

Geographically-Sensitive Link Analysis

Hyun Chul Lee Haifeng Liu Renée J. Miller
Department of Computer Science, University of Toronto, Toronto, Canada
{leehyun, hfliu, miller}@cs.toronto.edu

Abstract

Many web pages and resources are primarily relevant to certain geographic locations. For example, in many queries web pages on restaurants, hotels, or movie theaters are mostly relevant to those users who are in geographic proximity to these locations. Moreover, as the number of queries with a local component increases, searching for web pages which are relevant to geographic locations is becoming increasingly important. The performance of geographically-oriented search is greatly affected by how we use geographic information to rank web pages. In this paper, we study the issue of ranking web pages using geographically-sensitive link analysis algorithms. More precisely, we study the question of whether geographic information can improve search performance. We propose several geographically-sensitive link analysis algorithms which exploit the geographic linkage between pages. We empirically analyze the performance of our algorithms.

1 Introduction

In the current web, it is not uncommon to find web pages that are characterized by their geographic locality. For instance, web pages about local restaurants, or local car dealers providing information about their services, products and addresses are prevalent on the Web. On the user side, there is strong anecdotal evidence that a considerable number of search engine queries are geographically oriented.¹ Queries like “find the movie theaters closest to LAX” or “find the best pediatricians in Riverside” are examples of such queries. We refer to the problem of finding and retrieving web pages that match the subject in which the user is interested, and that are also relevant to a particular geographic area the user has specified as *geographically-oriented search*. Geographically-oriented search (or “local search”) has recently gained substantial attention from industry. Although there are several available commercial local search engines (e.g., Google Local), we were not able to find any public domain information on the ranking algorithms used by such local search engines. Furthermore, it is not even clear whether they are actually ranking web pages rather than businesses.

In this paper, we study the issue of ranking web pages using geographically-sensitive link analysis algorithms. Specifically, we propose a set of alternative geographically-oriented search algorithms that are based on incorporating geographic semantics in different ways into link analysis algorithms. We first

propose *location-dependent* geographically-sensitive link analysis algorithms. That is, we precompute multiple importance scores for each page relative to a given set of geographic locations. Then, these scores are combined at query time based on the intended location associated with a user’s query to produce the final ranking of pages. Next, we propose *location-independent* geographically-sensitive link analysis algorithms which are computationally less expensive approximations of the original ones. These approximations first compute offline a single geographic importance score, “independent” of any specific geographic location, for each page. Again, this score is used at query time based on a simple treatment of the intended location associated with a user’s query. All of our geographically-sensitive link analysis algorithms are based on the hypothesis that when the semantics of geographic locality are combined with the linkage relation between pages, the performance of geographically-oriented search is greatly enhanced. We empirically study our hypothesis through a set of experiments on real web data. Importantly, our techniques do not assume perfect knowledge of geographic scopes (something that will rarely be available), but rather they rely on a simple, but efficient technique for automatically computing geographic scope. Moreover, we assume that during query time, the geographical context associated with user’s query is directly available.

We assume that if a page contains a geographic entity (e.g., an address) as part of its content, then the page can be treated as a *geographically-aware page*. Furthermore, if a page points to a geographically-aware page or is pointed to by a geographically-aware page, then it can be viewed, to some degree, as a geographically-aware page as well. In traditional link analysis, the linkage relation between two pages is viewed as an endorsement of content. In a similar manner, a geographic entity shared by two pages can be viewed as an endorsement of geographically sensitive content. We model the presence of a geographic entity within the content of a page using a *geographic link*. Additionally, a geographic entity shared by two pages can be represented as a *co-citation link*. In summary, for geographically-oriented search, the quality of a page should be determined by (1) the quality of pages that point to the page and (2) the quality of geographic entities that are found within the content of page. Our main contributions are:

- We propose several alternative graph models capturing the semantics of the relations between geographic entities and pages in addition to the page-to-page relations. For each graph model, we propose a geographically-oriented link analysis algorithm. In particular, we present a new algo-

¹Yahoo estimates that 20-25% of all search engine queries have a local component. (See <http://searchenginewatch.com/searchday/article.php/3389591>)

rithm, *GeoLink*, that is based on a new model of geographic importance. GeoLink directly models the semantics of geographic locality embedded in pages. It distinguishes the role geographic entities play on different pages, in contrast to the other algorithms (GeoRank, an extension of Page Rank and GeoHits, an extension of the HITS algorithm).

- Our evaluation study shows the benefit of using the geographic-entity-to-page relation with the page-to-page relation in computing page reputations (as done by all three algorithms GeoRank, GeoHITS and GeoLink). In addition, the evaluation highlights the additional benefit of directly modelling and exploiting the semantics of geographic entities (as done in GeoLink). Our experimental results also show that the performance gap (that is, the difference in the rankings) between the location-dependent algorithms and their more efficient approximations is relatively small, suggesting that both types of algorithms are efficient in practice.

