Geographically Focused Collaborative Crawling

Weizheng Gao
Genieknows.com
Halifax, NS, Canada
wgao@genieknows.com

Hyun Chul Lee*
University of Toronto
Toronto, ON, Canada
leehyun@cs.toronto.edu

Yingbo Miao
Genieknows.com
Halifax, NS, Canada
ymiao@genieknows.com

ABSTRACT

A collaborative crawler is a group of crawling nodes, in which each crawling node is responsible for a specific portion of the web. We study the problem of collecting geographically-aware pages using collaborative crawling strategies. We first propose several collaborative crawling strategies for the geographically focused crawling, whose goal is to collect web pages about specified geographic locations, by considering features like URL address of page, content of page, extended anchor text of link, and others. Later, we propose various evaluation criteria to qualify the performance of such crawling strategies. Finally, we experimentally study our crawling strategies by crawling the real web data showing that some of our crawling strategies greatly outperform the simple URL-hash based partition collaborative crawling, in which the crawling assignments are determined according to the hash-value computation over URLs. More precisely, features like URL address of page and extended anchor text of link are shown to yield the best overall performance for the geographically focused crawling.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*performance measures*

General Terms

Measurement, Performance, Experimentation

Keywords

Collaborative crawling, geographically focused crawling, geographic entities

1. INTRODUCTION

While most of the current search engines are effective for *pure keyword-oriented searches*, these search engines are not fully effective for *geographic-oriented keyword searches*. For instance, queries like "restaurants in New York, NY" or "good plumbers near 100 milam street, Houston, TX" or "romantic hotels in Las Vegas, NV" are not properly managed by traditional web search engines. Therefore, in recent years, there has been surge of interest within the search industry on the *search localization* (e.g., Google Local[1], Yahoo Local[2]). The main aim of such search localization is to allow the user to perform the search according his/her keyword input as well as the geographic location of his/her interest.

Due to the current size of the Web and its dynamical nature, building a large scale search engine is challenging and it is still active area of research. For instance, the design of efficient crawling strategies and policies have been extensively studied in recent years (see [9] for the overview of the field). While it is possible to build geographically sensitive search engines using the full web data collected through a standard web crawling, it would rather be more attractive to build such search engines over a more focused web data collection which are only relevant to the targeted geographic locations. Focusing on the collection of web pages which are relevant to the targeted geographic location would leverage the overall processing time and efforts for building such search engines. For instance, if we want to build a search engine targeting those users in New York, NY, then we can build it using the web collection, only relevant to the city of New York, NY. Therefore, given intended geographic regions for crawling, we refer the task of collecting web pages, relevant to the intended geographic regions as *geographically focused crawling*.

The idea of focusing on a particular portion of the web for crawling is not novel. For instance, the design of efficient topic-oriented or domain-oriented crawling strategies has been previously studied [8, 23, 24]. However, there has been little previous work on incorporating the geographical dimension of web pages to the crawling. In this paper, we study various aspects of crawling when the geographical dimension is considered.

While the basic idea behind the standard crawling is straightforward, the collaborative crawling or parallel crawling is often used due to the performance and scalability issues that might arise during the real crawling of the web [12, 19]. In a collaborative or parallel crawler, the multiple crawling nodes are run in parallel on a multiprocessor or in a distributed manner to maximize the download speed and to further improve the overall performance especially for the scalability of crawling. Therefore, we study the geographically focused crawling under the collaborative setting, in which the targeted geographic regions are divided and then assigned to each participating crawling node. More precisely,

*This work was done while the author was visiting Genieknows.com

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

[1]http://local.google.com
[2]http://local.yahoo.com

in a *geographically focused collaborative crawler*, there will be a set of geographically focused crawling nodes in which each node is only responsible for collecting those web pages, relevant to its assigned geographic regions. Furthermore, there will be additional set of general crawling nodes which aim to support other geographically focused crawling nodes through the general crawling (download of pages which are not geographically-aware). The main contributions of our paper are follows:

1. We propose several geographically focused collaborative crawling strategies whose goal is to collect web pages about the specified geographic regions.
2. We propose several evaluation criteria for measuring the performance of a geographically focused crawling strategy.
3. We empirically study our proposed crawling strategies by crawling the real web. More specifically, we collect web pages pertinent to the top 100 US cities for each crawling strategy.
4. We empirically study *geographic locality*. That is pages which are geographically related are more likely to be linked compared to those which are not.

The rest of the paper is organized as follows. In Section 2, we introduce some of the previous works related to our geographically focused collaborative crawling. In Section 3, we describe the problem of geographically focused collaborative crawling and then we propose several crawling policies to deal with this type of crawling. In Section 4, we present evaluation models to measure the performance of a geographically focused collaborative crawling strategy. In Section 5, we present results of our experiments with the real web data. Finally, in Section 6, we present final remarks about our work.

2. RELATED WORKS

A focused crawler is designed to only collect web pages on a specified topic while transversing the web. The basic idea of a focused crawler is to optimize the priority of the unvisited URLs on the crawler frontier so that pages concerning a particular topic are retrieved earlier. Bra et al. [4] propose a focused web crawling method in the context of a client-based real-time search engine. Its crawling strategy is based on the intuition that relevant pages on the topic likely contain links to other pages on the same topic. Thus, the crawler follows more links from relevant pages which are estimated by a binary classifier that uses keyword and regular expression matchings. In spite of its reasonably acceptable performance, it has an important drawback as a relevant page on the topic might be hardly reachable when this page is not pointed by pages relevant to the topic.

Cho et al. [11] propose several strategies for prioritizing unvisited URLs based on the pages downloaded so far. In contrast to other focused crawlers in which a supervised topic classifier is used to control the way that crawler handles the priority of pages to be be downloaded, their strategies are based on considering some simple properties such as linkage or keyword information to define the priority of pages to be downloaded. They conclude that determining the priority of pages to be downloaded based on their PageRank value yield the best overall crawling performance.

