Automatic Construction of a Geo-Referenced Search Engine Index

Lefteris Kozanidis Sofia Stamou
Computer Engineering and Informatics Department, Patras University, 26500 GREECE
kozanid@ceid.upatras.gr stamou@ceid.upatras.gr

ABSTRACT
Locality is an important aspect for a considerable fraction of web searches, which intend the retrieval of pages that are geographically relevant to the query entities. So far, research on answering location-sensitive queries has been addressed as a problem of automatically determining the geographic orientation of web searches. However, identifying localizable queries is a tedious and time-consuming task that requires processing large volumes of web transaction logs before being able to decipher the geographic orientation of the query keywords. Most importantly, localizable queries even if successfully identified cannot be effectively answered unless these are searched against a geographic-aware data index. In this article, we address the problem of location-sensitive web searches from the search engine perspective and we propose a novel framework for handling geography-oriented queries. Our framework integrates two distinct yet complementary modules, namely a geo-focused web crawler and a geo-referenced search engine index. In particular, we have implemented a crawler that focuses its web walkthroughs on geographic content via the assistance of a geographic ontology and a number of tools that we leverage from the NLP community. Then, we build scoring techniques that encapsulate the geo-spatial aspect of URLs into the crawler's priority queue. Based on the location-descriptive elements in the page URLs and anchor text, the crawler directs the pages to a location-sensitive downloader. This downloading module resolves the geographical references of the URL location elements and organizes them into indexable hierarchical structures. The location-aware URL hierarchies are linked to their respective pages, resulting into a geo-referenced index against which location-sensitive queries can be answered. The experimental evaluation of our techniques indicates that our geo-focused crawler outperforms other focused crawling methods in terms of both coverage and accuracy. Moreover, our study reveals that our geo-referenced index guarantees that geographically-relevant pages are prioritized over topically-relevant ones for location-sensitive queries.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: Indexing Methods; H.3.3 [**Information Search and Retrieval**]: Search Process; H.3.5 [**Online Information Services**]: Web-Based Services.

General Terms
Measurement, Performance, Experimentation.

Keywords
Location-sensitive web searches, focused crawling, geo-referenced index.

1. INTRODUCTION

Locality is an important aspect for a considerable fraction of web searches. According to the study of (Wang et al., 2005a) 14% of the web queries have geographical intentions, i.e. they pursue the retrieval of information that relates to a geographical area. Moreover, (Wang et al., 2005b) found that 79% of the web pages in .gov domain contain at least one geographical reference. Although location-sensitive web searches are gaining ground (Himmelstein, 2005), still search engines are not very effective in identifying localizable queries (Welch and Cho, 2008). This is partially because users rarely add location constraints to their geographic searches and partially because most search mechanisms treat all queries uniformly as sets of keywords regardless of their semantic orientation, i.e. refereed concepts.

As an example, consider the query *[pizza restaurant in Lisbon]* over Google and assume that the intention of the user issuing the query is to obtain pages about pizza restaurants that are located in Lisbon, Portugal. However, the page that Google retrieves first (as of March 2009) in the list of results is about a pizza restaurant in Lisbon New Jersey, although New Jersey does not appear in the query keywords. Likewise, for the query *[Athens city public schools]* the pages that Google returns (up to position 20) are about schools in Athens City Alabama rather than Athens (Greece), although Alabama is not specified as a search keyword. As both examples demonstrate, ignoring the geographic scope of web queries and the geographic orientation of web pages, results into favoring pages of increased PageRank popularity over location-relevant pages in the search engine results. Thus, retrieval effectives is significantly harmed for a large number of localizable queries.

Currently, there are two main strategies towards dealing with location-sensitive web requests. The first approach implies the annotation of the indexed pages with geospatial information and the equipment of search engines with geographic search options (e.g. Northern Light GeoSearch). In this direction, researchers explore the services of available gazetteers (i.e. geo-

graphical indices) in order to associate toponyms (i.e. geographic names) to their actual geographic coordinates in a map (Markowetz, et al., 2004) (Hill, 2000). Then, they store the geo-tagged pages in a spatial index against which geographic information retrieval is performed. The main drawbacks of this approach are: First, traditional gazetteers do not encode spatial relationships between places and as such their applicability to web retrieval tasks is limited. Most importantly, general-purpose search engines perform retrieval simply by exploring the matching keywords between documents and queries and without discriminating between topically and geographically relevant data sources.

The second strategy suggests processing both queries and query matching pages in order to identify the geographic orientation of web searches (Yu and Cai, 2007). Upon detecting the geographic scope of queries, researchers have proposed different functions for scoring the geographical vicinity between the query and the search results in order to enable geographically-sensitive web rankings (Martins et al., 2005) Again, such techniques, although useful, they require extensive computations on large volumes of data before deciphering the geographic intention of queries and thus they have a limited scalability in serving dynamic queries that come from different users with varying geographic intentions.

In this article, we address the problem of handling geographically-sensitive web searches from the perspective of a conventional search engine that performs keyword rather than spatial searches. In particular, we propose a novel framework that automatically builds a geo-referenced search engine index against which localizable queries are searched. The novelty of our framework lies on the fact that, instead of post-processing the indexed documents in order to derive their location entities (e.g. cities, landmarks), we introduce a geo-focused web crawler that automatically identifies pages of geographic orientation, before these are downloaded and processed by the engine's indexing modules. In brief, our crawler operates as follows. Given a seed list of URLs and a number of tools that are leveraged from the NLP community the crawler looks for location-specific entities in the page URLs and anchor text. For the identification of location entities, the crawler explores the data encoded in GeoWordNet (Buscaldi and Roso, 2008). Based on the location-descriptive elements in the page's structural content, the crawler directs the pages to a location-sensitive downloading module. This module resolves the geographic references of the identified location elements and organizes them into indexable hierarchical structures. Every structural hierarchy maintains URLs of pages whose geographic references pertain to the same place. Moreover, location-aware pages are linked to each other according to the proximity of their location names. Based on the above process, we end up with a geo-referenced search engine index against which location-sensitive queries can be searched. The experimental evaluation of our techniques indicates that our geo-focused crawler outperforms other focused crawling methods in terms of both coverage and accuracy. Moreover, our study reveals that our geo-referenced index guarantees that for localizable queries geographically-relevant pages are prioritized in the search results.

The rest of the article is organized as follows. We begin our discussion with an overview of relevant works. In section 3, we introduce our geo-focused crawler and we describe how it operates for populating a search engine index with geo-referenced data. In section 4, we discuss our approach towards building a geo-referenced search engine index. In particular, we describe how to organize location-specific web pages into geo-spatial hierarchical structures and we present a new metric for estimating the geographic-proximity between the query and the page location entities. In Section 5, we discuss the details of the experiments we carried out in which we assessed both the crawler's accuracy in detecting geography-specific web pages and the index's effectiveness in answering localizable queries. We also discuss obtained results. Finally, we conclude the article in Section 6.

