fusion" below. Second, multiple names may represent the same location, but some of these names might not exist in the database. Thus lists of synonyms will need to be developed (e.g. "Jack's Market" = "1425 N 13th St."). Certain contextual clues in the HTML document, link patterns, and access patterns can help automate this process.

Analysis of indirect references.

Potent natural language understanding techniques have been developed for specialized dialog systems (e.g., flight reservations and situation reports), automated abstracting in restricted contexts, information retrieval, and language translation. At first, we will rely more on simple co-occurrence patterns than on subtle semantic analysis. For the example mentioned earlier of the 2002 Winter Olympics, we expect that the frequent co-occurrence of "Salt Lake City" and "2002" in Web pages under the subject heading "Winter Olympics" will suffice to establish the geographic focus of the original page (that did not mention SLC}.

If a page contains the statement "I graduated from Ladue High School in St. Louis in 1985 and then moved to Boston.", then this increases the likelihood that other pages related to this one are geographically related to St. Louis, which can help resolve geographic ambiguities. But to utilize such information effectively, it is best to extract the keywords in the correct context, e.g. the proximity of a city or school name to words such as "born", "moved", "resided", "attended", etc. Development of templates or ontologies can help in this process (as discussed above), but given the large number of contexts, new approaches will be required. One effective method is using manually-labeled data sets to infer hidden Markov models (HMMs) for text extraction. The learned HMMs parse the content of the web pages and extract relevant keywords and their associations (contexts), along with a confidence value in the associations.

After extracting relevant information and its context from the web page, the information can be used to learn rules governing georelations among web pages (e.g. hierarchical relationships, equality relationships, etc.). So when analyzing a web page based on indirect references, one can parse the page with an HMM and apply the rules to find the web pages (with known geoclassification) that are most closely related.

Tracing of in and out links.

Distance on the Web is measured in "clicks". The systematic tracing, graph-theoretic analysis, and visual display of link configurations constitute the rapidly growing discipline of web cartography. Although these concepts do not bear directly on real geography, they provide a ready set of tools determine the geographical correlates of link topology. We expect a high degree of spatial correlation among both in and out links of geographically focused sites. By exploiting the principle of geospatial locality that seems to exist among web accesses (as mentioned earlier), we can combine knowledge of in and out links with that of geoclassification of a subset of the linked pages to infer geoclassification of the remainder of the pages.

Analysis of accesses.

Unlike the relatively static link topology, site access is a highly dynamic phenomenon. Web servers log all accesses to the pages they serve, including date and time of access, the domain the browser is running on, and the URL of the page that the browser had just visited previously. Such information, if made available to our system, would help resolve ambiguities by exploiting any known geoinformation about the pages previously visited and about the web browser's domain. When the identity of the browser is hidden from the server of the accessed page, statistical information is collected by means of indirect means like cookies. Among all our potential sources of information, this may be the weakest. On the other hand, hit frequency is frequently considered an indication of the importance of a site. Some adverse social consequences of this practice are discussed in [Introna&Nussenbaum 00].

Using multiple labels for the data

In machine learning, recent work has involved so-called multi-label learning, where an instance (data item) is classified as belonging in several classes simultaneously. For example, a web page at UNL can have a label of "UNL" as well as "Lincoln", "Nebraska", and "USA", which implies a hierarchy of georeferences for a particular page. This exciting new area in machine learning tries to infer rules for such multi-label classifications, and recent work has explored assigning confidence values to each label.

Combining the evidence (information fusion).

In simple terms, this step requires either choosing the most reliable information or combining the facts from different sources, weighted according to their estimated reliability. The main goal of fusion is to improve the quality of geo-index. This is accomplished by increasing the confidence/relevance measure for some terms, adding new terms, removing factious geographic terms, and resolving potential conflicts. Some example of fusion are given below:

> A "Waterloo" reference in a page may be inferred to be "Waterloo, Iowa" instead of "Waterloo, Ontario" if the server is in Iowa and/or all the links point to various sites in Iowa or many of the accesses to the page are from domains in Iowa.
>
> A Web page describing the assassination of Abraham Lincoln has no direct geographical association with the city of Lincoln and should not have Lincoln as a geo-index. This conclusion may be reached by analysis of direct and indirect references, the linkage pattern and other auxiliary geo information. Of course, if the Web page is maintained by the Department of History at the University of Nebraska-Lincoln, with many links to other UNL pages, then Lincoln may be considered as a geo-index.
>
> A geo-index for "Miami" may be added to a page with no direct reference to Miami (or Florida) if it has a large number of links to other pages which are either located in Miami, have Miami as a geo-index, or if most of the accesses are from Miami. If the access and linkage pattern is spread through out the state of Florida, then it, instead of Miami, should be added as a geo-index term.