2 Geographic Entities

Note that the first step toward geographically-oriented search is the estimation of the geographical scopes of web pages. Several possible solutions have been proposed which use both the linkage relation and semantic information [1, 6, 9]. However, none of them are completely efficient for estimating geographical scopes of web pages. We use a simple yet efficient technique in which address information found within the page content is used as the basic unit for estimating geographical scope. By a *geographic entity* or simply *geo-entity*, we refer to the address description of a physical organization or entity². Our assumption is that the presence of an address within a page is a strong indication that the page is associated with the physical entity corresponding to that particular address. While there might be several representations for an address (and therefore a geo-entity) depending on the region and the country, in this paper, we restrict our attention to those geo-entities that correspond to standard US addresses for our experiments. All geo-entities considered in this paper are text sequences of the form “*Street Number, Street Address, City Name, State Name*”. We downloaded all common city name references whose population is greater than 20,000 from <http://www.city-data.com> obtaining a list of 4581 unique city name references. We constructed a database of all possible street names for each city from the freely available 2004 TIGER/Line files.³ We refer to this database as the *StName Database*.

Furthermore, we define a regular expression to represent patterns for geo-entity extraction. Let $D = [s, south, n, north, w, \dots]$ be the set of street directions; $ST = [street, st, drive, dr, road, \dots]$ be the set of street abbreviations; $States$ be the set of state names; and $Cities$ be the set of city names. Each set also contains all possible abbreviation forms. We define an extraction pattern for addresses using the following regular expression:

$$\begin{aligned} \sigma_{StNumber} &= [1-9]([0-9])^* \\ \sigma_{StName} &= [a-Z]([a-Z0-9\backslashs])^* \\ \sigma_{Street} &= \sigma_{StNumber}([D])?\sigma_{StName}[ST] \\ \sigma_{city} &= [Cities], \sigma_{state} = [States] \\ \sigma_{address} &= \sigma_{Street}\backslash\sigma_{city}\backslash\sigma_{state}? \end{aligned}$$

In the above regular expressions, \backslash represents the character “\” and $\backslash s$ represents any separator (e.g., “,” or white space). To extract a geo-entity, we match the content of a given page against the above regular expression, and once extracted, we validate whether the extracted geo-entity is correct or not using the StName Database (i.e., checking whether the extracted street name is in the database). When we extract city name references, we might have problems due to *aliasing* (when different names or abbreviations are used for the same city) and *ambiguity* (when the possible city name candidate can refer to city names in different states). We deal with these problems following the approach proposed by Ding et al. [9]. Based on the random sample of 500 extracted geo-entities, we manually assessed the performance of our gazetteer-based extractor and obtained 97% accuracy. Certainly, different or more sophisticated techniques (with higher recall) could be used to estimate geographic scopes. However, one of our goals is to see if even a simple (yet highly efficient) technique such as this is sufficient to reveal the semantics of geographic locality encoded in web pages.

3 Algorithms

A query’s dominant location (QDL) refers to the intended geographic location(s) associated with the given query. How to obtain it from a search query is already addressed in [15] or it can sometimes be obtained from user’s IP-address. Therefore, we simply assume that a QDL is available to our algorithms at query time. In our approach, we first precompute the importance scores of pages relative to geographic locations that are possibly relevant to QDLs. That is, we compute offline multiple importance scores for each page with a given set of geographic locations that are likely QDLs. Let Ω denote the set of given geographic locations used for such precomputation. There are many possible ways of defining Ω (e.g., street level, zip code level, city level, state level) and our algorithms can be run over any arbitrarily constructed Ω . In our work, however, we construct Ω at the city level due to the nature of experiments that we conducted. We define Ω as a set of all (disambiguated) city names that are used in geo-entities extracted from the dataset and used for ranking. At query time, multiple importance scores for each page are combined based on the actual QDL of the query to form a composite ranking score for pages matching the query keyword. Certainly, this score can be used in conjunction with other IR-based scoring schemes (e.g., TF/IDF) to produce a final rank for the resulting pages with respect to a geographically-oriented search query.

3.1 Location-Dependent Models

The first step toward our ranking is to generate a set of ranking scores for each $l \in \Omega$. For this purpose, we propose three types of geographically-sensitive link-analysis algorithms. The basic intuition for our algorithms is as follows. Given $l \in \Omega$ (e.g., Houston), we refer to a geo-entity that is relevant to location l as an “ l -geoentity”. Next, we augment the web page graph to contain a node for each l -geoentity. Moreover, the presence of an l -geoentity within the content of a page is viewed as the

²Conceptually a geographic entity can be any representation of a geographical scope associated with a given page. We can run our proposed algorithms with any type of geographic entity. However, in this paper, we only consider those geographic entities in their address form for the purpose of experiments.