2. RELATED WORK

Related work falls into two main categories, namely focused crawling and Geographic Information Retrieval (GIR). GIR deals with indexing, searching and retrieving geo-referenced information sources. Most of the works in this direction identify the geographical aspects of the web either by focusing on the physical location of web hosts (Borges et al., 2007) or by processing the web pages' content in order to extract toponyms (Smith and Mann, 2003) or other location-descriptive elements (Amitay et al., 2004). Ding et al. (2000) adopted a gazetteer-based approach and proposed an algorithm that analyzes resource links and content in order to detect their geographical scope. To obtain spatial data from the web pages' contents, Silva et al. (2006) and Fu et al., (2005) rely on geographic ontologies. One such ontology is GeoWordNet (Buscaldi and Roso, 2008) that emerged after enriching WordNet (Fellbaum, 1998) toponyms with geographical coordinates. Besides the identification of geographically-oriented elements in the pages' contents, researchers have proposed various schemes for ranking search results for location-aware queries according to their geographical relevance (Yu and Cai, 2007).

Our study relates also to existing works on focusing web crawls to specific web content. The main idea of a focused crawler is to optimize the priority of the unvisited URLs so that pages about a topic of interest are downloaded earlier. In this respect, most of the proposed approaches are concerned with focusing web crawls on pages dealing with specific topics (Chakrabarti et al., 1999) (Chung and Clarke, 2002). Early focused crawling techniques (Bra et al., 1994) used supervised topic classifiers to control the way in which the crawler assigns the downloading priority to the unvisited pages. Later on, researchers (Cho et al., 1998) suggested considering simple page properties such as linkage, PageRank values and keywords for determining the downloading priority of URLs in the focused crawler's frontier. In a different approach, Chung et al. (2002) studied the distributed topic-oriented crawling, where every crawling node handles a specific set of topics and examines pages that relate to the topics at hand. For finding topically-relevant pages they employ a simple Naïve-Bayes classifier.

In the recent years, the exploitation of focused crawlers has been addressed in the context of geographically-oriented data. Exposto et al. (2005) studied distributed crawling by means of the geographical partition of the web and by considering the

multi-level partitioning of the reduced IP web link graph. Later, Gao et al. (2006) proposed a method for geographically focused collaborative crawling. Their crawling strategy considers features like the URL address of a page, content, anchor text of links, etc. to determine the way and the order in which unvisited URLs are listed in the crawler's queue so that geographically focused pages are retrieved.

Although, our study shares a common goal with existing works on geographically-focused crawls, our approach for identifying location-relevant web content is novel in the following. First, our crawler integrates the GeoWordNet ontology for deriving the location entities in the pages' structural content. This way, our method eliminates any prior need for training the crawler with a set of pages that are pre-classified in terms of their location denotations. Some preliminary results of our method, reported in (Kozanidis et al., 2009) suggest that our crawler, although completely unsupervised, it outperforms supervised crawling approaches. Our current study complements and significantly extends our previous work on geo-focused web crawls since we also study the implementation of a geo-referenced search engine index in which the downloaded pages are organized into geo-spatial hierarchies. Thus, the scope of our study goes beyond focused crawling per se and touches upon issues pertaining to indexing and searching the downloaded focused web sources. We believe that our current study complements our previous work and introduces a unified indexing framework for serving localizable queries.

3. GEO-FOCUSED WEB CRAWLING

To build our geo-focused crawler, there are two challenges that we need to address: how to make the crawler identify web sources of geographic orientation, and how to organize the unvisited URLs in the crawler's frontier so that pages of great relatedness to the concerned locations are retrieved first. In this section, we present our geo-focused crawler and we discuss in detail how we addressed each of the above challenges. In Section 3.1, we describe the process that the crawler follows for automatically determining whether an unvisited URL points to geography-relevant content or not. In Section 3.2, we present the downloading policy that our crawler assigns to every unvisited geo-focused URL so as to ensure that it will not waste resources trying to download pages of little geographic relevance. Before delving into the details of our geo-focused crawling approach, we schematically illustrate the crawler's architecture in Figure 1.

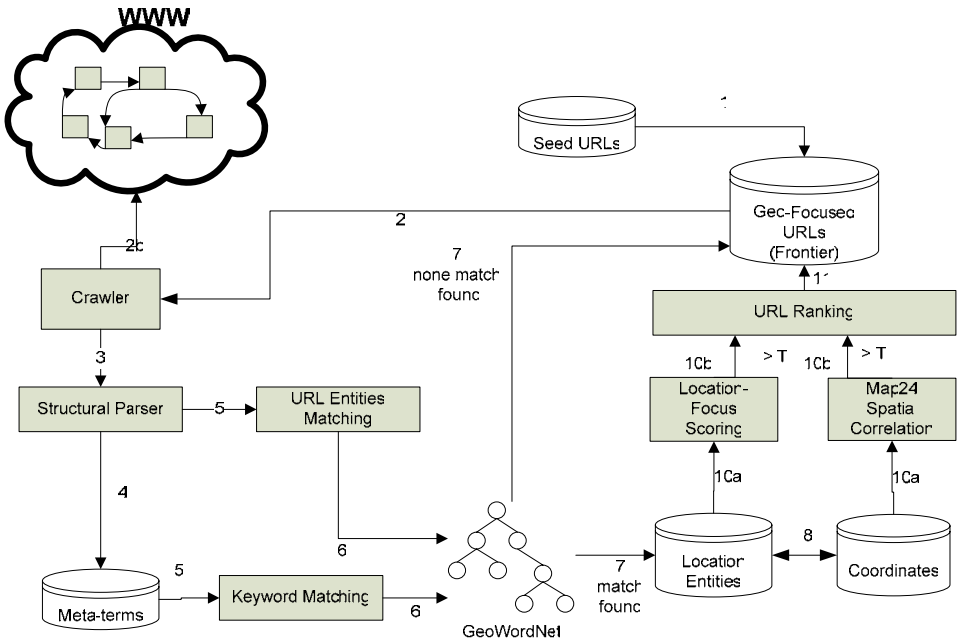

Figure 1: The geo-focused crawler's architecture and main components.

As the figure shows, the main components that our crawler integrates are: (i) a structural parser that extracts meta-terms from the page URLs and anchor text, (ii) the GeoWordNet ontology against which meta-terms that correspond to location entities are identified, (iii) the MAP24 tool that estimates the spatial distance between the location entities' coordinates and (iv) a ranking function that organizes the geo-focused URLs in the crawler's frontier in terms of their probability of guiding the crawler to highly focused geography pages.

3.1 Identifying Geography-Focused URLs

In the course of our study, we relied on a general-purpose web crawler that we parameterized in order to focus its web walkthroughs on geographically specific data. In particular, we integrated to a generic crawler a URL and anchor text parser in order to extract lexical elements from the pages' structural content and we used GeoWordNet as the crawler's backbone resource against which to identify which of the extracted meta-terms correspond to location entities. GeoWordNet contains a subset of WordNet synsets that correspond to geographical entities and which are inter-connected via hierarchical semantic relations. In addition, every GeoWordNet location entity is annotated with its corresponding geographical coordinates.

Given a seed list of URLs, the crawler needs to identify which of these correspond to pages of geographic orientation and thus they should be visited. To judge that, the crawler incorporates a structural parser that looks for the presence of location entities in the page URL and the anchor text of the page links. To identify location entities in the URL, the parser simply processes the admin-c section of the whois entry of a URL, since in most cases this section corresponds exactly to the location for which the information in the page is relevant (Markowetz et al., 2004). In addition, to detect location entities in the anchor text of a link in a page, the parser operates within a sliding window of 50 tokens surrounding the link in order to extract the lexical elements around it. To attest which of the terms in the page URL and anchor text represent location entities, we rely on the data encoded in GeoWordNet. The basic steps that the crawler follows to judge if a page is geographically-focused are illustrated in Figure 2.

```
Input: seed list of URLs (U), parser (P), GeoWordNet (GWN)
Output: annotated URLs with geographic orientation G(U)
For each URL u in U do
   Use parser P to identify meta-terms
   /*detect location entiries*/
        For each meta-term t in u do
           Query GWN using t
           If found
              Add t(u) in G(u)
           end
        end
end
```

Figure 2: Identifying URLs of geographic orientation.