Chakrabarti et al. [8] propose another type of focused crawler architecture which is composed of three components, namely classifier, distiller and crawler. The classifier makes the decision on the page relevancy to determine its future link expansion. The distiller identifies those hub pages, as defined in [20], pointing to many topic related pages to determine the priority of pages to be visited. Finally, the crawling module fetches pages using the list of pages provided by the distiller. In the subsequent work, Chakrabarti et al. [7] suggest that only a fraction of URLs extracted from a page are worth following. They claim that a crawler can avoid irrelevant links if the relevancy of links can be determined by the local text surrounding it. They propose alternative focused crawler architecture where documents are modeled as tag trees using DOM (Document Object Model). In their crawler, two classifiers are used, namely the "baseline" and the "apprentice". The baseline classifier refers to the module that navigates through the web to obtain the enriching training data for the apprentice classifier. The apprentice classifier, on the other hand, is trained over the data collected through the baseline classifier and eventually guides the overall crawling by determining the relevancy of links using the contextual information around them.

Diligenti et al. [14] use the context graph to improve the baseline best-first focused crawling method. In their approach, there is a classifier which is trained through the features extracted from the paths that lead to the relevant pages. They claim that there is some chance that some off-topic pages might potentially lead to highly relevant pages. Therefore, in order to mediate the hardness of identifying apparently off-topic pages, they propose the usage of context graph to guide the crawling. More precisely, first a context graph for seed pages is built using links to the pages returned from a search engine. Next, the context graph is used to train a set of classifiers to assign documents to different categories using their estimated distance, based on the number of links, to relevant pages on different categories. Their experimental results reveal that the context graph based focused crawler has a better performance and achieves higher relevancy compared to an ordinary best-first crawler.

Cho et al. [10] attempt to map and explore a full design space for parallel and distributed crawlers. Their work addresses issues of communication bandwidth, page quality and the division of work between local crawlers. Later, Chung et al. [12] study parallel or distributed crawling in the context of topic-oriented crawling. Basically, in their topic-oriented collaborative crawler, each crawling node is responsible for a particular set of topics and the page is assigned to the crawling node which is responsible for the topic which the page is relevant to. To determine the topic of page, a simple Naive-Bayes classifier is employed. Recently, Exposto et al. [17] study distributed crawling by means of the geographical partition of the web considering the multi-level partitioning of the reduced IP web link graph. Note that our IP-based collaborative crawling strategy is similar to their approach in spirit as we consider the IP-addresses related to the given web pages to distribute them among participating crawling nodes.

Gravano and his collaborators study the geographically-aware search problem in various works [15, 18, 5]. Particularly, in [15], how to compute the geographical scope of web resources is discussed. In their work, linkage and seman-

tic information are used to assess the geographical scope of web resources. Their basic idea is as follows. If a reasonable number of links pertinent to one particular geographic location point to a web resource and these links are smoothly distributed across the location, then this location is treated as one of the geographic scopes of the corresponding web resource. Similarly, if a reasonable number of location references is found within a web resource, and the location references are smoothly distributed across the location, then this location is treated as one of the geographical scopes of the web resource. They also propose how to solve aliasing and ambiguity. Recently, Markowotz et al. [22] propose the design and the initial implementation of a geographic search engine prototype for Germany. Their prototype extracts various geographic features from the crawled web dataset consisting of pages whose domain name contains "de". A geographic footprint, a set of relevant locations for page, is assigned to each page. Subsequently, the resulting footprint is integrated into the query processor of the search engine.

3. CRAWLING

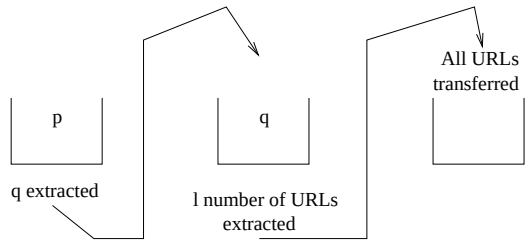
3.1 Problem Description

Even though, in theory, the targeted geographic locations of a geographically focused crawling can be any valid geographic location, in our paper, a geographic location refers to a city-state pair for the sake of simplicity. Therefore, given a list of city-state pairs, the goal of our geographically focused crawling is to collect web pages which are "relevant" to the targeted city-state pairs. Thus, after splitting and distributing the targeted city-state pairs to the participating crawling nodes, each participating crawling node would be responsible for the crawling of web pages relevant to its assigned city-state pairs.

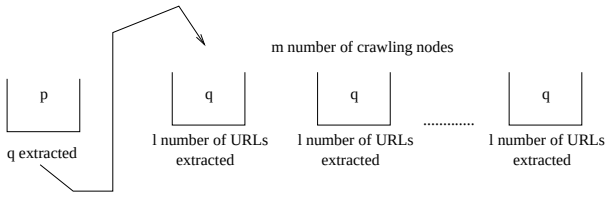
EXAMPLE 1. *Given {(New York, NY), (Houston, TX)} as the targeted city-state pairs and 3 crawling nodes {Cn_1, Cn_2 , Cn_3}, one possible design of geographically focused collaborative crawler is to assign (New York, NY) to Cn_1 and (Houston, TX) to Cn_2.*

Particularly, for our experiments, we perform the geographically focused crawling of pages targeting the top 100 US cities, which will be explained later in Section 5. We use some general notations to denote the targeted city-state pairs and crawling nodes as follows. Let $TC = \{(c_1, s_1), \ldots, (c_n, s_n)\}$ denote the set of targeted city-state pairs for our crawling where each (c_i, s_i) is a city-state pair. When it is clear in the context, we will simply denote (c_i, s_i) as c_i. Let $CR = \{Cn_1, \ldots, Cn_m\}$ denote the set of participating crawling nodes for our crawling. The main challenges that have to be dealt by a geographically focused collaborative crawler are the following:

- How to split and then distribute $TC = \{c_1, \ldots, c_n\}$ among the participating $CR = \{Cn_1, \ldots, Cn_m\}$
- Given a retrieved page p, based on what criteria we assign the extracted URLs from p to the participating crawling nodes.



a) All l URLs extracted from q are transferred to another crawling node (*the worst scenario for policy A*)



b) Page q is transferred to the m number of crawling nodes, but all URLs extracted from each q of the crawling nodes are not transferred to other crawling nodes (*the best scenario for policy B*)

Figure 1: Exchange of the extracted URLs

3.2 Assignment of the extracted URLs

When a crawling node extracts the URLs from a given page, it has to decide whether to keep the URLs for itself or transfer them to other participating crawling nodes for further fetching of the URLs. Once the URL is assigned to a particular crawling node, it may be added to the node's pending queue. Given a retrieved page p, let $pr(c_i|p)$ be the probability that page p is about the city-state pair c_i. Suppose that the targeted city-state pairs are given and they are distributed over the participating crawling nodes. There are mainly two possible policies for the exchange of URLs between crawling nodes.