³<http://www.census.gov/geo/www/tiger/tiger2004se/tgr2004se.html>

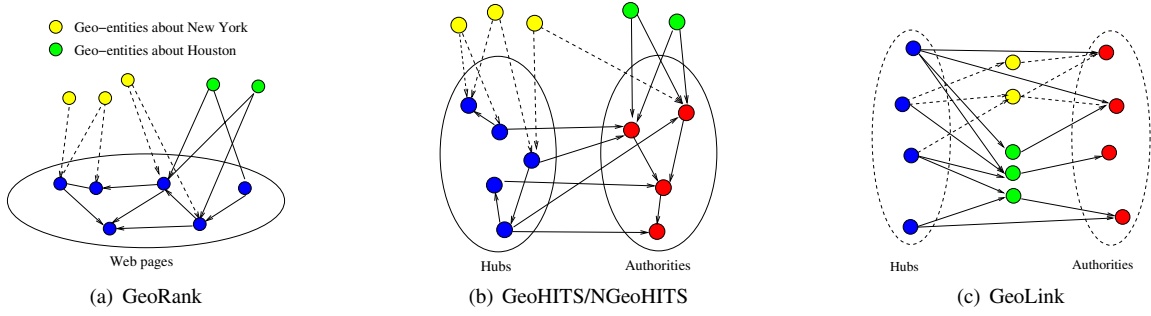


Figure 1. Graph Models (*Houston, TX*)

endorsement of authority from the corresponding l -geoentity to the page. Therefore, each l -geoentity-to-page relation is represented as a link between an l -geoentity and page. Consequently, two pages which are geographically similar with respect to l are *co-cited* by an l -geoentity. Therefore, the quality of a page with respect to the geographic location l depends on the quality of pages pointing to the page as well as the quality of l -geoentities pointing to the page. The quality of an l -geoentity, in turn, would depend on the quality of pages this geo-entity points to.

3.1.1 GeoRank

The GeoRank algorithm is similar to the original PageRank algorithm [5] in spirit. However, unlike PageRank which only considers page-to-page links, GeoRank performs a random walk over a graph with two types of edges (and nodes), namely page-to-page links and geo-entity-to-page links. The page-to-page relation is represented in the same way as in traditional link-analysis algorithms. That is, there is a link between two pages if there exists a hyperlink between these pages. Additionally, there is a link between an l -geoentity and a page if the page contains the l -geoentity in its content. Each random walk is performed with respect to those geo-entities relevant to l . In Figure 1(a), geo-entities relevant to *Houston* point to those pages containing addresses about Houston and these are used for the computation of GeoRank with respect to Houston. Each time the random surfer jumps uniformly at random to one of the page nodes with probability ϵ or decides to follow a “link” with probability $1 - \epsilon$. If the surfer decides to follow a “link” then it can either follow a page-to-page link with probability α or an l -geoentity-to-page link with probability $1 - \alpha$. In the former case, the random surfer follows uniformly at random one of the page-to-page links from the current page. In the latter case, the random surfer follows uniformly at random one of the l -geoentity-to-page links from the current page, and then it picks one of l -geoentity-to-page links from the chosen l -geoentity node to select the next page node. This defines a Markov chain on pages where the stationary distribution of this Markov chain is defined as the GeoRank page values for l .

Formally, for page j , let $F(j)$ denote the number of pages that are pointed to by page j and let $BG(j, l)$ denote the number of l -geoentities that point to page j . For l -geoentity k , let $FG(k, l)$ denote the number of pages that are pointed to by l -geoentity k then the GeoRank value of page i for l , $gr(i, l)$, is given as

$$gr(i, l) = \frac{\epsilon}{n} + (1 - \epsilon) \cdot \left(\alpha \sum_{\{j|j \rightarrow i\}} \frac{gr(j, l)}{F(j)} + (1 - \alpha) \sum_{\{k|k \Rightarrow i\}} \sum_{\{j|k \Rightarrow j, j \neq i\}} \frac{gr(j, l)}{BG(j, l)FG(k, l)} \right)$$

where $j \rightarrow i$ means that page j points to page i and $k \Rightarrow j$ means that l -geoentity k points to page j . The Markov chains of our Georank algorithm can be shown to be uniformly ergodic therefore convergent using the method of Breyer [4]. Therefore, the standard power-iteration method, used for the computation of principal eigenvectors, can also be applied to the computation of each GeoRank value.

3.1.2 GeoHITS and NGeoHITS

Unlike the PageRank algorithm, which assigns a single rank value to each page, the HITS algorithm assigns two rank values (authority and hub values) to each page. In the original HITS formulation, hub and authority values are calculated from a mutually reinforcing relationship where good hubs are those that point to good authorities, and good authorities are those that are pointed to by good hubs. We extend this mutually reinforcing relationship for relations among web pages and geographic entities. More precisely, page-to-page links and l -geoentity-to-page links are constructed following the GeoRank algorithm except that we distinguish hub pages from authority pages. In Figure 1(b), we illustrate an example of the graph model for the GeoHITS algorithm. Thus, for the GeoHITS algorithm, a good authority with respect to l is a page that is pointed to by both good hubs and good geographic entities relevant to l . Similarly, a good hub relevant to l is a page that points to good authorities relevant to l and is pointed to by good geographic entities relevant to l . Let A and H be the set of authorities and hubs, respectively. Let $H(j, l)$ and $A(i, l)$ denote the hub value of page j and the authority value of page i respectively. Let $G(k, l)$ denote the geographic rank value of l -geoentity k . Then, the GeoHITS algorithm calculates each hub/authority and geographic rank values for l using the following equations:

$$\begin{aligned} H(j, l) &= \alpha \sum_{\{i|j \rightarrow i\}} A(i, l) + (1 - \alpha) \sum_{\{k|k \Rightarrow j\}} G(k, l) \\ A(i, l) &= \alpha \sum_{\{j|j \rightarrow i\}} H(j, l) + (1 - \alpha) \sum_{\{k|k \Rightarrow i\}} G(k, l) \\ G(k, l) &= \beta \sum_{\substack{i \in A \\ k \Rightarrow i}} A(i, l) + (1 - \beta) \sum_{\substack{j \in H \\ k \Rightarrow j}} H(j, l) \end{aligned}$$

Bharat and Henzinger [3] point out that it is possible that the simple mutual reinforcing relationship approach of HITS may give undue weight to some pages. Therefore, a highly connected set of pages can dominate the results of the HITS algorithm. Note that similar phenomenon might occur in the geographically-oriented search context. In other words, it would be desirable for each l -geoentity/page to have the same influence on other pages to which it is connected. Thus, following [3], we modify

the GeoHITS algorithm by normalizing the number of forward links from a page or an l -geoentity using the number of pages to which it points. In a similar manner, all backward links are normalized. We refer to this modified GeoHITS algorithm as NGeoHITS. Similar iterative computation as that of the HITS algorithm can be done. Therefore, one can easily prove the convergence of GeoHITS and NGeoHITS algorithms in a similar manner as that of [3, 16].

3.1.3 GeoLink

Both GeoRank and GeoHITS exploit the existence of geographic entities in ranking web pages by using geographic content to endorse pages. However, they do not strictly exploit the semantics of geographic entities as they also consider those linkage relations that are not directly related to the quality of geographic entities. In the GeoLink algorithm, though, we take a stricter view by only considering the linkage structure that is directly derived from the semantics of geographic entities. In the GeoLink algorithm, we classify pages containing at least one l -geoentity into two categories, namely *l -strong geographic hubs* and *l -strong geographic authorities*. We say a page is a *l -strong geographic authority* if it contains exactly one unique l -geoentity in its content, and we say it is a *l -strong geographic hub* if it contains more than one unique l -geoentity in its content. As mentioned in Section 1, pages that point to or are pointed to by those pages with geo-entities embedded into their content can be viewed, in some degree, as geographically-aware pages. Therefore, we say that a page is a *l -weak geographic authority* if it is pointed to by some *l -strong geographic hub*, and we say that it is a *l -weak geographic hub* if it points to a *l -strong geographic authority*. An l -geographic authority is a page that is either an l -strong or l -weak geographic authority, and similarly, an l -geographic hub is a page that is either an l -strong and l -weak geographic hub. Intuitively, an l -geographic authority is a page which is possibly geographically-aware for a physical entity relevant to l . It contains the address information relevant to l for a single physical entity or it is pointed to by an l -strong geographic hub, which is associated with multiple physical entities relevant to l . Similarly, an l -geographic hub is a page which is possibly geographically-aware for multiple physical entities relevant to l , whose status is gained through multiple addresses relevant to l embedded in its content or by pointing to a strong geographic authority.

The graph model for the GeoLink algorithm is constructed using the set of l -geographic hubs and authorities as follows. The page-to-page graph is first constructed following the conventional web graph model. There is a link from page i to page j if there is a hyperlink from i to j . Next, only the edges (hyperlinks) from l -geographic hubs to l -geographic authorities are kept while the rest of the edges are discarded. For the l -geoentity-to-page relation, only forward links from the l -geographic hubs are allowed, and similarly only backward links from the l -geographic authorities are allowed. More specifically, there is a link from an l -geographic hub j to l -geoentity k , if j contains the l -geoentity k as part of its content along with other geographic entities. Similarly, there is a link from l -geoentity k to l -geographic authority i , if l -geoentity k is the only l -geoentity found in the content of page i . In Figure 1(c), we illustrate an example of the graph model for the GeoLink

algorithm.