The intuition behind applying structural parsing to the URLs in the crawler's seed list is that pages containing location entities in their URLs and anchor text links, have some geographic orientation and as such they should be visited by the crawler. Based on the above steps, the crawler filters its seed list and removes URLs of non-geographic orientation. The remaining URLs, denoted as G(U) are considered to be geographically-focused and are those on which the crawler focuses its web visits.

Having selected the seed URLs on which the crawler's web walkthroughs should focus, the next step is to organize the geographically-oriented URLs in the crawler's frontier. URLs' organization practically translates into ordering URLs according to their probability of guiding the crawler to other location-relevant pages. Next, we present our approach towards ordering unvisited URLs in the crawler's queue so as to ensure crawls of maximal coverage.

3.2 Ordering URLs in the Crawler's Frontier

A key element in all focused crawling applications is ordering the unvisited URLs in the crawler's frontier in a way that reduces the probability that the crawler will visit irrelevant sources. To account for that, we have integrated in our geo-focused crawler a probabilistic algorithm that estimates for every URL in the seed list the probability that it will guide the crawler to geographically-relevant pages. In addition, our algorithm operates upon the intuition that the links contained in a page with high geographic-relevance have increased probability of guiding the crawler to other geographically oriented pages. To derive such probabilities, we designed our algorithm based on the following dual assumption. The stronger the spatial correlation is between the identified URL location entities, the increased the probability that the URL points to geographic content. Moreover, the more location entities are identified in the anchor text of a link, the greater the probability that the link's visitation will lead the crawler to other geographically-oriented pages.

To estimate the degree of spatial correlation between the location entities identified for a seed URL, we proceed as follows. We map the identified location entities to their corresponding GeoWordNet nodes and we compute the distance of their coordinates, using the Map 24 AJAX API 1.28 (Map 24). Considering that the shortest the distance between two locations the increased their spatial correlation, we compute the average distance between the URL location entities and we apply the following formula:

$$\text{Spatial Correlation}(l_i, l_j) = 1 - \text{Avg.Distance}(l_i, l_j) \quad (1)$$

to derive the average spatial correlation between pairs of location entities (denoted as *l*) in the meta-terms of a URL. We then normalize average values so that these range between 1 (indicating high spatial correlation) and 0 (indicating no spatial correlation) and we rely on them for inferring the probability that the considered URL points to geographically-relevant content. This is done by investigating how the average spatial correlation of the URL location entities is skewed towards 1. Intuitively, a

highly skewed spatial correlation suggests that the URL has a clear geographic orientation and thus it should be retrieved. Our intuition is verified in Figure 3 where we show the average spatial correlation of 1,000 random URLs listed under each of the Dmoz top level categories[1]. As the figure illustrates, the spatial correlation of URLs listed under the Dmoz topic Regional is much more skewed towards rank one compared to the spatial correlation of the URLs listed under non-geography Dmoz topics. This practically suggests that pages categorized under geography-relevant topics contain more location entities compared to pages listed under other topical clusters.

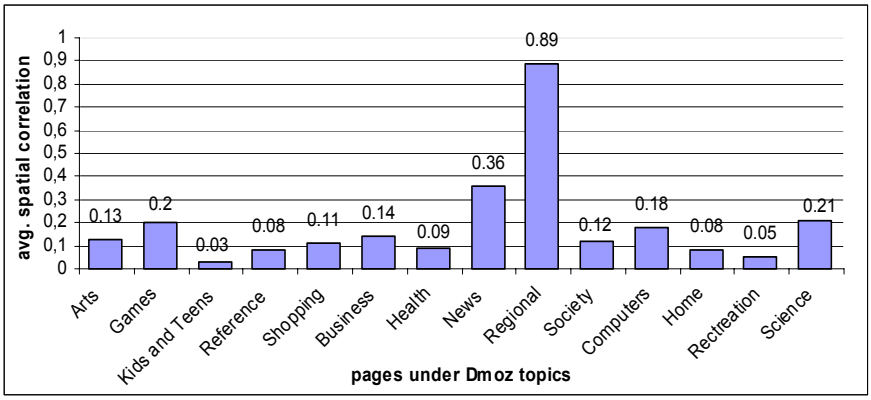

Figure 3. Spatial correlation among pages listed under different Dmoz topics

The estimated spatial correlation of the meta-terms in an unvisited URL indicates the probability that the URL's visitation will guide the crawler to a geography-relevant page. Nevertheless, spatial correlation is not sufficient per se for ordering the URLs in the crawler's frontier since it does not account the probability that the visitation of a geography-oriented URL will lead the crawler to other geography-focused pages.

To fill this void, we take a step further and we introduce a metric that computes the probability with which every geography-focused URL points to other geography-relevant pages. For our computations, we rely on the distribution of location entities in the anchor text of the links that the page URL contains. Recall that while processing anchor text, we have already derived the location entities that are contained in it. Our intuition is that the more location entities the anchor text of a link contains, the more likely it is that following this link will guide the crawler to a geographically oriented page. Formally, to quantify the probability that a link (u) in a geography-relevant page points to some other geography-relevant page, we estimate the fraction of meta-terms in the anchor text of (u) that correspond to toponyms in GeoWordNet and we derive the location focus of (u) as:

$$\text{Location Focus}(u) = \frac{|\#\text{of location entities in anchor}(u)|}{|\#\text{of terms in anchor}(u)|} \qquad (2)$$

Where Location Focus(u) indicates the probability that the visitation of (u) will result into the retrieval of a geography-relevant page. Location Focus scores are normalized, taking values between 0 and 1; with 0 indicating that the visitation of (u) is improbable to guide the crawler to a geography-relevant page and 1 indicating that the visitation of (u) is highly probable to guide the crawler to a geography-relevant page. In simple words, the increased the fraction of toponyms in the anchor text of a link, the greater the probability that this link points to a geography-oriented page.

For far, we have presented a metric for quantifying the probability with which every seed URL has a geographic orientation (cf. Spatial Correlation) and we have also introduced a measure for estimating the probability that the visitation of a geographically-oriented URL will guide the crawler to other geography-focused pages (cf. Location Focus). Now we turn our attention on how we can combine the two metrics in order to rank the URLs in the crawler's frontier so that the URL with the highest geographic orientation and the highest probability of pointing to other geography-focused pages is visited earlier. To rank the seed URLs in the crawler's frontier according to their probability of focusing web visits to highly geographic-relevant content, we rely on the URLs' Spatial Correlation and Location Focus values and we compute their downloading priority as:

$$\text{Rank}(u) = \text{Avg.Spatial Correlation}(l, u) + \text{Location Focus}(u, l) \qquad (3)$$

Where Rank (u) indicates the priority score assigned to every seed URL, so that the increased the Rank value of a URL (u) the greater downloading priority it has. Based on the above formula, we estimate for every URL in the crawler's seed list its download priority and we order them in descending priority values. This way, the crawler starts its web visits from the URL with the highest priority. Figure 4, summarizes the steps of our algorithm for ordering geographically-specific URLs in the crawler's frontier.

