FOCUSING WEB CRAWLS ON LOCATION-SPECIFIC CONTENT

Lefteris Kozanidis, Sofia Stamou, George Spiros
Computer Engineering and Informatics Department, Patras University, Greece
{kozanid, stamou, spiros}@ceid.upatras.gr

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

Abstract: Retrieving relevant data for location-sensitive keyword queries is a challenging task that has so far been addressed as a problem of automatically determining the geographical orientation of web searches. Unfortunately, identifying localizable queries is not sufficient per se for performing successful location-sensitive searches, unless there exists a geo-referenced index of data sources against which localizable queries are searched. In this paper, we propose a novel approach towards the automatic construction of a geo-referenced search engine index. Our approach relies on a geo-focused crawler that incorporates a structural parser and uses GeoWordNet as a knowledge base in order to automatically deduce the geo-spatial information that is latent in the pages' contents. Based on 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.

1 INTRODUCTION

Locality is an important parameter in web search. According to the study of (Wang et al., 2005a) 14% of 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). 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 second (as of October 2008) in the list of results is about a pizza restaurant in Lisbon New Hampshire, although New Hampshire 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 favouring popular pages over location –relevant pages in the search engine results. Thus, retrieval effectives is harmed for a large number of 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. geographical 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 iden-

tify 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 paper, we address the problem of improving the geographically-oriented web searches from the perspective of a conventional search engine that performs keyword rather than spatial searches. In particular, we propose a technique that automatically builds a geo-referenced search engine index against which localizable queries are searched. The novelty of our approach 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 (either spatial or semantic) 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 rest of the paper is organized as follows. We start 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 the advantages of our crawler and we report some preliminary results.

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. 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).

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 that it integrates GeoWordNet for deriving the location entities that the crawler considers. 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.

3 GEO-FOCUSED 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 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 the hierarchical semantic relations. In addition, all location entities in GeoWordNet are annotated with their 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 1.

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.

```
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 entities*/
        For each meta-term t in u do
           Query GWN using t
           If found
              Add t(u) in G(u)
           end
        end
end
```

Figure 1: Identifying URLs of geographic orientation.

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.

3.1 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 [1-avg. distance]

formula to derive their average spatial correlation. We then normalize average values so that these range between 1 (indicating high correlation) and 0 (indicating no 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.

On the other hand, to estimate the probability that a geographically-focused URL will guide the crawler to other geographically-oriented pages, 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 computations rely on the intuition that the more location entities the anchor text of a link contains, the more likely it is that the given link will guide the crawler to a geographically oriented page. To quantify the probability that a link in a page points to a location-relevant resource, we compute the percentage of the anchor text terms that correspond to toponyms in GeoWordNet. This way, the increased the fraction of toponyms in the anchor text of a link, the greater the probability that this link points to a geographically oriented page. Based on the average values that the combination of the above metrics deliver, our algorithm computes an overall ranking score for every URL in the crawler's seed list and prioritizes URLs in the crawler's queue accordingly (i.e. the URL of the highest rank value appears first in the list). Figure 2, illustrates the steps of our algorithm for ordering geographically-specific URLs in the crawler's frontier.

Based on the above process, the algorithm goes over the data in the crawler's seed list and estimates for every seed URL the probability that it points to a location-relevant page. Then, the crawler starts its web visits from the URL with the highest probability of being geographically-focused.

Moreover, 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 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.