- **Policy A**: Given the retrieved page p, let c_i be the most probable city-state pair about p, i.e. $arg\,max_{c_i \in TC}$ $pr(c_i|p)$. We assign each extracted URL from page p to the crawling node Cn_j responsible on c_i
- **Policy B**: Given the retrieved page p, let $\{c_{p_1}, \ldots, c_{p_k}\}$ $\subset TC$ be the set of city-state pairs whose $Pr(c_{p_i}|p) \neq 0$. We assign each extracted URL from page p to EACH crawling node Cn_j responsible on $c_{p_i} \in TC$,

LEMMA 2. *Let b be the bandwidth cost and let c be the inter-communication cost between crawling nodes. If $b > c$, then the Policy A is more cost effective than the Policy B.*

Proof: *Given an extracted URL q from page p, let m be the number of crawling nodes used by the Policy B (crawling nodes which are assigned to download q). Since the cost for the policy A and B is equal when $m = 1$, we suppose $m \geq 2$. Let l be the total number of URLs extracted from q. Let $C(A)$ and $C(B)$ be the sum of total inter-communication cost plus the bandwidth cost for the Policy A and Policy B respectively. One can easily verify that the cost of download for q and all URLs extracted from q is given as $C(A) \leq b+l\cdot(c+b)$ as shown in Figure 1a) and $C(B) \geq m\cdot b+l\cdot m\cdot b$.*

as shown in Figure 1b). Therefore, it follows that $C(A) \leq C(B)$ *since* $m \geq 2$ *and* $b > c$.

The assignment of extracted URLs for each retrieved page of all crawling collaboration strategies that we consider next will be based on the Policy A.

3.3 Hash Based Collaboration

We consider the hash based collaboration, which is the approach taken by most of collaborative crawlers, for the sake of comparison of this basic approach to our geographically focused collaboration strategies. The goal of hash based collaboration is to implementing a distributed crawler partition over the web by computing hash functions over URLs. When a crawling node extracts a URL from the retrieved page, a hash function is then computed over the URL. The URL is assigned to the participating crawling node responsible for the corresponding hash value of the URL. Since we are using a uniform hash function for our experiments, we will have a considerable data exchange between crawling nodes since the uniform hash function will map most of URLs extracted from the retrieved page to remote crawling nodes.

3.4 Geographically Focused Collaborations

We first divide up CR, the set of participating crawling nodes, into *geographically sensitive* nodes and *general* nodes. Even though, any combination of geographically sensitive and general crawling nodes is allowed, the architecture of our crawler consists of five geographically sensitive and one general crawling node for our experiments. A geographically sensitive crawling node will be responsible for the download of pages pertinent to a subset targeted city-state pairs while a general crawling node will be responsible for the download of pages which are not geographically-aware supporting other geographically sensitive nodes.

Each collaboration policy considers a particular set of features for the assessment of the geographical scope of page (whether a page is pertinent to a particular city-state pair or not). From the result of this assessment, each extracted URL from the page will be assigned to the crawling node that is responsible for the download of pages pertinent to the corresponding city-state pair.

3.4.1 URL Based

The intuition behind the URL based collaboration is that pages containing a targeted city-state pair in their URL address might potentially guide the crawler toward other pages about the city-state pair. More specifically, for each extracted URL from the retrieved page p, we verify whether the city-state pair c_i is found somewhere in the URL address of the extracted URL. If the city-state pair c_i is found, then we assign the corresponding URL to the crawling node which is responsible for the download of pages about c_i.

3.4.2 Extended Anchor Text Based

Given link text l, an extended anchor text of l is defined as the set of prefix and suffix tokens of l of certain size. It is known that extended anchor text provides valuable information to characterize the nature of the page which is pointed by link text. Therefore, for the extended anchor text based collaboration, our assumption is that pages associated with the extended anchor text, in which a targeted city-state pair c_i is found, will lead the crawler toward those pages about c_i. More precisely, given retrieved page p, and the extended anchor text l found somewhere in p, we verify whether the city-state pair $c_i \subset TC$ is found as part of the extended anchor text l. When multiple findings of city-state occurs, then we choose the city-state pair that is the closest to the link text. Finally, we assign the URL associated with l to the crawling node that is responsible for the download of pages about c_i.

3.4.3 Full Content Based

In [15], the location reference is used to assess the geographical scope of page. Therefore, for the full content based collaboration, we perform a content analysis of the retrieved page to guide the crawler for the future link expansion. Let $pr((c_i, s_i)|p)$ be the probability that page p is about city-state pair (c_i, s_i). Given TC and page p, we compute $pr((c_i, s_i)|p)$ for $(c_i, s_i) \in TC$ as follows:

$$pr((c_i, s_i)|p) = \alpha \cdot \#((c_i, s_i), p) + (1 - \alpha) \cdot pr(s_i|c_i) \cdot \#(c_i, p) \quad (1)$$

where $\#((c_i, s_i), p)$ denotes the number of times that the city-state pair (c_i, s_i) is found as part of the content of p, $\#(c_i, p)$ denotes the number of times (independent of $\#((c_i, s_i), p)$) that the city reference c_i is found as part of the content of p, and α denotes the weighting factor. For our experiments, $\alpha = 0.7$ was used.

The probability $pr(s_i|c_i)$ is calculated under two simplified assumptions: (1) $pr(s_i|c_i)$ is dependent on the real population size of (c_i, s_i) (e.g., Population of Kansas City, Kansas is 500,000). We obtain the population size for each city city-data.com[3]. (2) $pr(s_i|c_i)$ is dependent on the number of times that the state reference is found (independent of $\#((c_i, s_i), p)$) as part of the content of p. In other words, our assumption for $pr(s_j|c_i)$ can be written as

$$pr(s_i|c_i) \propto \beta S(s_i|c_i) + (1 - \beta)\tilde{S}(s_i|p) \quad (2)$$

where $S(s_i|c_i)$ is the normalized form of the population size of (c_i, s_i), $\tilde{S}(s_i|p)$ is the normalized form of the number of appearances of the state reference s_i, independent of $\#((c_i, s_i), p))$, within the content of p, and β denotes the weighting factor. For our experiments, $\beta = 0.5$ was used. Therefore, $pr((c_i, s_i)|p)$ is computed as

$$\begin{aligned} pr((c_i, s_i)|p) &= \alpha \cdot \#((c_i, s_i), p) + (1 - \alpha) \cdot (\beta S(s_i|c_i) \\ &+ (1 - \beta)\tilde{S}(s_i|p)) \cdot \#(c_i, p) \end{aligned} \quad (3)$$

Finally, given a retrieve page p, we assign all extracted URLs from p to the crawling node which is responsible for pages relevant to $arg\ max_{(c_i,s_i) \in TC}\ Pr((c_i, s_i)|p)$.