Formally, let $F(i)$ denote the number of pages that are pointed to by page i , and let $B(j)$ denote the number of pages that point to page j . Let $FG(k, l)$ denote the number of pages that are pointed to by l -geoentity k and let $BG(k, l)$ denote the number of pages that point to l -geoentity k . Let $FS(i, l)$ denote the number of l -geoentities pointed to by page i and let $BS(i, l)$ denote the number of l -geoentities that point to page i . Let $H(j, l)$ be the hub value of page j for l , let $A(i, l)$ be the authority value of page i for l , and let $G(k, l)$ be the geographic rank value of l -geoentity k . Finally, let n denote the number of l -geographic hub pages, let m denote the number of l -geographic authority pages, and let s denote the number of l -geoentities. The GeoLink algorithm consists of the following self-consistent equations:

$$\begin{aligned} H(j, l) &= \frac{\epsilon}{n} + (1 - \epsilon)(\alpha \sum_{j \rightarrow i} \frac{A(i, l)}{B(i)} + (1 - \alpha) \sum_{k, j \rightarrow k} \frac{G(k, l)}{BG(k, l)}) \\ A(i, l) &= \frac{\epsilon}{m} + (1 - \epsilon)(\beta \sum_{j \rightarrow i} \frac{H(j)}{F(j)} + (1 - \beta) \sum_{k \rightarrow i} \frac{G(k, l)}{FG(k, l)}) \\ G(k, l) &= \frac{\epsilon}{s} + (1 - \epsilon)(\gamma \sum_{k \rightarrow i} \frac{A(i)}{BS(i, l)} + (1 - \gamma) \sum_{j \rightarrow k} \frac{H(j, l)}{FS(j, l)}) \end{aligned}$$

The computation of hub/authority values and geographic rank values can be performed in a similar manner to that of the previous algorithms. The convergence proof of the algorithm can be done using similar techniques to those used in LinkFusion [16].

4 Location-Independent Models

Our geographically-sensitive link analysis algorithms presented in the previous section require the computation of a page importance score for each $l \in \Omega$. Since the size of Ω might be very large,⁴ the computation cost for our algorithms might be expensive. In a dataset consisting of 10 million pages,⁵ we found 1083 different cities from the extracted geo-entities. We also found through our experiments (presented in the later section) that the convergence rate of our algorithms is slower than that of the traditional link-analysis algorithms. For instance, for the dataset “Hotel in Austin, TX” which consists of 9544 pages, 88478 links, 1743 l -geoentities, and 4421 geo-links. We found that both HITS and PageRank reach the desired residual ratio (≤ 0.001) after 18 and 13 iterations, respectively, while the geographically-sensitive link analysis algorithms require more than 30 iterations. Therefore, in what follows we propose a simple heuristic method for approximating our (location-dependent) geographically-sensitive link-analysis algorithms which eliminates the need for computing a page’s ranking with respect to each location (e.g., city). Without loss of generality, we only focus on the GeoLink algorithm since the approximation of others can be done in a similar manner. Our heuristic algorithms treat “all” geo-entities within a page as being relevant, something which will rarely be true. However, this simplification will allow us to compute a ranking more efficiently. In a page relevant to “Los Angeles”, our heuristic will assume that most geo-entities found within the page are relevant to “Los Angeles” (or equivalently that the presence of a few geo-entities about other

⁴Cities with population size greater than 20,000 well exceed few thousands.

⁵Using “Regional/North America: United States” from DMOZ (<http://www.dmoz.org>) as seed pages, we ran a small scale crawler for 5 days to download 10 million pages.

cities does not skew the ranking too much). Therefore, approximations of our original geographically-sensitive link analysis algorithms are *location-independent*. More precisely, the basic assumption for the approximation is that GeoLink values satisfy:

$$a(p, l) \propto gt(p, l) \cdot \tilde{a}(p), h(p, l) \propto gt(p, l) \cdot \tilde{h}(p)$$

where $\tilde{a}(p)$ and $\tilde{h}(p)$ refer to location-independent GeoLink values, and $gt(l, p)$ is defined as the ratio between the number of l -geentities found within page p and the total number of geo-entities found within the page.⁶ Location-independent GeoLink values are computed using the original GeoLink algorithm with the only exception being that the construction of page-to-page and geo-entity-to-page links is not limited to any particular $l \in \Omega$.

5 Query-Time Page Importance

Static rankings like those produced by link-analysis-based ranking algorithms are normally combined with the traditional IR-type content-based rankings to produce the final rank for the page. In this section, we describe one possible way of combining the rankings produced by our geographically-sensitive link-analysis algorithms with the content-based rankings to produce the final rank of a page given a geographically-sensitive query.

To compute the final rank of a page using our original **location-dependent** rankings, we do the following. We only describe this step for the GeoLink algorithm as the other algorithms are similar. Given a geographically-oriented search query q , let $k(q)$ be the set of keyword terms and let $l(q)$ be the query’s dominant location (QDL). For example, given the query q , “Hotel in Los Angeles”, $k(q)$ would be “Hotel” and $l(q)$ would be “Los Angeles”. Using a text index, we retrieve all pages containing the original keyword terms, $k(q)$. The GeoLink values of each page p with respect to $l(q)$ are computed as:

$$a(p) = \sum_{l \in \Omega} pr(l|l(q))a(p, l), h(p) = \sum_{l \in \Omega} pr(l|l(q))h(p, l)$$