This way, we order URLs in the crawler's queue so as to 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
  /*Compute coordinates of location entities in u*/
  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
  /*Compute avg. spatial correlation of location entities in u*/
  For all c(t) in u do
    Compute paired c(t) distance using Map24
    Compute avg. distance of all c(T) in u
    Use 1-avg.distance as avg. spatial correlation of c (T) in u
    If avg. spatial correlation (u,T) >0.5
        Add u to crawler's frontier
    end
end
/*Compute location focus in the anchor text of links in u*/
Extract anchor text for all links in u L(u)
For every link l in L(u) do
    Compute location focus (l) =
    = |# location entities in anchor (l)| / |# terms in anchor (l)|
    If location focus (l) >0
      Add l to crawler's frontier
    end
end
/* Compute ranking values for URLs u in frontier*/
For each u in frontier do
    Compute Rank (u) =
    = avg. spatial correlation (u) + location focus (u, l) /2
end
Return URLs ordered by Rank (u) values
```

Figure 2: Ordering URLs in the focused crawler's frontier.

3.2 Toward a Geo-Referenced 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 a vector space model (Salton et al., 1975) for representing their contents. Every page is modeled as a vector of terms that constitute the indexing keywords of the page. To build our geo-referenced index, we start with the identification of the page keywords that denote location entities. In this respect, we look the keywords up in GeoWordNet and those found are extracted and further processed in order to resolve their geographic references. Geographic references' resolution practically translates into determining the geographical orientation of the page that contains the identified location entities. To

derive that, we map the location keywords of a page to their corresponding GeoWordNet nodes and we explore their hierarchical relations. Terms that are linked to each other under a common location node are deemed to be geographically relevant, i.e. they refer to the same place.

To verbalize the geographic reference of a page, we use the name of the location node under which the page location entities are organized. Then, we rely on the hierarchical relations among the location entities of common geographic references in order to represent the geographic orientation of the entire page contents. That is, we model the geographic orientation of a page as a small location hierarchy, the root of which denotes the broad geographic area to which the page refers (e.g. country name) and the intermediate and leaf nodes represent specific locations in that area that the page discusses (e.g. city names, landmarks, etc.). Having modelled the geographic orientation of every retrieved page as a structural hierarchy, we label the hierarchies' nodes with their corresponding geographic coordinates that we take from GeoWordNet.

At the end of this process, we end up with a set of location hierarchies, each one representing a different geographical area. Every location hierarchy constitutes a hierarchical index under which we store the URLs of the pages that refer to that place. This way, we end up with a geo-referenced index that groups pages into location hierarchies according to the relatedness of their geographic entities. This index can then be utilized for answering location-sensitive queries.

4. DISCUSSION

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. 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.

To evaluate the performance of our geo-focused crawler, we run two preliminary experiments. In the first experiment, we validated the crawler's accuracy in identifying geographically-relevant data, whereas in the second experiment, we assessed the geographic-coverage of the crawler's visits. 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. 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 geographically specific data sources. Based on those pages, we measured the accuracy of our geo-focused crawler in retrieving geographically-relevant data. To quantify the crawler's accuracy, we estimated the fraction of the pages that the crawler retrieved as geographically relevant from all the visited pages that have some geographic orientation. Formally, we define accuracy as: $\text{accuracy} = |P_{\text{retrieved}}| / |P_{\text{visited}}|$ where $|P_{\text{retieved}}|$ denotes the number of pages that the crawler retrieved as geographically-focused and $|P_{\text{visited}}|$ denotes the number of geographically-oriented pages that the crawler visited. To assess which of the pages (visited and retrieved) do have geographic orientation, we processed their contents and looked for location entities among their elements, using GeoWordNet as our reference source. Pages containing location entities in their contents are deemed as geographically-oriented. Results, reported in Table 1, indicate that our crawler has overall 89.28% accuracy in identifying geographically-relevant data in its web visits.

Table 1: Geo-focused crawling accuracy.

Geographically-oriented visited pages	2.8 million
Geographically-oriented retrieved pages	2.5 million
Geo-focused crawling accuracy	89.28%