3.4.4 Classification Based

Chung et al. [12] show that the classification based collaboration yields a good performance for the topic-oriented collaborative crawling. Our classification based collaboration for the geographically crawling is motivated by their work. In this type of collaboration, the classes for the classifier are the partitions of targeted city-state pairs. We train our classifier to determine $pr(c_i|p)$, the probability that the retrieved page p is pertinent to the city-state pair c_i. Among various possible classification methods, we chose the Naive-Bayes classifier [25] due to its simplicity. To obtain training

[3]http://www.city-data.com

data, pages from the Open Directory Project (ODP)[4] were used. For each targeted city-state pair, we download all pages under the corresponding city-state category which, in turn, is the child category for the "REGIONAL" category in the ODP. The number of pages downloaded for each city-state pair varied from 500 to 2000. We also download a set of randomly chosen pages which are not part of any city-state category in the ODP. We download 2000 pages for this purpose. Then, we train our Naive-Bayes classifier using these training data. Our classifier determines whether a page p is pertinent to either of the targeted city-state pairs or it is not relevant to any city-state pair at all. Given the retrieved page p, we assign all extracted URLs from p to the crawling node which is responsible for the download of pages which are pertinent to $arg\ max_{c_i \in TC} pr(c_i|p)$.

3.4.5 IP-Address Based

The IP-address of the web service indicates the geographic location at which the web service is hosted. The IP-address based collaboration explores this information to control the behavior of the crawler for further downloads. Given a retrieved page p, we first determine the IP-address of the web service from which the crawler downloaded p. With this IP-address, we use the IP-address mapping tool to obtain the corresponding city-state pair of the given IP, and then we assign all extracted URLs of page p to the crawling node which is responsible on the computed city-state pair. For the IP-address mapping tool, freely available IP address mapping tool, *hostip.info*(API)[5] is employed.

3.5 Normalization and Disambiguation of City Names

As indicated in [2, 15], problems of *aliasing* and *ambiguity* arise when one wants to map the possible city-state reference candidate to an unambiguous city-state pair. In this section, we describe how we handle these issues out.

- **Aliasing:** Many times different names or abbreviations are used for the same city name. For example, Los Angeles can be also referred as LA or L.A. Similar to [15], we used the web database of the United States Postal Service (USPS)[6] to deal with aliasing. The service returns a list of variations of the corresponding city name given the zip code. Thus, we first obtained the list of representative zip codes for each city in the list using the US Zip Code Database product, purchased from ZIPWISE[7], and then we obtain the list of possible names and abbreviations for each city from the USPS.

- **Ambiguity:** When we deal with city names, we have to deal with the ambiguity of the city name reference. First, we can not guarantee whether the possible city name reference actually refers to the city name. For instance, New York might refer to New York as city name or New York as part of the brand name "New York Fries" or New York as state name. Second, a city name can refer to cities in different states. For example, four states, New York, Georgia, Oregon and California, have a city called Albany. For both cases, unless we fully analyze the context in which the reference was made, the city name reference might be inherently ambiguous. Note that for the full content based collaboration, the issue of ambiguity is already handled through the term $pr(s_i|c_i)$ of the Eq. 2. For the extended anchor text based and the URL based collaborations, we always treat the possible city name reference as the city that has the largest population size. For instance, *Glendale* found in either the URL address of page or the extended anchor text of page would be treated as the city name reference for Glendale, AZ.[8].

4. EVALUATION MODELS

To assess the performance of each crawling collaboration strategy, it is imperative to determine how much geographically-aware pages were downloaded for each strategy and whether the downloaded pages are actually pertinent to the targeted geographic locations. Note that while some previous works [2, 15, 18, 5] attempt to define precisely what a geographically-aware page is, determining whether a page is geographically-aware or not remains as an open problem [2, 18]. For our particular application, we define the notion of geographical awareness of page through geographic entities [21]. We refer the address description of a physical organization or a person as *geographic entity*. Since the targeted geographical city-state pairs for our experiments are the top 100 US cities, a geographic entity in the context of our experiments are further simplified as an address information, following the standard US address format, for any of the top 100 US cities. In other words, a geographic entity in our context is a sequence of *Street Number, Street Name, City Name* and *State Name*, found as part of the content of page. Next, we present various evaluation measures for our crawling strategies based on geographic entities. Additionally, we present traditional measures to quantify the performance of any collaborative crawling. Note that our evaluation measures are later used in our experiments.

- **Geo-coverage**: When a page contain at least one geographic entity (i.e. address information), then the page is clearly a geographically aware page. Therefore, we define the *geo-coverage* of retrieved pages as the number of retrieved pages with at least one geographic entity, pertinent to the targeted geographical locations (e.g., the top US 100 cities) over the total number of retrieved pages.

- **Geo-focus**: Each crawling node of the geographically focused collaborative crawler is responsible for a subset of the targeted geographic locations. For instance, suppose we have two geographically sensitive crawling nodes Cn_1, and Cn_2, and the targeted city-state pairs as {(New York, NY),(Los Angeles, CA)}. Suppose Cn_1 is responsible for crawling pages pertinent to (New York, NY) while Cn_2 is responsible for crawling

[4]http://www.dmoz.org
[5]http://www.hostip.info
[6]http://www.usps.gov
[7]http://www.zipwise.com

[8]Note that this simple approach does minimally hurt the overall crawling. For instance, in many cases, even the incorrect assessment of the state name reference New York instead of the correct city name reference New York, would result into the assignment of all extracted URLs to the correct crawling node.

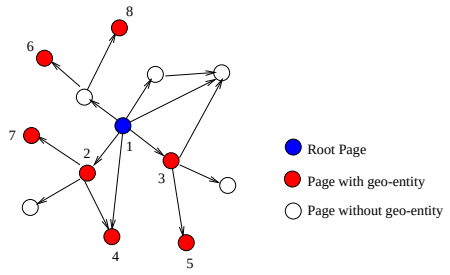


Figure 2: An example of geo-centrality measure

pages pertinent to (Los Angeles, CA). Therefore, if the Cn_1 has downloaded a page about Los Angeles, CA, then this would be clearly a failure of the collaborative crawling approach.