It is useful to assign relevance measures to the geo-index terms in a page. We must associate beliefs with each assertion and devise an approach to combine them. Many systematic approaches have been proposed in AI literature, including Fuzzy Logic, Bayesian Certainty Factors, Bayesian Networks, Dempster-Shafer Theory, Stanford Certainty Theory, Odds Propagation. Samal will compare these approaches and others in our application.

2.4 Geo-indexing

Geographic context must be codified at least by type, location, and extent. Earth-centered queries (we resist the temptation to index the entire Universe) should be based on the well-established geodesic (lat-long) coordinates. We expect further to retain the customary GIS distinction

between point objects (a mountain peak or a cinema), line/curve objects (roads and streams), and areal objects (characterized by a closed curvilinear boundary).

The *type* specification allows deriving many different relations between objects. Examples are distance and orientation of point pairs, geometric relations (e.g., parallelism) between curves, and topological or set-theoretic relations (inclusion, union) between areas. We intend to draw heavily on related geo-indexing projects initiated by the National Center on Geographic Information and Analysis (NCGIA). We will also consider the linear constraints proposed in [Revesz 98] to describe points, polylines and regions uniformly. Buffer zones about points or lines may be used to describe their extent or to take into account uncertainty [Goodchild 97].

To begin with, we will implement a simple index of the form
(type <t>, center <x,y>, extent <r>), with *type* as point / line / area, *x* and *y* as lat-long coordinates, and *r* in meters. We will extend this to a hierarchical scheme, where each page will have a top-level location and extent with subsidiary foci. This will allow describing a dealership with multiple branches, a municipal school system, or a river with tributaries.

Queries may also require combined information from several pages. Consider (a) a match, contested by sports teams from different towns; (b) factors that influence a political or a military campaign; (c) alternative production and transportation facilities; (d) the distance of a prospective dwelling from a workplace and shopping areas. The pertinent geographic information can be readily derived from our index by current database algorithms.

The codification itself will be performed by worms and crawlers ("agents") as in the current search engines. Depending on the pace of progress, we will either adapt XML standards for geographic indexing, or participate in the development of such standards. The technology is already well established and we can adapt existing directory structures. Our schema will, however, be more similar to that of a multi-scale cartographic archive or a GIS than to a library subject catalog. In contrast to the current topical directories, we believe that at least the geographical focus and extent of each page can be codified in an objective manner rooted in geodesy.

A substantial number of pageviews are hits on dynamic Web pages whose content is generated in response to a user query. It is not yet clear to us whether we need to capture such pages, or whether a static page-centered indexing scheme will suffice.

Since we eventually need to combine geographic focus with standard subject categories, the efficient intersection of results from independent search engines is an issue. Another important consideration is the "grain" of indexing: to what level of geographical detail should the content of each page be classified? The trade-off between depth and breadth of search, and stored or built-on-the-fly multiple views to speed-up search, have already been addressed in many successful search engines. Data structures for spatial (e.g. range) queries were developed in computational geometry and applied to GIS. We will reexamine these models in our application.

2.5 Prototype geographic server and browser

The construction of a prototype (Java?) geo-server is perhaps the least speculative part of this research, since several other groups are already combining GIS query facilities with Web browsers. There are also operational search engines available for a fee. We will probably have to write only a user-friendly front-end that facilitates geo-browsing.

Our server will simply constrain Boolean queries according to spatial context. The retrieved information will be combined, when appropriate, with GIS displays of structured data. Hits will be ordered by a weighted combination of the search topic and the location constraint. Seth will examine some of the options:

(a) Preindex all Web pages (as do most current search engines), then decide the order of topical versus geospatial search according to the number of hits for each case. If the focus of the query is educational institutions in Red Cloud, Nebraska, then it is better to apply the spatial constraint first. On the other hand, if it is Tokay grape in California, then it is faster to begin with the Tokay.

(b) Geo-index the pages on the fly. As in the Clever project, we start with a cluster of pages, obtained by topical search, and expand it using to-from pointers. The resulting set is geo-indexed and analyzed in the geographic context of the query.

(c) Instead of intersecting the complete set of responses that satisfy the topical and spatial constraints separately, as in (a), start with a cluster that does, and follow links as in (b).

Some queries will be served best by a link to an existing map. Additional geographic information could be presented using some cartographic metaphor. Linking site maps is already common practice on the Web, but, like lists of local facilities, the links are seldom compiled automatically. Only pre-specified items are mapped. Here too our intent is to make the geo-context graphically available even if the author of the page did not provide it.