[1] The Dmoz top level category World was not considered since it contains multilingual content.

Having decided on the visitation order of the seed URLs, the crawler starts its web visits. As the crawler comes across new links in the contents of the geographically-focused pages that it retrieves, our algorithm examines the anchor text of these links and estimates for every link the probability (e.g. Location Focus) that it points to a location-relevant page. Links with some probability of pointing to location-relevant content are added in the crawler's frontier so that their pages are retrieved in future crawls. The crawling priority of the newly added links (i.e. the ordering of the URLs in the crawler's frontier) is determined by their Rank(u) values, which are re-computed as new URLs are added in the frontier. By re-ordering the URLs in the crawler's queue, we ensure that every web visit remains focused on location-specific content and that the crawler's frontier gets updated with new URLs that point to geographically-specific content.

```
Input: G(U), GWN, Map24 Resource
Output: ordered URLs in the crawler's frontier
For each URL u in U do
   Extract all location entities t from u
   For each t in u do
      Query GWN using t
      Retrieve coordinates of t
      Add coordinates to t, c(t)
   end
   For all c(t) in u do
      Compute paired c(t) distance using Map24
      Compute avg. distance of all c(T) in u
      Compute avg. spatial correlation of c(T) in u
      Return
      end
end
Extract anchor text for all links in u L(u)
For every link l in L(u) do
      Compute LocationFocus (u)
      Return
      end
end
For each u in frontier do
      Compute Rank (u)
end
Return URLs ordered by Rank (u) values
```

Figure 4: Ordering URLs in the focused crawler's frontier.

4. TOWARDS A GEO-REFERENCED SEARCH ENGINE INDEX

So far, we have presented our geo-focused crawler and we have described the algorithm that the crawler integrates for organizing URLs in the frontier, so as to ensure successful and affordable geographically-specific crawls. We now turn our discussion on how we process the crawled pages in order to index them into a geo-referenced repository of data sources.

Crawled web pages are directed to a downloading module that retrieves their contextual data and employs the vector space Geo-thematic model (Cai, 2002) to represent every page as a term vector associated with a geographic footprint. The term vector of a page contains the location entities that appear in the page contextual elements, while the geographic footprint of a page indicates the geographical coordinate space in which the page location entities span. To build the term vector of a page we start by processing the page's textual content in order to identify location entities. In this respect, we apply HTML parsing (to remove markup), tokenization, Part-of-Speech tagging and lemmatization to the page's content. Then, we rely on the lemmatized terms that are morphologically annotated as Proper Nouns and we look them up in GeoWordNet. Named entities identified in GeoWordNet are characterized as location entities and are the ones that will be considered for building the term vector of the page.

The next step is to derive the location reference(s) of the identified location entities, in order to be able to infer the geographical area(s) that the page discusses and thus be able to index it under the appropriate geographic hierarchy (-ies). For deducing the location reference(s) of the identified geography entities, we map the location keywords of a page to their corresponding GeoWordNet nodes and we explore their geographic relations. This is done by investigating whether there are overlapping named entities in the definitions of location entities. For example, the location entity *Rhode Island*: defined as a state in New England and the location entity *Brown University*: defined as a university in Rhode Island, are deemed as geographically related since the name of the second location entity appears in the definition sentence of the first. Location entities that are found to be geographically related are grouped together under a common location reference. To verbalize the geographic reference of a page, we use the name of the location entity under which the page location entities are organized. On the other hand, in case the location entities in a page are not geographically related in GeoWordNet, we speculate that the page refers to many geographic areas and we use the names of all identified location entities to denote the geographic references of the page. Figure 5(a) shows an example of location entities that are geographically related, i.e. refer to the same geographic area, while

Figure 5(b) shows an example of location entities that refer to distinct geographic areas. As the examples suggest, a page that contains the location entities of Figure 5(a) discusses a single geographic area, i.e. *Virginia*, while a page that contains the location entities of Figure 5(b) discusses multiple geographic areas. Therefore, the location reference of the page containing the entities of Figure 5(a) is *Virginia*, while the location references of the page containing the entities of Figure (b) are *Germany*, *Berkeley* and *Brown University*.

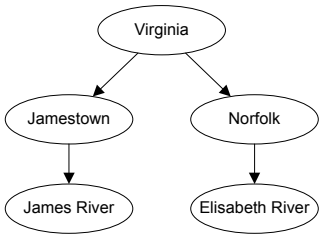

Figure 5(a). Example of location entities with a single geographic reference.

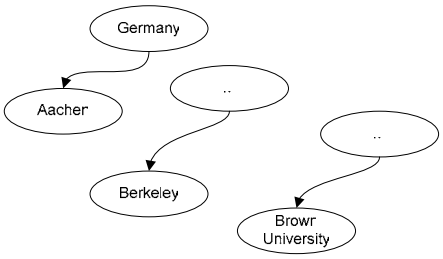

Figure 5(b). Example of location entities with multiple geographic references.

Having extracted the location entities from the downloaded pages' contents and having also deduced the geographic area(s) to which every page refers, we now turn our attention on how to associate location entities in the page vector with an appropriate weight that indicates how much the location entity represents the page's contents. For our estimations, we use the TF.IDF weighting scheme (Robertson & Sparck-Jones, 1976) that computes the importance of every location entity for the page in which it appears. At the end of this process, we derive the term vector of every focused page that the crawler has downloaded.

All terms represented in the downloaded pages' vectors constitute the geographic keywords of our index. The last step we take before structuring our index is to associate the geographic keywords in every page with their corresponding geographic footprint. In this respect, we rely on the geographic coordinates that GeoWordNet encodes for location entities and we estimate the spatial distance between the geography keywords in every page. Spatial distance is again computed via the Map24 tool. Then, we rely on the pair of location keywords in a page that share the maximum distance and use their geographic coordinates to specify the geographic footprint of the respective page. This way, in case a page has a single geographic reference (i.e. its location entities are geographically related) the geographic coordinates of that location reference will constitute the geographic footprint of the page. On the other hand, in case a page has multiple geographic references, the geographic coordinates of the most distant location references will be used as the page's geographic footprint. Based on the above process (schematically illustrated in Figure 6) we represent every page the geo-focused crawler has downloaded as a vector of weighted geography terms accompanied by a geographic footprint of the page's location reference(s).