To formalize this notion, we define the *geo-focus* of a crawling node, as the number of retrieved pages that contain at least one geographic entity of the assigned city-state pairs of the crawling node.

- **Geo-centrality:** One of the most frequently and fundamental measures used for the analysis of network structures is the *centrality measure* which address the question of how central a node is respect to other nodes in the network. The most commonly used ones are the degree centrality, eigenvector centrality, closeness centrality and betweenness centrality [3]. Motivated by the closeness centrality and the betweenness centrality, Lee et al. [21] define novel centrality measures to assess how a node is central with respect to those geographically-aware nodes (pages with geographic entities). A *geodesic path* is the shortest path, in terms of the number of edges transversed, between a specified pair of nodes. Geo-centrality measures are based on the geodesic paths from an arbitrary node to a geographically aware node.

  Given two arbitrary nodes, p_i, p_j, let $GD(p_i, p_j)$ be the geodesic path based distance between p_i and p_j (the length of the geodesic path). Let $w_{GD(p_i,p_j)} = 1/m^{GD(p_i,p_j)}$ for some $m \in \Re$ and we define $\delta(p_i, p_j)$ as

$$\delta(p_i, p_j) = \begin{cases} w_{GD(p_i,p_j)} & \text{if } p_j \text{ is geographically aware node} \\ 0 & otherwise \end{cases}$$

  For any node p_i, let $\Omega_k(p_i) = \{p_j | GD(p_i, p_j) < k\}$ be the set nodes of whose geodesic distance from p_i is less than k.

  Given p_i, let $GCt_k(p_i)$ be defined as

$$GCt_k(p_i) = \sum_{p_j \in \Omega_k(p_i)} \delta(p_i, p_j)$$

  Intuitively the geo-centrality measure computes how many links have to be followed by a user which starts his navigation from page p_i to reach geographically-aware pages. Moreover, $w_{GD(p_i,p_j)}$ is used to penalize each following of link by the user.

EXAMPLE 3. *Let consider the graph structure of Figure 2. Suppose that the weights are given as* $w_0 = 1, w_1 = 0.1, w_2 = 0.01$, *i.e. each time a user navigates a link, we penalize it with* 0.1. *Given the root node* 1 *containing at least one geo-entity, we have* $\Omega_2(node\ 1) = \{1, \ldots, 8\}$. *Therefore, we have* $w_{GD(node\ 1, node\ 1)} = 1$, $w_{GD(node\ 1, node\ 2)} = 0.1$, $w_{GD(node\ 1, node\ 3)} = 0.1$, $w_{GD(node\ 1, node\ 4)} = 0.1$, $w_{GD(node\ 1, node\ 5)} = 0.01$, $w_{GD(node\ 1, node\ 6)} = 0.01$, $w_{GD(node\ 1, node\ 7)} = 0.01$, $w_{GD(node\ 1, node\ 8)} = 0.01$. *Finally,* $GCt_k(node\ 1) = 1.34$.

- **Overlap:** The Overlap measure is first introduced in [10]. In the collaborative crawling, it is possible that different crawling nodes download the same page multiple times. Multiple downloads of the same page are clearly undesirable. Therefore, the *overlap* of retrieved pages is defined as $\frac{N-I}{N}$ where N denotes the total number of downloaded pages by the overall crawler and I denotes the number of unique downloaded pages by the overall crawler. Note that the hash based collaboration approach does not have any overlap.

- **Diversity:** In a crawling, it is possible that the crawling is biased toward a certain domain name. For instance, a crawler might find a crawler trap which is an infinite loop within the web that dynamically produces new pages trapping the crawler within this loop [6]. To formalize this notion, we define the *diversity* as $\frac{S}{N}$ where S denotes the number of unique domain names of downloaded pages by the overall crawler and N denotes the total number of downloaded pages by the overall crawler.

- **Communication overhead:** In a collaborative crawling, the participating crawling nodes need to exchange URLs to coordinate the overall crawling work. To quantify how much communication is required for this exchange, the communication overhead is defined in terms of the exchanged URLs per downloaded page [10].

5. CASE STUDY

In this section, we present the results of experiments that we conducted to study various aspects of the proposed geographically focused collaborative crawling strategies.

5.1 Experiment Description

We built an geographically focused collaborative crawler that consists of one general crawling node, Cn_0 and five geographically sensitive crawling nodes, $\{Cn_1, \ldots, Cn_5\}$, as described in Section 3.4. The targeted city-state pairs were the top 100 US cities by the population size, whose list was obtained from the city-data.com[9].

We partition the targeted city-state pairs according to their time zone to assign these to the geographically sensitive crawling nodes as shown in Table 1. In other words, we have the following architecture design as illustrated in Figure 3. Cn_0 is general crawler targeting pages which are not geographically-aware. Cn_1 targets the Eastern time zone with 33 cities. Cn_2 targets the Pacific time zone with 22 cities. Cn_3 targets the Mountain time zone with 10 cities.

[9]www.city-data.com

Time Zone	State Name	Cities
Central	AL	Birmingham,Montgomery, Mobile
Alaska	AK	Anchorage
Mountain	AR	Phoenix, Tucson, Mesa, Glendale, Scottsdale
Pacific	CA	Los Angeles , San Diego , San Jose San Francisco, Long Beach, Fresno Oakland, Santa Ana, Anaheim Bakersfield, Stockton, Fremont Glendale,Riverside , Modesto Sacramento, Huntington Beach
Mountain	CO	Denver, Colorado Springs, Aurora
Eastern	DC	Washington
Eastern	FL	Hialeah
Eastern	GA	Atlanta, Augusta-Richmond County
Hawaii	HI	Honolulu
Mountain	ID	Boise
Central	IL	Chicago
Central	IN	Indianapolis,Fort Wayne
Central	IA	Des Moines
Central	KA	Wichita
Eastern	KE	Lexington-Fayette, Louisville
Central	LO	New Orleans, Baton Rouge Shreveport
Eastern	MD	Baltimore
Eastern	MA	Boston
Eastern	MI	Detroit, Grand Rapids
Central	MN	Minneapolis, St. Paul
Central	MO	Kansas City , St. Louis
Central	NE	Omaha , Lincoln
Pacific	NV	Las Vegas
Eastern	NJ	Newark , Jersey City
Mountain	NM	Albuquerque
Eastern	NY	New York, Buffalo,Rochester,Yonkers
Eastern	NC	Charlotte, Raleigh,Greensboro
Eastern	OH	Durham , Winston-Salem Columbus , Cleveland
Central	OK	Cincinnati , Toledo , Akron Oklahoma City, Tulsa
Pacific	OR	Portland
Eastern	PA	Philadelphia,Pittsburgh
Central	TX	Houston,Dallas,San Antonio,Austin El Paso,Fort Worth Arlington, Corpus Christi Plano , Garland ,Lubbock , Irving
Eastern	VI	Virginia Beach , Norfolk Chesapeake, Richmond , Arlington
Pacific	WA	Seattle , Spokane , Tacoma
Central	WI	Milwaukee , Madison