The geographic context will be provided to the browser either explicitly, or by a default option set for each sign-on. We would also like to experiment with intelligent agents to determine the user's intent to focus the search. For example, given the query "Springfield" (one of the most common city names in the US), the agent can determine which particular city was intended, or if the user was searching instead for references to the 1980s rock star, or to "The Simpsons".

3. RELEVANT WORK ELSEWHERE

Due to space limitations this review is limited to the following topics that are most closely related to the proposed activities: geospatial metadata and standards, information extraction from the web, web page classification via machine learning, and web-search methods. We refrain from reviewing the large body of relevant work in statistical pattern recognition, geographic information systems, spatial data structures and algorithms, and database management.

Metadata. Metadata describes the content, quality, condition, and other characteristics of data. Much effort is now being devoted to including digital geospatial metadata in content standards [CSDGM1, FGDC] and to linking the terms to the tags in USMARC, a well-established standard for representing and communicating bibliographic information in machine readable form [Alexandria Crosswalk]. As a result, the amount of geo-information in digital libraries and on the web is growing [Hill 96]. As in digital libraries, most significant changes in the web environment are likely to occur as its meta-information environment becomes richer with the improved use of tagging [Smith 96].

XML, while relatively new, is emerging as the primary vehicle for content tagging on the web [XML]. Already a very substantial body of literature exists on this textual vehicle of expressing structured data [Harold 99]. XML became popular so quickly because it is easy to index, it supports a wide variety of applications including geographic, it is compatible with SGML and HTML, and it is easy to translate. Resource Description Framework (RDF) is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web. RDF emphasizes facilities for automated processing of Web resources. RDF can have a class system much like any object-oriented programming and modeling system ([Brickly & Guha 99, Beckett]).

Information Extraction. Our work on information extraction differs fundamentally from the approach others have taken, because we provide a document-independent target description. The most common approach to information extraction from the Web has been through page-specific wrappers, written by hand [Chawathe 94, Atzeni 97b, Gupta 97], or with the aid of a tool kit [Sahuguet 99], hand-coded specialized grammars [Abiteboul 97], wrapper generators based on HTML and other formatting information [Ashish 97, Hammer 97], page grammars [Atzeni 97a], landmark grammars [Muslea 98] concept definition frames [Smith 97], or some form of supervised learning [Adelberg 98, Ashish 97, Doorenbos 97, Kushmerick 97, Soderland 97, Freitag 98, Craven 98]. A disadvantage of these wrapper-generation techniques is the work required to create the initial wrapper (a disadvantage we also share in the sense that we have to create a target description), and the rework required to update the wrapper when the source document changes (a disadvantage that we do not share).

The approach of [Smith 97] using *concept definition frame* and [Craven 98] using *an ontology describing classes and relations* are closest to our approach. Our notion of a *data frame* [Embley 80] is similar to a *concept definition frame*, but embodies a richer description of the data to be recognized and extracted, and our notion of an *ontology* is similar to *ontology* of [Craven 98], but goes much further in describing the application of interest. The work reported in [Brin 98] is also similar to ours in the sense that it is robust with respect to source document changes. The technique in [Brin 98], which extracts author/title pairs, requires very little supervision for the machine-learning approach it takes, and need not be altered either for new pages or when pages change. This approach, however, appears to be limited to small, tightly coupled application domains such as author/title pairs for which it was used.

Another approach that has been used for information extraction is natural language processing (NLP) [Lehnert 94, Cowie 96, Soderland 97] NLP approaches use part-of-speech tagging, semantic tagging, building relationships among phrasal and sentential elements, and producing a coherent framework for extracted information fragments. Our work does not attempt to understand the text in the NLP sense, nor does it depend upon sentential elements as the NLP approach does (which are often missing, particularly for Web pages of classified ads, and partially formatted data found, for example, in forms and census records).

Web Page Classification. It has been shown that content-based web page classification can be automated via machine learning approaches that try to infer classification rules from the data. However, we know of no such approaches to geo-based classification. Content-based classification has been extensively studied at CMU, where learning is applied to adaptive web crawlers, customized feature extraction, and classification (e.g. [McCallum et al. 99, Nigam et al. to appear]). Feature extraction via hidden Markov models [Rabiner 89] was used in [McCallum et al. 99], and we believe that a similar procedure can be used more generally for context extraction when analyzing indirect references. After extracting this (and other) information, numerous approaches exist for classification, including classic machine learning algorithms such as naive Bayes approaches [Lang 95] and the expectation maximization (EM) algorithm [Nigam et al. to appear]. Other learning approaches attempt to infer relationships between web pages [Craven et al. 98, Freitag 98], which is particularly important to us when analyzing indirect references and establishing a geographic hierarchy. Algorithms for inferring relationships are available (e.g. [Quinlan 90, Goldman et al 95]), but will require some adaptation to fit our needs.