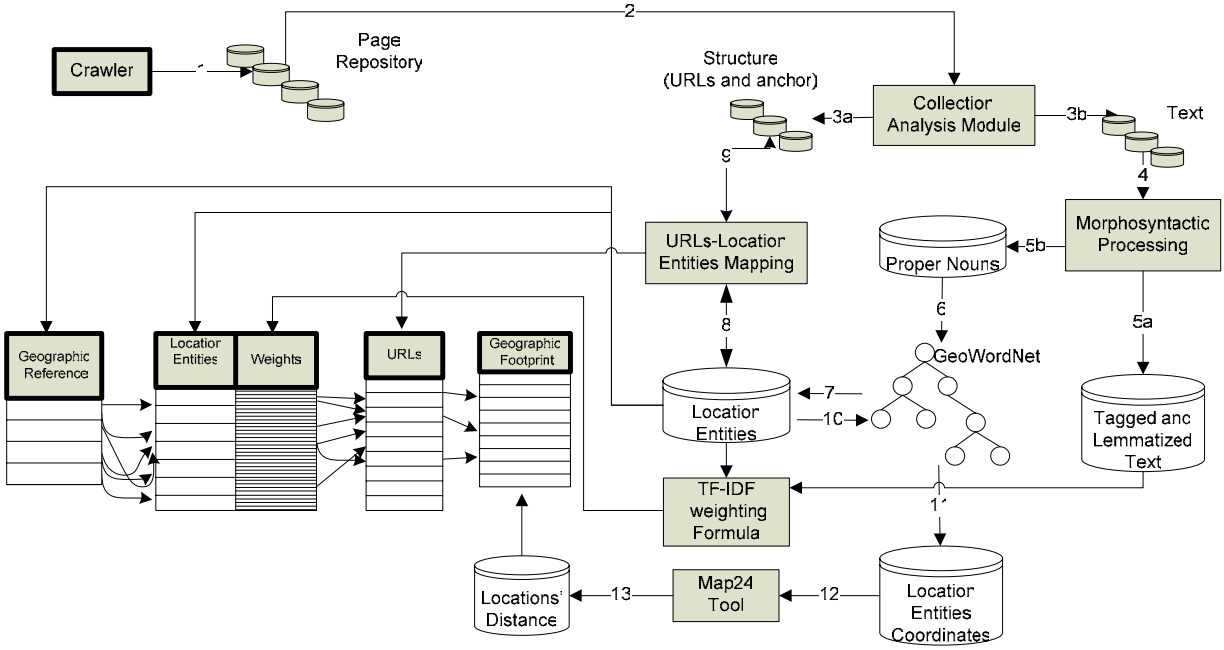

Figure 6: Geo-referenced indexing process.

As a final step, we organize the data stored in the geographic inverted index into a geo spatial hierarchical structure. In this respect, we store the geographic footprints of the crawled URLs (previously computed) into an R-Tree structure and we associate them with their corresponding indexed location entities. We also associate the latter with the URLs of the indexed pages in which they appear. Figue 7, illustrates the structure of our geo-spatial index.

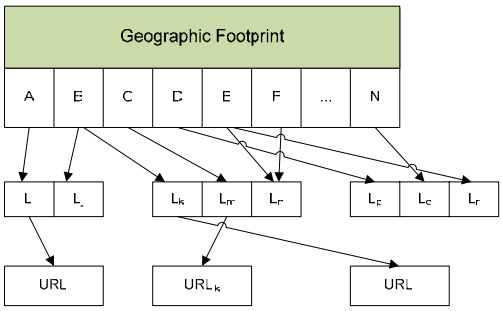

Figure 7: The R-tree geo-referenced index.

As Figure 7 shows, every page is indexed under the geographic footprint(s) of the location entities that it contains. This way, a page with a single geographic reference is indexed under a single footprint tree node (recall that geographic footprints correspond to coordinates of type (x, y)), whereas a page with multiple geographic references is indexed under all its corresponding geographic footprints.

At the end of indexing, we maintain the downloaded pages into two distinct yet complementary indexing structures. That is, we build a geo-thematic inverted index (cf. Figure 6) against which localizable keyword queries can be searched and we also build a geographic hierarchical index (cf. Figure7) against which purely geographic searches can be performed. Although our model suggests implementing two indices for storing the geo-focused pages that the crawler downloads, nevertheless there is not significant overhead associated with indexing, given that pages need only be processed once. Nevertheless, in case of space constraints one could consider maintaining only one of the two index structures depending on the nature of searches to be executed towards the indexed documents.

4.1 Searching Geographic Queries against the Geo-Referenced Index

Given our geo-referenced index of web data, the next step is to describe the process of searching localizable queries against the index and retrieving results that are both geographically and thematically relevant to the query intention. The first issue we need to tackle in this respect is how to determine whether a keyword query has a geographic orientation. In simple words, we need a method to automatically detect whether the intention of the query is to retrieve geography-specific pages or not. In the course of our study, the straightforward method that we adopted for identifying whether a query intends the retrieval of geography-oriented results is presented in Figure 8. Specifically, we deem the intention of the query to be geographic in case the query contains a geography keyword, determined after mapping query terms against GeoWordNet.

```
Input: search queries (q), GeoWordNet (GWN)
Output: geographic queries G(q)
For each term ∈ q do
        Query GWN using t
        If found
           Return q as G(q)
        end
end
```

Figure 8: Identifying queries of geographic orientation.

At this point, we should note that there are other approaches towards identifying localizable queries (e.g. that proposed by Welch and Cho, 2008) that one could employ. Nevertheless, investigating the most suitable method for detecting the geographic orientation of web queries goes beyond the scope of our study, which concentrates on how to automatically construct a geo-referenced search engine index.

Having determined queries of geographic intention, our next step is to retrieve pages that are most relevant to both the geographic scope and the thematic orientation of the queries. In our work, this is done in three steps: first we use the inverted geographic index to find all the documents that contain the terms in the query. For every document found, we compute a query-document similarity value that indicates the degree to which every query-matching page relates to the query intention.

To derive the query-document similarity scores, we rely on the TF.IDF vector space model to represent both the document *d* and the query *q* terms and we employ the cosine distance metric, which gives their similarity based on:

$$\mathrm{Sim}(q, d) = \frac{\sum_{j=1}^{t} w_{qj} * w_{dij}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 * \sum_{j=1}^{t} (w_{dij})^2}} \qquad (4)$$

where w_{qj} and w_{dij} represent the index weights of the query and the matching document terms. Similarity values indicate the degree to which each of the matching documents relates to the query.

Second, once we have the relevant documents, we compute the geographic similarity between the documents' geographic footprints and the query's geographic coordinates. In this respect, we rely on the GeoWordNet nodes that match the query geography keywords and we extract the geographic coordinates of the query location entities, denoted as $(x,y) \in q$. We also explore the geographic footprints of the query relevant documents, denoted as $(x', y') \in d$ and which are stored in the spatial index. Then, we derive the query-document geographic similarity by measuring the inverse of the Euclidian distance between their coordinates, as:

$$\mathrm{SimCoord}(q, d) = \frac{1}{\mathrm{dist}(q, d)} \qquad (5)$$

Geographic similarity values indicate the geographic proximity between the query and the document location entities. Finally, we combine the similarity values between the keywords and the geographic coordinates of the query and the matching documents, to calculate the rank score of each document with respect to the query and we return to the user the ordered list of result pages. Formally, the final ranking score of every matching document is a function of its term and geographic similarity to the query, given by:

$$\mathrm{Rank}(d, q) = \mathrm{Sim}(q, d) + \mathrm{SimCoord}(q, d) \qquad (6)$$

Based on the above values, we order retrieved pages for a geography-oriented query so that pages of high thematic and geographic similarity to the query are prioritized in the query results.

5. EXPERIMENTAL EVALUATION

In this section, we validate the effectiveness of our proposed method towards building a geo-referenced search engine index against which localizable queries can be searched. For our evaluation, we carried out two distinct yet complementary experiments. In the first experiment, presented in Section 5.1 we assess our geo-focused crawler's performance in effectively discovering and downloading geographically-oriented pages. In the second experiment, discussed in Section 5.2 we evaluate the performance of our geo-referenced index in serving geography-specific keyword requests.