Table 1: Top 100 US cities and their time zone

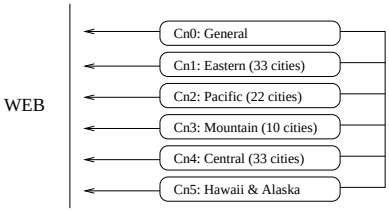


Figure 3: Architecture of our crawler

Cn_4 targets the Central time zone with 33 cities. Finally, Cn_5 targets the Hawaii-Aleutian and Alaska time zones with two cities.

We developed our collaborative crawler by extending the open source crawler, *larbin*[10] written in C++. Each crawling node was to dig each domain name up to the five levels of depth. The crawling nodes were deployed over 2 servers, each of them with 3.2 GHz dual P4 processors, 1 GB of RAM, and 600 GB of disk space. We ran our crawler for the period of approximately 2 weeks to download approximately 12.8 million pages for each crawling strategy as shown in Table 2. For each crawling process, the usable bandwidth was limited to 3.2 mbps, so the total maximum bandwidth used by our crawler was 19.2 mbps. For each crawling, we used the category "Top: Regional: North America: United States" of the ODP as the seed page of crawling. The IP mapping tool used in our experiments did not return the corresponding city-state pairs for Alaska and Hawaii, so we ignored Alaska and Hawaii for our IP-address based collaborative crawling.

[10]http://larbin.sourceforge.net/index-eng.html

Type of collaboration	Download size
Hash Based	12.872 m
URL Based	12.872 m
Extended Anchor Text Based	12.820 m
Simple Content Analysis Based	12.878 m
Classification Based	12.874 m
IP Address Based	12.874 m

Table 2: Number of downloaded pages

5.2 Discussion

5.2.1 Quality Issue

As the first step toward the performance evaluation of our crawling strategies, we built an extractor for the extraction of geographic entities (addresses) from downloaded pages. Our extractor, being a gazetteer based, extracted those geographic entities using a dictionary of all possible city name references for the top 100 US cities augmented by a list of all possible street abbreviations (e.g., street, avenue, av., blvd) and other pattern matching heuristics. Each extracted geographic entity candidate was further matched against the database of possible street names for each city that we built from the 2004 TIGER/Line files[11]. Our extractor was shown to yield 96% of accuracy out of 500 randomly chosen geographic entities.

We first analyze the *geo-coverage* of each crawling strategy as shown in Table 3. The top performers for the geo-coverage are the URL based and extended anchor text based collaborative strategies whose portion of pages downloaded with geographic entities was 7.25% and 7.88%, respectively, strongly suggesting that URL address of page and extended anchor text of link are important features to be considered for the discovery of geographically-aware pages. The next best performer with respect to geo-coverage was the full content based collaborative strategy achieving geo-coverage of 4.89%. Finally, the worst performers in the group of geographically focused collaborative policies were the classification based and the IP-address based strategies. The poor performance of the IP-address based collaborative policy shows that the actual physical location of web service is not necessarily associated with the geographical scopes of pages served by web service. The extremely poor performance of the classification based crawler is surprising since this kind of collaboration strategy shows to achieve good performance for the topic-oriented crawling [12]. Finally, the worst performance is observed with the URL-hash based collaborative policy as expected whose portion of pages with geographical entities out of all retrieved pages was less than 1%. In conclusion, the usage of even simple but intuitively sounding geographically focused collaborative policies can improve the performance of standard collaborative crawling by a factor of 3 to 8 for the task of collecting geographically-aware pages.

To check whether each geographically sensitive crawling node is actually downloading pages corresponding to their assigned city-state pairs, we used the geo-focus as shown in

[11]http://www.census.gov/geo/www/tiger/tiger2004se/tgr2004se.html

Type of collaboration	Cn0	Cn1	Cn2	Cn3	Cn4	Cn5	Average	Average (without Cn0)
URL-Hash Based	1.15%	0.80%	0.77%	0.75%	0.82%	0.86%	0.86%	0.86%
URL Based	3.04%	7.39%	9.89%	9.37%	7.30%	13.10%	7.25%	8.63%
Extended Anchor Text Based	5.29%	6.73%	9.78%	9.99%	6.01%	12.24%	7.88%	8.58%
Full Content Based	1.11%	3.92%	5.79%	6.87%	3.24%	8.51%	4.89%	5.71%
Classification Based	0.49%	1.23%	1.20%	1.27%	1.22%	1.10%	1.09%	1.21%
IP-Address Based	0.81%	2.02%	1.43%	2.59%	2.74%	0.00%	1.71%	2.20%

Table 3: Geo-coverage of crawling strategies

Type of collaboration	Cn1	Cn2	Cn3	Cn4	Cn5	Average
URL based	91.7%	89.0%	82.8%	94.3%	97.6%	91.1%
Extended anchor text based	82.0%	90.5%	79.6%	76.8%	92.3%	84.2%
Full content based	75.2%	77.4%	75.1%	63.5%	84.9%	75.2%
Classification based	43.5%	32.6%	5.5%	25.8%	2.9%	22.1%
IP-Address based	59.6%	63.6%	55.6%	80.0%	0.0%	51.8%

Table 4: Geo-focus of crawling strategies

Type of collaboration	Cn0	Cn1	Cn2	Cn3	Cn4	Cn5	Average
URL-hash based	0.45	0.47	0.46	0.49	0.49	0.49	0.35
URL based	0.39	0.2	0.18	0.16	0.24	0.07	0.18
Extended anchor text based	0.39	0.31	0.22	0.13	0.32	0.05	0.16
Full content based	0.49	0.35	0.31	0.29	0.39	0.14	0.19
Classification based	0.52	0.45	0.45	0.46	0.46	0.45	0.26
IP-Address based	0.46	0.25	0.31	0.19	0.32	0.00	0.27