Generally, machine learning algorithms require large amounts of labeled data to learn. However, manually labeling web pages can be a tedious process, and XML-annotated pages are relatively rare. Thus work has been done in learning using small sets of labeled data supplemented with large sets of unlabeled data [Nigam et al 98, Blum & Mitchell 98]. Generally, these approaches work by training a pair of classifiers, each on a different half of the labeled data. Then each of these labels some of the unlabeled data, which becomes more labeled training for the other classifier. Such co-training approaches work surprisingly well in practice, and have a theoretical basis. We shall investigate this model further, and improve and adapt it for geo-classification.

Finally, we note that simultaneously learning several levels of a geographic hierarchy for a page might be amenable to an approach in the multiple label-learning model (e.g. simultaneously learning that UNL's pages are in UNL, Lincoln, Nebraska, and the USA). This new notion of learning has undergone little research to date (e.g. [Schapire & Singer b]) but has shown promise in text classification [Schapire & Singer a].

Search Engine Methods. Conventional text searches of web pages are apt to miss a relevant document because of synonymy or bring back too many irrelevant pages because of polysemy. Human-built Semantic networks [WordNet 98] may solve the synonyms problem but can aggravate polysemy [Clever 99]. Therefore automated indexing methods are increasingly turning to non-textual information associated with web pages. The hyperlink structure of the web has been analyzed in two quite different ways by the Clever project at IBM [Chakrabarti el al. 99, Kleinberg 99] and by Google [Google] both of which may be adapted to geo-indexing, as

indicated in the last section. The basic notion behind Clever search is that some web pages serve as "authorities" (endorsed by a large number of other pages) and some serve as "hubs" (with links to many authorities). Clever search algorithm starts with a set of pages derived from the term search, augments with pages linked to them by to and from links, and through an iterative process identifies the "hubs" and "authorities" (a page could be both). Google also examines the hyperlink structure of the web pages but assigns each a static weight, independent of a query, indicating the weighted sum of scores of other locations pointing to it. The weights are related to the frequency with which a page would be visited in random traversals of the web.

We expect further rapid development in all of these areas and will remain alert to emerging ideas and techniques.

4. COLLABORATIVE ASPECTS

Each member of our research team has expertise and prior publications on some aspect of Web-based information retrieval. Four of us have conducted research on geographic information systems. We have been visiting each other regularly in the course of on-going collaborations and advisory panels and already have well-established network routines for information exchange, brain storming, decision making, and documentation. In various combinations, we have co-authored over thirty papers on related topics.

To integrate the students fully, however, we intend to hold a three-day retreat or mini-symposium 18 months into the project. This will give us a chance for intensive interaction in the course of presentations of work to date and final planning for completing the project. The timing will coincide with the students' time-frame for graduate school application. The meeting will be held in centrally-located Lincoln during a school break. Since 5 of the 12 us are in Lincoln, this will also minimize travel costs.

We partitioned the research tasks according to our interests and expertise. These assignments reflect, however, only leadership responsibilities. We intend to meddle extensively and to discuss all major decisions.

David Embley, Brigham Young University, Computer Science:
- Conceptual modeling of geographic ontologies;
- Text analysis.

Mukkai Krishnamoorthy, RPI, Computer Science:
- Utilization of Web registries;
- Design and implementation of indexing agents.

George Nagy, RPI, Electrical, Computer, and Systems Engineering:
- Taxonomy of geo-context;
- Web census tools;
- Evaluation of utility of geographic meta-data.

Ashok Samal, UN-Lincoln, Computer Science and Engineering:
Fusion of geographic attributes;
Ancillary databases;
Browser implementation.

Stephen Scott, UN-Lincoln, Computer Science and Engineering:
Machine learning for geographic attribute extraction.

Sharad Seth, UN-Lincoln, Computer Science and Engineering:
User model and browser interface;
Combination of geographic and subject searches.

Embley, Nagy and Seth have long-standing collaborations, based on months-long mutual visits and exchanges, with researchers in Italy (M. Ancona, C. Arcelli, L. Cordella, B. Falcidieno, G. Sanniti di Baja), France (F. Lebourgois), Germany (A. Dengel, J. Biskup, N. Fuhr, Sweden (G. Borgefors), Norway (E. Aas), India (J. Jacob, S.K. Nandy, R.A. Parekhji) and Japan (H. Fujisawa, T. Wakabayashi). We are maintaining these contacts and will initiate more formal collaborations on various aspects of our current undertaking. We also have good contacts with the US GIS research community.