5.1 Geo-Focused Crawling Performance

To evaluate the performance of our geo-focused crawler, we rely on the following measures: accuracy, geographic coverage and quality of crawls. Specifically, to assess the crawler's accuracy we measure the absolute acquisition rate of geographic pages in order to see how effective our crawler is in targeting its web visits on geographic content. In addition, to evaluate coverage we assess the crawler's effectiveness in identifying pages of different geographic orientations. Finally, to estimate the quality of crawls we compared the performance of our geo-focused crawler to the performance of other focused crawling strategies and we comparatively analyze obtained results.

To begin with our experiments, we compiled a seed list of URLs from which the crawler would start its web walkthroughs. In selecting the seed URLs, we relied on the pages organized under the Dmoz categories [*Regional*: *North America*: *United States*] out of which we picked a total set of 10 random URLs and we used them for compiling the crawler's seed list. Recall that we have previously examined the pages listed under the Dmoz topic *Regional* and found that they contain location entities. Therefore, by picking those page URLs as the crawler's seed data, we ensure that the crawler's first visit will pertain to geographic content. Based on these 10 seed URLs, we run our crawler for a period of one week during which the crawler downloaded a total set of 2.5 million pages as geography-specific data sources. Note that the crawler's visits adhere to the depth-fist paradigm, which implies that the crawler follows all the links from the first link of the starting page before proceeding with the next seed URL. However, in the course of our experiment we limited the depth of the crawler's visits to level 5, i.e. starting form a page the crawler examined its internal links up to 5 levels down in their connectivity graph. This limitation was imposed in order not to waste the crawler's resources before the latter exhausts the list of seed URLs.

To estimate our crawler's accuracy in focusing its visits on geography-relevant pages, we essentially measured the fraction of pages that the crawler downloaded as geographic from all the geography-relevant pages that it examined during its web walkthroughs. However, to be able to judge which of the pages that the crawler visited are indeed geography-oriented, we built a

utility index in which the crawler directed the URLs it encountered but did not download, because it considered them as non-geography relevant. In simple words, every URL that the crawler examined but not downloaded was recorded in the utility index, whereas every URL that the crawler downloaded was recorded in the geographic index. We then relied on the two sets of URLs, namely *visited-but-not-downloaded* and *visited-and-downloaded* and we supplied them as input to a generic crawler in order to retrieve their contents. Thereafter, we processed their contents (as previously described) in order to identify location entities among their contextual elements. Page URLs containing geographic keywords (identified against GeoWordNet) were deemed as geography-relevant, while page URLs without any geographic keywords in their contents were deemed as non-geography. Thereafter, we relied on the set of pages that the crawler downloaded as geographic (i.e. pages stored in the geography index) and the set of pages that have a geographic-orientation (i.e. geography pages stored either in the utility or the geography index) and we estimated our crawler's accuracy in retrieving geographically-relevant data as:

$$\text{Accuracy}(c) = \frac{\left|P_{\text{geography-downloaded}}\right|}{\left|P_{\text{geography-visited}}\right|} \qquad (7)$$

That is, we formally quantify the crawler's accuracy as the fraction of geographic pages that it downloaded from all the geography pages that it visited. Obtained results summarize to the following: our geo-focused crawler visited a total set of 14.1 million pages of which 2.8 million are geography-relevant. From those 2.8 million geography pages, the crawler managed to correctly identify and thus download 2.5 million pages. Results, reported in Table 1, indicate that our crawler has overall 89.28% accuracy in identifying geographically-relevant data in its web visits. This finding suggests that given a geographic ontology and simple parsing techniques, our method in quite effective in focusing web crawls on location specific content without any prior need for building training examples.

Table 1. Geo-focused crawling accuracy.

Geography visited pages	2.8 million
Geography downloaded pages	2.5 million
Geo-focused crawling accuracy	89.28%

Besides testing the crawler's accuracy, we also estimated its geographic coverage, i.e. the different location references on witch the crawler can focus its web visits. This in order to assess whether the crawler's accuracy is equally attributed to all location entities encoded in GeoWordNet or rather the crawler's focus is limited within a small set of location names. To quantify the crawler's geographic coverage, we measured the fraction of distinct geographic entities in the downloaded pages' contents, formally given by:

$$\text{GeoCoverage}(c) = \frac{\left|E_{\text{distinct}}\right|}{\left|E_{\text{total}}\right|} \qquad (8)$$

Where $E_{distinct}$ denotes the number of unique location entities and E_{total} denotes the total number of location entities that appear in the contents of the downloaded pages. Results, reported in Table 2, show that our geo-focused crawler can successfully retrieve sources pertaining to distinct geographical areas, ensuring thus complete and focused web crawls.

Table 2. Crawler's coverage of location entities

Number of all location entities identified	1,265
Number of distinct location entities	1,029
Crawler's geographic coverage	81.34%

Another factor that we assessed during our crawling experiment concerns the quality of crawls performed by our geo-focused crawler. In this respect, we compared the accuracy of our unsupervised focused crawler to the accuracy of supervised crawling techniques. For this experiment, we built a Naïve Bayes classifier (Duda and Hart, 1973) and we integrated it into a generic web crawler so as to assist the latter in focusing its web visits on specific web content. The reason for selecting to integrate a Bayesian classifier into a generic crawler is because Naïve Bayes classifiers are both effective and accurate for web scale processing. To build the classifier, we relied on the 1,000 random URLs listed under the top level Dmoz topics, previously examined (cf. Section 3.2). For each topic's URLs we explored the distribution of location elements in their corresponding pages' contents. Recall that we have already processed those pages and extracted their location entities based on GeoWordNet by following the steps presented in Section 3.2. Based on the above data, we estimated the geographic focus of every page URL as:

$$\text{Geo Focus}(p) = \frac{\left|\#\text{of location entities in p}\right|}{\left|\#\text{ of terms in p}\right|} \qquad (9)$$

GeoFocus values indicate the fraction of location entities among the pages' contextual elements, so that the more location terms a page contains the increased its GeoFocus value. We then normalized GeoFocus scores so that they range between 0

and 1; with zero values indicating the absence of location entities in the page's content and values close to one indicating the presence of many location entities in the page's contents. Pages of GeoFocus values above threshold T (T=0.5) are deemed as on-topic (i.e. geography) documents, whereas pages with GeoFocus values below T are deemed as off-topic (i.e. non-geography) documents. After creating the set of geography and non-geography pages we trained the classifier so that it is able to judge for a new page it comes across whether it is on-topic or not.

Having trained the classifier, we integrated it in a crawling module, which we call baseline focused crawler. To perform our comparative evaluation, we run the baseline focused crawler using the same seed list of URLs that our geo-focused crawler explored in the previous experiment. Note that the execution period for the baseline focused crawler was again set to one week and the depth of the crawling path was again limited to five levels down the internal links' hierarchy. For every page visited by the baseline crawler, we recorded in the utility index the page URL and its classification label, i.e. geography or non-geography. Based on the set of geography pages that the baseline crawler visited and the set of geography pages that the baseline focused crawler downloaded, we employed our accuracy metric (cf. formula 7) in order to estimate the harvest rate of the baseline geo-focused crawler. Obtained results indicate that the baseline focused crawler visited a total of 12.3 million pages, of which 2.1 million have a geographic orientation. From these geography-relevant pages, it correctly identified and downloaded nearly 1.5 million pages. Finally, we compared the accuracy of the baseline focused crawler in retrieving geography-specific pages to the accuracy of our geo-focused crawler. Table 3 reports the comparison results.