Table 5: Number of unique geographic entities over the total number of geographic entities

Table 4. Once again, the URL-based and the extended anchor text based strategies show to perform well with respect to this particular measure achieving in average above 85% of geo-focus. Once again, their relatively high performance strongly suggest that the city name reference within a URL address of page or an extended anchor text is a good feature to be considered for the determination of geographical scope of page. The geo-focus value of 75.2% for the content based collaborative strategy also suggests that the locality phenomena which occurs with the topic of page also occurs within the geographical dimension as well. It is reported, [13], that pages tend to reference (point to) other pages on the same general topic. The relatively high geo-focus value for the content based collaborative strategy indicates that pages on the similar geographical scope tend to reference each other. The IP-address based policy achieves 51.7% of geo-focus while the classification based policy only achieves 22.7% of geo-focus. The extremely poor geo-focus of the classification based policy seems to be due to the failure of the classifier for the determination of the correct geographical scope of page.

In the geographically focused crawling, it is possible that pages are biased toward a certain geographic locations. For instance, when we download pages on Las Vegas, NV, it is possible that we have downloaded a large number of pages which are focused on a few number of casino hotels in Las Vegas, NV which are highly referenced to each other. In this case, quality of the downloaded pages would not be that good since most of pages would contain a large number of very similar geographic entities. To formalize the notion, we depict the ratio between the number of unique geographic entities and the total number of geographic entities from the retrieved pages as shown in Table 5. This ratio verifies whether each crawling policy is covering sufficient number of pages whose geographical scope is different. It is interesting

Type of collaboration	Geo-centrality
Hash based	0.0222
URL based	0.1754
Extended anchor text based	0.1519
Full content based	0.0994
Classification based	0.0273
IP-address based	0.0380

Table 6: Geo-centrality of crawling strategies

Type of collaboration	Overlap
Hash Based	None
URL Based	None
Extended Anchor Text Based	0.08461
Full Content Based	0.173239
Classification Based	0.34599
IP-address based	None

Table 7: Overlap of crawling strategies

to note that those geographically focused collaborative policies, which show to have good performance relative to the previous measures, such as the URL based, the extended anchor text based and the full content based strategies tend to discover pages with less diverse geographical scope. On the other hand, the less performed crawling strategies such as the IP-based, the classification based, the URL-hash based strategies are shown to collect pages with more diverse geographical scope.

We finally study each crawling strategy in terms of the geo-centrality measure as shown in Table 6. One may observe from Table 6 that the geo-centrality value provides an accurate view on the quality of the downloaded geo graphically-aware pages for each crawling strategy since the geo-centrality value for each crawling strategy follows what we have obtained with respect to geo-coverage and geo-precision. URL based and extended anchor text based strategies show to have the best geo-centrality values with 0.1754 and 0.1519 respectively, followed by the full content based strategy with 0.0994, followed by the IP based strategy with 0.0380, and finally the hash based strategy and the classification based strategy show to have similarly low geo-centrality values.

5.2.2 *Performance Issue*

In Table 7, we first show the overlap measure which reflects the number of duplicated pages out of the downloaded pages. Note that the hash based policy does not have any duplicated page since its page assignment is completely independent of other page assignment. For the same reason, the overlap for the URL based and the IP based strategies are none. The overlap of the extended anchor text

Type of collaboration	Diversity
Hash Based	0.0814
URL Based	0.0405
Extended Anchor Text Based	0.0674
Full Content Based	0.0688
Classification Based	0.0564
IP-address based	0.3887

Table 8: Diversity of crawling strategies

Type of collaboration	Communication overhead
URL-hash based	13.89
URL based	25.72
Extended anchor text based	61.87
Full content text based	46.69
Classification based	58.38
IP-Address based	0.15

Table 9: Communication-overhead

based is 0.08461 indicating that the extended anchor text of page computes the geographically scope of the corresponding URL in an almost unique manner. In other words, there is low probability that two completely different city name references are found within a URL address. Therefore, this would be another reason why the extended anchor text would be a good feature to be used for the partition of the web within the geographical context. The overlap of the full content based and the classification based strategies are relatively high with 0.173239 and 0.34599 respectively.

In Table 8, we present the diversity of the downloaded pages. The diversity values of geographically focused collaborative crawling strategies suggest that most of the geographically focused collaborative crawling strategies tend to favor those pages which are found grouped under the same domain names because of their crawling method. Especially, the relatively low diversity value of the URL based strongly emphasizes this tendency. Certainly, this matches with the intuition since a page like "http://www.houston-guide.com" will eventually lead toward the download of its child page "http://www.houston-guide.com/guide/arts/framearts.html" which shares the same domain.

In Table 9, we present the communication-overhead of each crawling strategy. Cho and Garcia-Molina [10] report that the communication overhead of the Hash-Based with two processors is well above five. The communication-overhead of the Hash-based policy that we have follows with what they have obtained. The communication overhead of geographically focused collaborative policies is relatively high due to the intensive exchange of URLs between crawling nodes.

In Table 10, we summarize the relative merits of the proposed geographically focused collaborative crawling strategies. In the Table, "Good" means that the strategy is expected to perform relatively well for the measure, "Not Bad" means that the strategy is expected to perform relatively acceptable for that particular measure, and "Bad" means that it may perform worse compared to most of other collaboration strategies.

5.3 Geographic Locality

Many of the potential benefits of topic-oriented collaborative crawling derive from the assumption of *topic locality*, that pages tend to reference pages on the same topic [12, 13]. For instance, a classifier is used to determine whether the child page is in the same topic as the parent page and then guide the overall crawling [12]. Similarly, for geographically focused collaborative crawling strategies we make the assumption of *geographic locality*, that pages tend to reference pages on the same geographic location. Therefore, the performance of a geographically focused collaborative crawling strategy is highly dependent on its way of exploiting the geographic locality. That is whether the corresponding strategy is based on the adequate features to determine the geographical similarity of two pages which are possibly linked. We empirically study in what extent the idea of geographic locality holds. Recall that given the list of city-state pairs $G = \{\tilde{c}_1, \ldots, \tilde{c}_k\}$ and a geographically focused crawling collaboration strategy (e.g., URL based collaboration), $pr(\tilde{c}_i|p_j)$ is the probability that page is p_j is pertinent to city-state pair c_i according to that particular strategy. Let $gs(p, q)$, geographic similarity between pages p, q, be

$$gs(p,q) = \begin{cases} 1 & \text{if } (arg\ max_{\tilde{c}_i \in G} Pr(\tilde{c}_i|p) \\ & = arg\ max_{\tilde{c}_j \in G} Pr(\tilde{c}_j|q)) \\ 0 & \text{otherwise} \end{cases}$$