where $pr(l|l(q))$ denotes the probability that the query’s dominant location, $l(q)$, is related to the geographic location l . Since Ω consists of cities, $pr(l|l(q))$ will express the relevancy of $l(q)$ to the city l . Since geographically-oriented search queries used in our experiments are at the city level, we will have $pr(l|l(q)) = 1$ if $l = l(q)$ and $Pr(l|l(q)) = 0$ otherwise. We also compute the content-based rank of each page with respect to $k(q)$ using some traditional IR-approaches. Let $tr(p, k(q))$ denote such a rank value of page p . Finally, we combine $tr(p, k(q))$ of page with that $a(p)$ or $h(p)$ of page as a linear sum of these values (following [7] and others) to produce the final rank of page.⁷ We can do the same computation using our **location-independent** rankings, in which case we will use our query time approximations for $a(p)$ and $h(p)$:

$$a(p) = \sum_{l \in \Omega} pr(l|l(q))gr(p, l)\tilde{a}(p), h(p) = \sum_{l \in \Omega} pr(l|l(q))gr(p, l)\tilde{h}(p)$$

⁶This assumption is empirically supported. We observe in the MSN dataset (see Section 6) that the average similarity between location-independent rankings and location-dependent rankings is 0.7445. The similarity value is computed using a variant of the Kendall’s measure [10] with $n=500$.

⁷We expect that more sophisticated use of content-based ranking would yield even better rankings.

6 Experiments

We present the results of experiments that we ran to determine the feasibility of our algorithms in practice. Due to the unavailability of a benchmark for testing our algorithms, we decided to construct our own dataset for testing. We first chose the following 11 sample keywords: *Day Care, Financial Service, Fitness, Health, Shopping, Seafood, Hotel, Italian Restaurant, Plumbing, Real Estate, School*. Then, we chose the following 7 locations: “Austin, TX”, “Chicago, IL”, “Houston, TX”, “Miami, FL”, “Los Angeles, CA”, “New York, NY”, “Tucson, AZ”. Each keyword and location were combined to build a query string (e.g., School in Austin). Next, each constructed query string was sent to the MSN search engine⁸ and the top 200 returned pages were retained as the Root Set. For each page in the Root Set, we included all pages that are pointed to by this page again using the MSN search engine and the first 300 pages (in the order returned by the MSN search engine) that point to the page. The total number of pages collected in this way was around 665,000. We call our dataset *MSN dataset*.⁹

6.1 User Study

As our first evaluation study, we compared our geographically-sensitive link analysis algorithms (both location-independent and dependent) against those traditional link analysis algorithms using randomly chosen 20 sample queries from the MSN dataset (e.g., “Hotel in New York”). For each query, both geographically-sensitive and traditional link analysis algorithms like HITS and PageRank were run (without any IR-ranking) to produce the final top 10 ranked pages. We conducted a user study using the precision over the top-10 (p@10) as the evaluation measure. More precisely, let the *high relevance* ratio be the fraction of pages within the top 10 results returned by each algorithm that are highly relevant to the given query. Let *relevant ratio* be the fraction of pages within the top 10 results that are relevant, or highly relevant to the given query. Similar measurements are used in the literature [14] and the TREC conferences. The relevance rating of results was obtained through a user study. The set of results produced by the algorithms were shuffled and then displayed to each user according to what query he/she selected for evaluation. Without any prior knowledge about which algorithm was used to produce the results, the user was asked to rate the returned web pages as either “unknown”, “non-relevant”, “relevant”, or “highly relevant” to the query. For each query, we take the average of high relevancy and relevancy ratios over all participants. The number of participants per query varied from 3 to 8 while the average number of participants per query was 4.192. Most of participants had a computer science background and extensive experience with web search. In Table 1, we present the average high relevance (HR) and relevance (R) ratios of all algorithms. In the table, the algorithm name followed by “-A” refers to the approximation version (location-independent) of the corresponding algorithm. One can observe from Table 1 that when the semantics of geographic entities are combined with link analysis of pages, the performance of geographically-oriented search is clearly improved. The performance of traditional link-analysis algorithms

⁸<http://search.msn.com>

⁹The dataset construction was performed in March 2006.

	Avg HR ratio	Avg R ratio
PageRank	1.04	3.64
HITS	2.05	4.9
GeoRank-A	2.82	5.26
GeoRank	2.39	5.7
GeoHITS-A	3.09	5.71
GeoHITS	3.2	6
NGeoHITS-A	2.87	5.32
NGeoHITS	2.22	5.19
GeoLink-A	3.52	6.13
GeoLink	3.66	6.65

Table 1. Comparison of our algorithms against traditional link analysis algorithms

like the PageRank and the HITS algorithms, on the other hand, is substantially worse than geographically-sensitive algorithms.