Table 3. Comparison results

Geo-focused crawling accuracy	89.28%
Classification-aware crawling accuracy	71.25%

Results indicate that our geo-focused crawler clearly outperforms the baseline focused crawler. In particular, we observe that from all the geography-relevant pages that each of the crawlers' visited our module managed to accurately identify as geographic and thus download more pages than the baseline focused crawler did. This, coupled with the fact that our geo-focused crawler does not need to undergo a training phase imply the potential of our geo-focused crawler towards retrieving geographically-specific web data. Although we are aware of the fact that we cannot directly compare the performance of two crawlers, since we cannot guarantee that their visits will concentrate on the exact same set of pages, nevertheless by exploring the fraction of relevant documents downloaded from all the relevant documents visited, we believe that we can derive some valuable conclusions concerning the quality of crawls, where the latter is defined as the ability of the crawler to locate pages of interest.

5.2 Location-Sensitive Retrieval Performance

Besides evaluating our geo-focused crawler's performance in automatically identifying and downloading geography-specific web pages, we also assessed the effectiveness of our geo-referenced search engine index in retrieving relevant pages for geography-oriented search queries. In order to perform realistic experiments, we implemented a search engine prototype and relied on a real set of pages that both our geo-focused and the baseline focused crawler have visited. More specifically, we used the 14.1 million pages visited by our geo-focused crawler and the 12.3 million pages visited by the baseline focused crawler, we removed duplicates (based on their URL addresses) and we retained a total set of 16.6 million distinct pages, of which 3.7 million have a geographic orientation. Recall that during our crawling experiment (cf. Section 5.1) we examined the contents of all the pages visited by both crawlers and based on the presence of location entities within their contextual elements, we discriminated them into geography-specific and non-geography pages. Then, we used those 16.6 million pages as the data to be indexed in our engine prototype. In this respect, we represented the contents of the experimental pages according to the vector space model and we stored them into our geo-focused inverted index. Note that for non-geography pages, the index stores the pages' keyword terms rather than location entities and leaves their corresponding geographic footprint table empty. On the other hand, for geography pages the index maintains information about both their location entities and geographic footprints, as described in Section 3.2.

Having built our prototype geo-referenced index, the next step was to submit location-sensitive queries to our engine in order to evaluate the index effectiveness in serving geography-specific requests. In selecting our experimental queries, we relied on the data encoded in GeoWordNet from which we extracted location entities that span to distinct geographic areas. Specifically, to select our geographically-distinct location queries, we examined the location references among all the GeoWordNet entities, based on the approach described in Section 3.2., and we retained those with unique geographic references. From all the location entities encoded in GeoWordNet, 380 entities have unique geographic references. Therefore, we utilized those 380 location terms as the keywords of our geography-specific queries. We then submitted the 380 experimental queries to our prototype engine, which looked them up against the geo-referenced index and retrieved relevant pages. As query-relevant pages the engine retrieved all the indexed documents that contain the query keywords. Nevertheless, in the course of our study we deem a page to be relevant to a given experimental query if the following apply both: (i) the page contains the query terms and (ii) the page has been characterized as geography-specific during pre-processing. Retrieved results are ordered by our Ranking function (cf. equation 6) that prioritizes pages of increased topical and geographic similarity among the search results of the respective queries.

To evaluate the effectiveness of our geo-referenced index in answering location-sensitive queries, we computed the 11-point interpolated average precision of the retrieved results. That is, for each of the 380 geography queries, we measured the arithmetic mean of the interpolated precision at the 11 recall levels, i.e. 0.0, 0.1, …, 1.0. Note that the interpolated precision at a certain recall level r is formally defined as the highest precision found for any recall level $r' \geq r$, given by:

$$P_{interp} = \max_{r' \geq r} p(r') \qquad (10)$$

Obtained results, reported in Figure 9 indicate that both our geo-referenced index and the geo-thematic similarity ranking function are quite effective in prioritizing relevant results for localizable queries. In particular, we observe that the precision of retrieved results is considerably high at recall level 0.1; demonstrating that the first ten pages returned by our index in response to geographic queries are both relevant to the query keywords and the query geographic orientations. As Figure 9 shows precision drops at increasing recall levels, which comes naturally if we consider that as more pages are considered the increased the probability that there are query-matching but no geography-oriented pages among them. Nevertheless, the precision drop is mostly pronounced beyond recall level 0.4 which implies that the first forty returned pages are on average highly relevant to both the intention and the geographic orientation of the respective queries. Based on our findings and the observations of many researches that web information seekers rarely examine the retrieved results beyond point ten, suggest that our framework is quite effective in satisfying the users' geographic requests.

In overall, results indicate that our proposed retrieval method is quite effective not only towards matching geography queries against the geographic index by also towards identifying within the query matching pages the ones that have clear geographic orientations and prioritize them in the query results. Therefore, we believe that our method can be fruitfully explored by the research community that studies ways of effectively answering location-specific search requests.

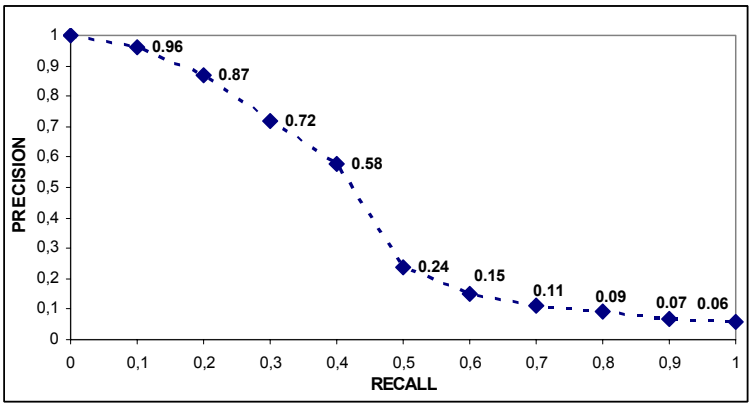

Figure 9: Averaged 11-point precision/recall graph across the 380 location-sensitive queries.

As a final note, we should mention that the only limitation our method imposes is the fact that its performance depends upon the GeoWordNet data. In other words, since our method relies entirely on GeoWordNet for identifying geographic entities it is apparent that location manes currently absent from GeoWordNet cannot be identified by our crawler and thus cannot be properly represented in our geo-referenced index. Nevertheless, considering that there are other geographic ontologies available, such as those proposed in (Arpinar et al., 2006), (Fu et al., 2005) we can rely on them to complement or substitute the GeoWordNet ontology in our implementation. As a matter of fact, we are currently experimenting with different geographic resources so as to ensure that both our crawler and indexer explore all the available geographic information sources towards facilitating geography-specific web searches.

6. CONCLUDING REMARKS

In this paper, we introduced a novel approach towards implementing a geo-focused web crawler and we presented a method for building a geo-referenced search engine index. Our crawler identifies pages of geographic orientation simply by exploring the presence of location entities in the page URLs and anchor text. In this direction, the crawler consults the data encoded in GeoWordNet and employs a number of heuristics for deducing the pages and the order in which these should be retrieved. Moreover, we have presented a method for automatically building a geo-referenced web index that conventional search engines could employ for answering location-sensitive queries. In this respect, we have also introduced a hybrid formula for ordering search results according to both their topical and geographic proximity to the search query. The innovations of our work pertain to the following. First, our crawler automatically identifies the geographic focus of a page without any prior need for processing the page's contents. Moreover, our focused crawler runs completely unsupervised, diminishing thus computational overheads associated with building training examples for learning the crawler to detect its visitations' foci. In addition, the crawler directs the retrieved pages to a downloading module, which processes their contents and indexes them into location-aware hierarchies. The geo-spatial index that we introduced is quite effective in answering geographically-focused keyword

queries, thus taking of the users the burden of going over long list of search results before satisfying their search pursuits. The experimental evaluation of our framework demonstrates that our system is both accurate and effective in performing geo-thematic information retrieval over the web data. In the future we plan to improve our geo-focused crawler so that is accounts of dynamic pages and to extend our geo-referenced index so as to be able to able to answer location-sensitive queries of multiple geographic references.