In other words, our geographical similarity determines whether two pages are pertinent to the same city-state pair. Given Ω, the set of retrieved page for the considered crawling strategy, let $\delta(\Omega)$ and $\tilde{\delta}(\Omega)$ be

$$\delta(\Omega) = \frac{|\{(p_i, p_j) \in \Omega \times \Omega | p_i, p_j \text{ linked and } gs(p,q) = 1\}|}{|\{(p_i, p_j) \in \Omega \times \Omega | p_i, p_j \text{ linked}\}|}$$

$$\tilde{\delta}(\Omega) = \frac{|\{(p_i, p_j) \in \Omega \times \Omega | p_i, p_j \text{ not linked and } gs(p,q) = 1\}|}{|\{(p_i, p_j) \in \Omega \times \Omega | p_i, p_j \text{ not linked}\}|}$$

Note that $\delta(\Omega)$ corresponds to the probability that a pair of *linked* pages, chosen uniformly at random, is pertinent to the same city-state pair under the considered collaboration strategy while $\tilde{\delta}(\Omega)$ corresponds to the probability that a pair of *unlinked* pages, chosen uniformly at random, is pertinent to the same city-state pair under the considered collaboration strategy. Therefore, if the geographic locality occurs then we would expect to have high $\delta(\Omega)$ value compared to that of $\tilde{\delta}(\Omega)$. We selected the URL based, the classification based, and the full content based collaboration strategies, and calculated both $\delta(\Omega)$ and $\tilde{\delta}(\Omega)$ for each collaboration strategy. In Table 11, we show the results of our computation. One may observe from Table 11 that those pages that share the same city-state pair in their URL address have the high likelihood of being linked. Those pages that share the same city-state pair in their content have some likelihood of being linked. Finally, those pages which are classified as sharing the same city-state pair are less likely to be linked. We may conclude the following:

- The geographical similarity of two web pages affects the likelihood of being referenced. In other words, geographic locality, that pages tend to reference pages on the same geographic location, clearly occurs on the web.
- A geographically focused collaboration crawling strategy which properly explores the adequate features for determining the likelihood of two pages being in the same geographical scope would expect to perform well for the geographically focused crawling.

Type of collaboration	Geo-coverage	Geo-Focus	Geo-Connectivity	Overlap	Diversity	Communication
URL-Hash Based	Bad	Bad	Bad	Good	Good	Good
URL Based	Good	Good	Good	Good	Bad	Bad
Extended Anchor Text Based	Good	Good	Good	Good	Not Bad	Bad
Full Content Based	Not Bad	Not Bad	Not Bad	Not Bad	Not Bad	Bad
Classification Based	Bad	Bad	Bad	Bad	Not Bad	Bad
IP-Address	Bad	Bad	Bad	Good	Bad	Good

Table 10: Comparison of geographically focused collaborative crawling strategies

Type of collaboration	$\delta(\Omega)$	$\tilde{\delta}(\Omega)$
URL based	0.41559	0.02582
classification based	0.044495	0.008923
full content based	0.26325	0.01157

Table 11: Geographic Locality

6. CONCLUSION

In this paper, we studied the problem of geographically focused collaborative crawling by proposing several collaborative crawling strategies for this particular type of crawling. We also proposed various evaluation criteria to measure the relative merits of each crawling strategy while empirically studying the proposed crawling strategies with the download of real web data. We conclude that the URL based and the extended anchor text based crawling strategies have the best overall performance. Finally, we empirically showed geographic locality, that pages tend to reference pages on the same geographical scope. For the future research, it would be interesting to incorporate more sophisticated features (e.g., based on DOM structures) to the proposed crawling strategies.

7. ACKNOWLEDGMENT

We would like to thank Genieknows.com for allowing us to access to its hardware, storage, and bandwidth resources for our experimental studies.

8. REFERENCES

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *WWW*, pages 96–105, 2001.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR*, pages 273–280, 2004.

[3] S. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005.

[4] P. D. Bra, Y. K. Geert-Jan Houben, and R. Post. Information retrieval in distributed hypertexts. In *RIAO*, pages 481–491, 1994.

[5] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. Exploiting geographical location information of web pages. In *WebDB (Informal Proceedings)*, pages 91–96, 1999.

[6] S. Chakrabarti. *Mining the Web*. Morgan Kaufmann Publishers, 2003.

[7] S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated focused crawling through online relevance feedback. In *WWW*, pages 148–159, 2002.

[8] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640, 1999.

[9] J. Cho. *Crawling the Web: Discovery and Maintenance of Large-Scale Web Data*. PhD thesis, Stanford, 2001.

[10] J. Cho and H. Garcia-Molina. Parallel crawlers. In *WWW*, pages 124–135, 2002.

[11] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. *Computer Networks*, 30(1-7):161–172, 1998.

[12] C. Chung and C. L. A. Clarke. Topic-oriented collaborative crawling. In *CIKM*, pages 34–42, 2002.

[13] B. D. Davison. Topical locality in the web. In *SIGIR*, pages 272–279, 2000.

[14] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. Focused crawling using context graphs. In *VLDB*, pages 527–534, 2000.

[15] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *VLDB*, pages 545–556, 2000.

[16] J. Edwards, K. S. McCurley, and J. A. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. In *WWW*, pages 106–113, 2001.

[17] J. Exposto, J. Macedo, A. Pina, A. Alves, and J. Rufino. Geographical partition for distributed web crawling. In *GIR*, pages 55–60, 2005.

[18] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *CIKM*, pages 325–333, 2003.

[19] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.

[20] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[21] H. C. Lee and R. Miller. Bringing geographical order to the web. private communication, 2005.

[22] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger. Design and implementation of a geographic search engine. In *WebDB*, pages 19–24, 2005.

[23] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. A machine learning approach to building domain-specific search engines. In *IJCAI*, pages 662–667, 1999.

[24] F. Menczer, G. Pant, P. Srinivasan, and M. E. Ruiz. Evaluating topic-driven web crawlers. In *SIGIR*, pages 241–249, 2001.

[25] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.