6.2 Link Based Ranking+IR-Ranking

We conducted an additional experimental study to assess the ranking quality of our algorithms when they are combined with traditional IR-ranking. We used 13 randomly chosen sample queries (“Seafood in Houston”) from the MSN dataset. We produced 2 types of rankings: (1) We combined rankings produced by our proposed algorithms with IR-ranking as follows. For each query, we first computed geographically-sensitive rankings (with respect to “Houston” in our example) using our algorithms. Then, we computed the content-based IR ranking of each page with respect to the query keyword (“Seafood” in our example). The cosine similarity and the TF/IDF weighting were used for IR-ranking. For every algorithm, the final rank of a page was a linear sum of geographically-sensitive ranking and content-based IR ranking as described in Section 6. (2) We wished to have a baseline ranking for comparison. For each sample query q (e.g., “Seafood in Houston”), we used PageRank and HITS unchanged to produce rankings, and then we filtered their results based on a simple notion of geographic relevance. Specifically, we removed those pages not containing any geo-entity relevant to the query’s dominant location. In this way, we were able to get rid of those pages that are not relevant to the given query’s dominant location. We computed a content-based ranking of each page with respect to the full set of query terms (“Seafood in Houston” in our example). The final rank of each page was produced in a similar manner as that of geographically-sensitive ranking case.

To evaluate the quality of the results returned by each algorithm, this time, we constructed a ground-truth set as follows. We first merged all top 10 pages returned by each algorithm and those from MSN into one single set. Without having any knowledge about which algorithm produced a page, we rate each page as either “non-relevant”, “relevant”, or “highly relevant” by carefully analyzing its content. We used the following basic criteria for this evaluation step: (1) A page is *highly relevant* if it contains information that are definitely relevant to the query term as well as to the query’s dominant location. (2) A page is *relevant* if it contains information that are related to, but not necessarily relevant to both the query term and the query domination location or it contains some sections that are relevant to both the query term and the query’s dominant location. (3) A page is *not relevant* if it is irrelevant to either the query term or the query’s dominant location. Using this ground-truth set, we assessed the quality of our rankings by comparing this set against the top 10 results returned by each algorithm. Once again, we used the pre-

	Avg HR ratio	Avg R ratio
PageRank-IR	1.65	3.65
HITS-IR	1.65	3.5
MSN	2.77	6.23
GeoRank-A	2.82	5.26
GeoRank-IR	4.73	7.15
GeoHITS-A	3.09	5.71
GeoHITS-IR	4.54	6.85
NGeoHITS-A	2.87	5.32
NGeoHITS-IR	4.77	6.85
GeoLink-A	3.52	6.13
GeoLink-IR	5.15	7.92

Table 2. Performance comparison of our link analysis algorithms against traditional link analysis algorithms and MSN (combined with IR-ranking)

cision over the top-10 ($p@10$) as the measure for the evaluation. In Table 2, we report the average high relevance and relevance ratios of our algorithms (with IR-ranking) against those traditional methods (with IR-ranking), and the original MSN results. Once again, one can observe from Table 2 that when the semantics of geographic entities are combined with a link analysis of pages, the performance of geographically-oriented search is clearly improved, reinforcing the conclusions of our first experimental study.

6.3 Discussion

Our algorithms that capture the relevant semantics of geographic locality in the web perform well for the given geographically-oriented queries. When the page is geographically-aware, the number of l -geentities found within its content and its linkage structure are valuable for determining what type of page it is. For instance, a page which contains reviews of local restaurants may contain several addresses as part of its content, and possibly links to homepages of local restaurants. On the other hand, there is a strong probability that the homepage of a local restaurant only contains one single address corresponding to the geographic location of the restaurant and may also be pointed to by a review page of local restaurants. Our GeoLink algorithm, which captures this subtle difference among geographically-aware pages by distinguishing l -strong/weak geographic authorities from l -strong/weak geographic hubs, emerges clearly as the best among all of our algorithms.

Furthermore, by comparing Tables 1 and 2, one can observe that the performance of our algorithms is strengthened when they are combined with IR-ranking. The performance of traditional link-analysis algorithms like PageRank and HITS, on the other hand, is still substantially worse, even when they are combined with IR-ranking, than geographically-sensitive algorithms. We believe that the main reason why classical link-based ranking methods failed was due to their inability to distinguish the influence of the query’s dominant location from the influence of the keyword search terms in computing the ranking. We have found in numerous query samples that the content-based ranking using the full set of query terms were biased due to the query’s dominant location term. For instance, in the query “Fitness in Chicago”, the top 5 pages returned by GeoLink algorithm were all highly relevant, while the top 5 pages returned by PageRank were all non-relevant but they had a high frequency of the term “Chicago” in their content as shown in Table 3.