As a second evaluation step, we estimated the geographic coverage of the crawler's visits, i.e. the number of different location names that the crawler identifies in the web pages it comes across. To quantify the crawler's geographic coverage, we measured the fraction of distinct geographic entities in the retrieved pages' contents. Formally, we compute the crawler's geographic coverage as:

Coverage = $|E_{distinct}| / |E_{total}|$ where $E_{distinct}$ denotes the number of unique location entities in the page contents and E_{total} denotes the total number of location entities. Results, reported in Table 2, show that our geo-focused crawler can successfully retrieve sources pertaining to distinct geographical areas, ensuring thus complete and qualitative 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%

Finally, we compared our crawler's accuracy in identifying geographically relevant web pages to the accuracy of a classification-aware crawler. For this experiment, we built a Bayesian classifier and we used it to score every URL in the crawler's seed list with respect to their corresponding geographic categories in the Dmoz directory. For scoring geographically-relevant URLs, we relied on the semantic relations between the Dmoz category names and the keywords extracted form the anchor text of the respective URLs, using WordNet. We then used the above set of scored URLs as the classifier's training data. Having trained the classifier we integrated it in a crawling module, which we run against the seed URLs of our previous experiment for one week. At the end of crawling, we computed the accuracy of the classification-aware crawler and we compared it to the accuracy of our geo-focused crawler. In Table 3, we report the comparison results.

Table 3: Comparison results.

Geo-focused crawling accuracy	89.28%
Classification-aware crawling accuracy	79.06%

Results indicate that our geo-focused crawler has improved performance compared to the performance of a classification-aware crawler. 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. Currently, we are running a large-scale focused crawling experiment in order to evaluate the effectiveness of our ranking algorithm in ordering URLs in the crawler's frontier.

ACKNOWLEDGMENTS

Lefteris Kozanidis is funded by the PENED 03ED_413 research project, co-financed 25% from the Greek Ministry of Development-General Secretariat of Research and Technology and 75% from E.U.-European Social Fund.

REFERENCES

Amitay, E., Har'El, N., Silvan, R., Soffer, A. 2004. Web-a-where: geo-tagging web content. In *Proceedings of the 27th Annual Intl. SIGIR Conference.*

Borges, K., Laender, A., Mederios, C., Davis, C. 2007. Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th Intl Workshop on GIR*

Buscaldi, D., Roso, P. 2008. Geo-WordNet: automatic geo-referencing of WordNet. In *Proceedings of the 6th Intl. LREC Conference.*

Chakrabarti, S., van den Berg, M., Dom, B. 2000. Focused crawling: a new approach to topic-specific web resources discovery. *Computer Networks*, 31(11-16): 1623-1640.

Chung, C., Clarke, C.L.A., 2002. Topic-oriented collaborative crawling, In *CIKM Conference*, pp. 34-42.

Ding, J., Gravano, L., Shivakumar, N. 2000. Computing geographical scopes of web resources. In *Proceedings of the VLDB Conference*.

Exposto, J., Macedo, J., Pina, A., Alves, A., Rufino, J. 2005. Geographical partition for distributed web crawling. In *Proceedings of the 2nd GIR Workshop.*

Fellbaum, Ch. (ed.) 1998. *WordNet: An Electronic Lexical Database*, MIT Press.

Fu, G., Jones, C.R., Abdelmoty, A. 2005. Building a geographical ontology for intelligent spatial search on the Web. In *Proceedings of the IASTED Intl. Conference on Databases and Applications*. pp. 167-172.

Gao, W., Lee, H.C., Miao, Y. 2006. Geographically focused collaborative crawling. In *Proceedings of the WWW Conference*.

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

Hill, L. 2000. Core elements of digital gazetteers: placements, categories and footprints. In *Research and Advanced Technology of Digital Libraries*.

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

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

Markowetz, A., Brinkhoff, T., Seeger, B., 2004. Geographic information retrieval. In *Web Dynamics.*

Martins, B., Silva, M.J., Andrade, L. 2005. Indexing and ranking on Geo-IR systems. In *Proceedings of the 2nd Intl. Workshop on GIR*.

Salton, G., Wong, A., Yang, S.C. 1975. A vector space model for automatic indexing. In *Communications of the ACM,* Vol.18, No.11, pp. 631-620.

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.

Smith, D., Mann, G. 2003. Bootstrapping toponyms classifiers. In *Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References*, pp. 45-49.

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 SIGIR Conference*.

Wang, C., Xie, X., Wang, L., Lu, Y.S., Ma, W,Y. 2005b. Detecting geographic locations from web sources. In *Proceedings of the 2nd Intl. Workshop on GIR*.

Welch, M., Cho, J. 2008. Automatically Identifying Localizable Queries. In Proceedings of the SIGIR Conference.

Yu, B., Cai, G. 2007. A query-aware document ranking method for geographic information retrieval. In *Proceedings of the 4th Intl Workshop on GIR.*