7. REFERENCES

[1] Amitay, E., Har'El, N., Silvan, R., Soffer, A. (2004). Web-a-where: geotagging web content. In Proceedings of the 27th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR), United Kingdom, pp. 273-280.

[2] Arpinar, B.I., Sheth, A., Ramakrishnan, C., Usery, E.L., Azami, M. (2006). Geospatial ontology development and semantic analysis. In Transactions of Geographic Information Systems, vol. 10, no.4, pp. 551-575.

[3] Borges, K., Laender, A., Mederios, C., Davis, C. (2007). Discovering geographic locations in web pages using urban addresses. In Proceedings of the 4th International Workshop on Geographic Information Retrieval (GIR), Portugal, pp. 31-36.

[4] Buscaldi, D., Roso, P. (2008). Geo-WordNet: automatic georeferencing of WordNet. In Proceedings of the 6th International Language Resources and Evaluation Conference (LREC), Morocco, pp. 1255-1258.

[5] Cai, G. 2002. GeoVSM: An integrated retrieval model for geographic information. In Geographic Information Science, pp. 65-79.

[6] Chakrabarti, S., van den Berg, M., Dom, B. (2000). Focused crawling: a new approach to topic-specific web resources discovery. In Computer Networks, vol.31, no. 11-16, pp. 1623-1640.

[7] Cho, J., Garcia-Molina, H., Page, L. (1998). Efficient crawling through url ordering. In Computer Networks vol. 30, no.1-7, pp. 161-172.

[8] Chung, C., Clarke, C.L.A. (2002). Topic-oriented collaborative crawling. In Proceedings of the Information and Knowledge Management Conference (CIKM), Virginia, USA, pp. 34-42.

[9] Ding, J., Gravano, L., Shivakumar, N. (2000). Computing geographical scopes of web resources. In Proceedings of the Conference on Very large Databases (VLDB), pp. 545-556.

[10] D.Bra, P., Geert-Jan Houben, Y.K., Post, R. (1994). Information retrieval in distributed hypertexts. In Proceedings of the Recherche d'Information Assistee par Ordinateur Conference (RIAO), New York, USA, pp. 481-491.

[11] Duda, R.O., Hart, P.E. (1973). Pattern classification and scene analysis. Willey, New York.

[12] Exposto, J., Macedo, J., Pina, A., Alves, A., Rufino, J. (2005). Geographical partition for distributed web crawling. In Proceedings of the 2nd Workshop on Geographic Information Retrieval (GIR), Germany, pp. 55-60.

[13] Fellbaum, Ch. (Ed.). (1998). WordNet: An Electronic Lexical Database, MIT Press.

[14] Fu, G., Jones, C.R., Abdelmoty, A. (2005). Building a geographical ontology for intelligent spatial search on the Web. In Proceedings of the IASTED International Conference on Database and Applications, Austria, pp. 167-172.

[15] Gao, W., Lee, H.C., Miao, Y. (2006). Geographically focused collaborative crawling. In Proceedings of the 15th International World Wide Web Conference (WWW), Scotland, pp. 287-296.

[16] GeoWordNet. Available at: http://www.dsic.upv.es/ grupos/nle/downloads-new.html

[17] Hill, L. (2000). Core elements of digital gazetteers: placements, categories and footprints. Research and Advanced Technology of Digital Libraries. In proceedings of the 4th European Conference on Digital Libraries (ECDL), Portugal, pp. 280-290.

[18] Himmelstein, M. (2005). Local search: the internet is yellow pages. In Computer, v.38, n.2, pp. 26-34.

[19] Kozanidis, L., Stamou, S., Spiros, G. (2009). Focusing web crawls on location specific content. In Proceedings of the 6th International Conference on Web Information Systems and Technology (WebIST), Portugal.

[20] Map 24. Available at: http://developer.navteq.com/site/ global/zones/ms/downloads.jsp.

[21] Markowetz, A., Brinkhoff, T., Seeger, B. (2004). Geographic information retrieval. In Proceedings of the 3rd International Workshop on Web Dynamics, New York, pp. 116-125.

[22] Martins, B., Silva, M.J., Andrade, L. (2005). Indexing and ranking on Geo-IR systems. In Proceedings of the 2nd International Workshop on Geographic Information Retrieval (GIR), Germany, pp. 31-34.

[23] Robertson, S., Sparck-Jones, K. (1976). Relevance weighting of search terms. In Journal of the American Society for Information Science, Vol. 27, pp. 129-146.

[24] Silva, M.J., Martins, B., Chaves, M., Cardoso, N., Afonso, A.P. (2006). Adding geographic scopes to web resources. In Computers, Environment and Urban Systems, vol. 30, pp. 378-399.

[25] Smith, D., Mann, G. (2003). Bootstrapping toponyms classifiers. In Proceedings of the Human Language Technology Conference, North American Association for Computational Linguistics Workshop on Analysis of Geographic References, pp. 45-49.

[26] Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y.S., Ma, W,Y., Li, Y. (2005a). Detecting dominant locations from search queries. In Proceedings of the 28th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR), Brazil, pp, 424-431.

[27] Wang, C., Xie, X., Wang, L., Lu, Y.S., Ma, W,Y. (2005b). Detecting geographic locations from web sources. In Proceedings of the 2nd International Workshop on Geographic Information Retrieval (GIR), Germany, pp. 17-24.

[28] Welch, M., Cho, J. (2008). Automatically identifying localizable queries. In Proceedings of the 31st Annual ACM Conference on Research and Development in Information Retrieval (SIGIR), Singapore, pp. 507-514.

[29] Yu, B., Cai, G. (2007). A query-aware document ranking method for geographic information retrieval. In Proceedings of the 4th International Workshop on Geographic Information Retrieval (GIR), Portugal, pp. 49-54.

Author biographies

Lefteris Kozanidis is a PhD student at the Computer Engineering and Informatics Department of Patras University. His area of expertise is web crawling, databases and language processing with an emphasis on semantic indexing. He received his M.Sc. and B.A. degrees from the Computer Engineering and Informatics Department of Patras University, in 2005 and 2007 respectively. He has published the results of his research in several international conferences and journals.

Sofia Stamou is an adjunct lecturer at the Computer Engineering and Informatics Department of Patras University. Her area of expertise is text analysis and language processing with an emphasis on semantic analysis of textual data. Her research interests include web searching, personalization, ontologies, text mining and data organization. She received her Ph.D. and M.Sc. degrees from the Computer Engineering and Informatics Department of Patras University in 2002 and 2006 respectively and her B.A. in Philosophy from the University of Ioannina in 1999. She has served as a program committee member in several conferences that relate to the areas of expertise and she has published the results of her research in several international conferences and journals.