Title	Webpage URL
Geolink Ranking	
Fitness Chicago	http://www.depofitness.com/fitnesschicago/
Fitness-Salon And Spa Beauty Fitness	http://fitness.researcheasy.com/salonandspabeautyfitness/
The Fitness Chicago resource!	http://www.alexfitness.com/fitnesschicago/
qovexfitness.com	http://www.qovexfitness.com/chicagofitness/
Chicago Crunch Fitness II	http://model.fitnesshub.info/chicago-crunch-fitness-il.html
PageRank	
Gapers Block:Author-James Allenspach	http://www.gapersblock.com/.author/jma/
Chicago - welcome to Bar Chicago	http://www.dchicago.com/barchicago/
Chicago - welcome to Salon Chicago	http://www.dchicago.com/salonchicago/
Chicago Hotels in Chicago Illinois	http://www.readio.com/.chicago-hotels.html
Chicago - welcome to Chicago Illinois	http://www.dchicago.com/chicagoillinois/

Table 3. Top 5 Results returned by Geolink and PageRank on the query “Fitness in Chicago IL”

7 Related Work

The first generation of link analysis algorithms like the HITS [11] and PageRank [5] algorithms only consider one type of linkage relation (the hyperlink relation between web pages) for their ranking. Recently, some researchers have started to consider linkage relations among different types of data objects. For example, Xi et al. [16] propose a unified link-analysis framework, called *link fusion*, for multi-type data objects by extending the traditional algebraic link-analysis ranking algorithms. However, they do not consider how to create good graph models for specific types of data objects, and notably they do not consider objects that model geographic locality. Other work has considered specific types of data objects, but most model these objects, and relationships between them, independently of the standard page linkage relation. This work includes a modified version of HITS, proposed by Miller et al. [12], that models both pages and user behavior as objects in a graph. Additionally, Davison [8] calculates authority and hub values of pages by representing the term-term, doc-doc, term-doc and doc-term relationships in a single matrix model. Cai et al. [7] propose an enhanced link-analysis ranking algorithm in which each web page is first partitioned into blocks using a vision-based page segmentation algorithm and then the link analysis is performed over block-to-page and page-to-block graph models. Balmin et al. [2] propose the *ObjectRank* system which applies authority-based ranking to keyword search in databases modeled as labeled graphs where the co-occurrence of keywords in a tuple corresponds to an edge in the graph. Nie et al. [13] propose *PopRank* in which objects relevant to a specific application domain are ranked in terms of their relevancy and popularity to answer user queries using the web information about these objects.

Our work follows the general theme of considering new types of data objects (i.e., geographic entities) to improve search performance. Our proposed approaches combine geographic locality and page linkage semantics to produce a single graph. Moreover, unlike most previous work, we compare different graph models to understand how to effectively capture the semantics of geographic entities embedded in web content. Hence, our geographically-sensitive link analysis algorithms can properly exploit correlations in the way page linkage and geographic linkage semantics are modeled. Zhang et al. [17] have also considered the use of a graph model that includes geographic entities. They propose a different graph model (which is hard to be generalized), and use the original HITS algorithm directly on this

model. In contrast, we compare the performance of several algorithms (including a HITS style algorithm) and consider the scalability issue. More importantly, we empirically show that our *GeoLink* algorithm, which is designed to fully use the correlation between geographic locality and page linkage semantics, outperforms the other algorithms.

8 Conclusion

We proposed several link-based algorithms to rank web pages in a geographically sensitive fashion. In addition to the hyperlinks between pages, the existence of an *l*-geontology within a page is represented as a link as well. Both linkage relations are explored by our algorithms showing that geographic content can be exploited to improve the performance of geographically-oriented search. As for future research, we plan to extend our link analysis approaches by considering more sophisticated geographic features.

References

- [1] E. Amitay, N. Har’El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR*, pages 273–280, 2004.
- [2] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *VLDB*, pages 564–575, 2004.
- [3] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR*, pages 104–111, 1998.
- [4] L. Breyer. Markovian page ranking distributions: some theory and simulations. *Preprint*, 2002.
- [5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.
- [6] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of web pages. In *WebDB*, pages 91–96, 1999.
- [7] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *SIGIR*, pages 440–447, 2004.
- [8] B. D. Davison. Toward a unification of text and link analysis. In *SIGIR*, pages 367–368, 2003.
- [9] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scope of web resources. In *VLDB*, pages 545–556, September 2000.
- [10] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.*, 15(4):784–796, 2003.
- [11] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [12] J. C. Miller, G. Rae, and F. Schaefer. Modifications of kleinberg’s hits algorithm using matrix exponentiation and weblog records. In *SIGIR*, pages 444–445, 2001.
- [13] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: bringing order to web objects. In *WWW*, pages 567–574, 2005.
- [14] P. Tsaparas. *Link Analysis Ranking*. PhD thesis, University of Toronto, 2004.
- [15] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting dominant locations from search queries. In *SIGIR*, pages 424–431, 2005.
- [16] B. Z. Wensi Xi, Z. Chen, Y. Lu, S. Yan, W.-Y. Ma, and E. A. Fox. Link fusion: A unified link analysis framework for multi-type interrelated data objects. In *WWW*, pages 319–327, 2004.
- [17] J. Zhang, Y. Ishikawa, and H. Kitagawa. Extended link analysis for extracting spatial information hubs. In *WIRI ’05*, pages 17–22. IEEE, 2005.