5. EDUCATIONAL ASPECTS

We are requesting funds mainly to support 6 students, with considerable cost sharing. Because some of the required project tasks require little theoretical background, we will recruit third-year students at the beginning of the project. We hope that these students will stay with us for all three years and enter a graduate program at one of the partner institutions in their final year. By then they will have the skills necessary to participate in the more abstract and mathematical aspects of the project.

We already know that this project is attractive to students because of its focus on the Web, on geography, and on data mining. We therefore hope that we will be able to recruit at least three minority or women students. Although our team is all-male, we have three recent woman PhDs graduates, and one more in the pipeline. The interaction with so many faculty with different academic backgrounds, interests and styles will provide our students with an unusually broad, and we hope attractive, perspective on research and scholarship.

We will endeavor to refine this perspective through local seminars on some of the critical issues facing researchers today. Many of these issues can be traced to the enormous increase in the potential for instantaneous and unstructured interaction. Privacy, security, intellectual property rights and obligations, and changing research ethics reflect different facets of the same

phenomenon. We shall document our progress in examining these issues critically through a common Web page or news group. Most of us have already faced some of these issues with ACM and IEEE student groups, and some of us have taught Computer & Society courses.

6. SCHEDULE

Year 1: Collect two hundred randomly selected Web pages and index them manually. Design interactive programs to assist in the extraction of geographic information in each of the five major categories, and to translate them to a geographic index. Prepare the collection of a larger (several thousand) unbiased sample based on registry access. Develop a scaleable data structure for the index. Firm up or initiate liaisons with researchers working on related problems.

Year 2: Download the larger sample. Automate the extraction of XML tags, addresses, links, and direct references. Analyze the larger sample to estimate the reliability of extraction methods and extracted attributes. Investigate machine-learning techniques for training an indexer on XML-tagged training data. Investigate information-fusion mechanisms and develop prototype algorithms. Develop candidates for measures of evaluation. Design a framework for combining traditional and geographic index terms. Review progress and bottlenecks at our minisymposium.

Year 3: Implement in-situ analysis (worms or crawlers) for at least the above categories. Design and implement the prototype server/browser combination, using off-the-shelf components to the extent possible. Design a query set of fewer than 100 queries divided into four classes of applications and execute these queries. Analyze and evaluate the results, with particular attention to scaleability. Demonstrate the mapping of query results on suitable structured geographic databases (especially maps). Report our findings. Decide whether to seek additional government or other funding for the project.

7. RESULTS FROM PRIOR NSF GRANTS

PI: **Sharad Seth**
Number: EPS 9720643 NSF Cooperative Agreement
Amount: $3,000,000 (UNL Share: $1,000,000)
Period: February 1, 1998 – January 31, 2001
Title: Nebraska EPSCoR Cooperative Agreement

Summary: The grant funds are being used to improve the computing and networking infrastructure at the University of Nebraska at Lincoln and Omaha and at Creighton University. The UNL component with which the PI is associated established a mid-level high-performance computing facility to serve the needs of researchers in science and engineering areas. Also included in the grant are funds to improve the networking environment, technical and user support services, and the laboratory facilities for new faculty in computational sciences.

PI: **Stephen D. Scott**
Number: NSF Award CCR-9877080
Amount: $170,192 (plus $15,000 in matching funds from UNL's Center for Communication and Information Sciences)
Period: July 1, 1999-June 30, 2002
Title: Applying Learning Theory to Systems Problems

S. Goldman and S. Scott. Multi-instance learning of fuzzy geometric concepts. Technical report UNL-CSE-99-006, University of Nebraska, 1999. In preparation for submission to the Machine Learning Journal.

> Presents an algorithm that is applicable to pattern recognition and drug discovery. It takes data that has real-valued (fuzzy) labels in [0,1] (e.g. a description of a molecule and a label of how well it binds at a particular site) and infers a function to assign [0,1] labels to new data.

D. Chawla, L. Li, and S. Scott. Exploring applications of MCMC methods to ensemble pruning and learning DNF formulas. Submitted to the Seventeenth International Conference on Machine Learning (ICML '00), 2000.

> Uses Markov chain Monte Carlo (MCMC) methods as a procedure in our algorithms for learning DNF formulas given labeled examples, and for pruning an ensemble of learned classifiers, produced e.g. by a boosting or bagging algorithm.