Adaptive Geospatially Focused Crawling

Dirk Ahlers
OFFIS – Institute for Information Technology
Oldenburg, Germany
ahlers@offis.de

Susanne Boll
University of Oldenburg
Germany
susanne.boll@uni-oldenburg.de

ABSTRACT

Location information on the Web is a precious asset for a multitude of applications and is becoming an increasingly important dimension in Web search. Even though more and more Web pages carry location information, they form only a small share of all pages and are scattered over the Web. To efficiently find and index location-related Web content, we propose an efficient crawling strategy that retrieves precisely those pages that are geospatially relevant while minimizing the amount of the non-spatially-relevant pages within the crawled pages. We propose to address this challenge by expanding the technique of focused crawling to exploit location references on Web pages to specifically retrieve geospatial topics on the Web. In this paper, we describe the design and development of a focused crawler with an adaptive geospatial focus that efficiently retrieves and identifies location-relevant documents on the Web. Drawing from geospatial features of both Web pages and the link graph, a crawl strategy based on Bayesian classifiers prioritizes promising links and pages, leading to a faster coverage of the desired geospatial topic as a means for fast creation of precise geospatial Web indexes. We present evaluations of the system's performance and share our findings on the geospatial Web graph and the distribution of location references on the Web.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Adaptive Focused Crawling, location-aware Web search, geographic Web information retrieval, resource discovery

CIKM'09, November 2–6, 2009, Hong Kong, China.

1. INTRODUCTION

Location-aware Web search has gained a lot of attention in both research and commercial search engines, corresponding to the high interest of users in location-based information. Several studies [34, 23] conclude that 5 to up to 20% of all user queries express a geographic information need. Matching this obvious need, an estimate of up to 20% of all Web pages contain location references [21, 25].

A geospatial search engine therefore has to identify location-relevant Web pages and extract their location semantics. For a single page, geoparsing techniques can be used to assess its location relevance. However, these pages are scattered among a multitude of Web pages and cannot easily be retrieved for creation of a geospatial Web index.

To build an independent geospatial search engine without having to rely on preprocessed data, several challenges have to be addressed. One major challenge is to design a resource-discovery process that can efficiently discover and retrieve the relevant location-related Web content [9]. The method of choice in the Web context is a crawler that traverses the Web to index its contents. For a specialized location-based search engine, a common broad crawling would be very resource-intensive and of limited use since only a low percentage of retrieved page is relevant to the geospatial domain and the location of interest [2, 4]. Along with other characteristics of geospatial information on the Web, this makes crawling for a geospatial search engine a challenging task.

To address this challenge, we propose to adapt the focused crawling strategies first described in [14] to utilize the Web's inherent link structure and to efficiently retrieve relevant pages. We extend them with the necessary techniques to recognize and exploit the specific geospatial characteristics of the Web to develop a pinpointed focused crawler.

The application scenario for the tailored Web crawler solution is a location-based information system for mobile or pedestrian users. We aim to identify location references at a fine granularity level of individual buildings or addresses that is directly applicable to a mobile user or retrieval and analysis tasks at this geographical granularity [2].

The remainder of this paper is structured as follows. In Section 2 we discuss related work. Section 3 describes the characteristics of geospatial information on the Web and the implications for a crawler. Section 4 covers the design of our adaptive geospatial focused crawler and its heuristics and link prediction methods for adaptation of the crawl strategy. The system is evaluated in Section 5 and following the discussion, Section 6 concludes the paper.

2. RELATED WORK

The search for spatially relevant pages can be understood as a topical fraction of the Web. Contrary to large-scale search engines with the goal to index substantial parts of the Web, topic-centered search engines are only indexing pages relevant to their topic. These aim to maximize the ratio of relevant pages retrieved by exploiting the link structure of the Web combined with content classification.

An analysis of the Web hyperlink graph [11] gives an estimate of the diameter of the Web, with the average link distance between arbitrary Web pages rather small compared to the overall size of the Web. For the design of Web crawlers, this means that due to the small diameter large parts of the Web are accessible in a breadth-first crawl strategy. Conversely, with each additional link depth step, the amount of reachable Web pages increases drastically so that it would soon encompass major parts of the Web. This is a convincing argument for a restriction of a crawl even at small distances from its seeds and the careful selection of links a crawler should follow. Geospatial pages are a common part of this structure.

The seminal work on the topical crawling method are the papers on focused crawling [14, 13] that describe the approach as an effective tool to quickly build up a topical index for specialized search. A later work [12] further explores certain aspects of the approach, concentrating on enhanced topic-based link prediction ranging from link anchor analysis to machine learning approaches.

Using the definitions of [1], focused crawling assumes both linkage locality and sibling locality of the Web's linkage structure. The former means that topical pages are more likely to link to further topical pages than other, the latter means that if some links on a page point to topical pages, the other links are also more likely to link to topical pages. The linkage or topical locality on the Web is confirmed by [16] with several evaluations that show that Web pages on a topic form densely connected clusters. These assumptions lead to an idealized structure of the Web in which topical pages are densely interlinked and have very short link distances between them, as they directly link towards each other. However, this idealized structure is usually not found on the Web. Following the critique of [1] who first doubted the strong linkage locality, we also find that for the geospatial topic, the short distance topic cohesion cannot be assumed [4]. However, a cohesion does exist that can be captured by a broader definition of linkage locality as we will explain in Section 3.

The exclusion of non-relevant pages and the inclusion of relevant pages are main design goals which refine the common breadth-first approach into a best-first link selection strategy. While breadth-first crawling is a preferred strategy for general search engines [29], it is insufficient for the focused crawling approach, which has to employ a focused best-first strategy. The impact of different crawling and link selection strategies is discussed in [15], who argue for the use of relevance metrics in a crawl strategy. The topic of evaluation of topical crawlers is discussed in [26] and [36] and various metrics are presented.

[6] have shown that crawling strategies using historical information yield good results when using the page importance as a metric. We will later use a variation of the described historical-parent strategy where newly discovered pages inherit a weight from their linking page. Furthermore, the work explores a certain geographical aspect of the Web, as it examines the strategies for the geographical concept of a whole country.

Part of focused crawling is the prediction of a link's relevance to the topic. [17] use context graphs as representations of the inlinks of relevant pages. A trained classifier determines whether a page could be the predecessor of relevant pages. The inlinks are generated by querying search engines, the classification of pages uses a naive Bayes technique. [19] discusses the use of machine learning for link prediction in focused crawling and present initial evaluations on a confined testset. [31] present an extensive study and discussion of classification techniques for topical crawling. [27] compare various methods of focused crawling and adaptation strategies and give an overview of the field. To reach indirectly connected pages, [8] use an adaptive cutoff in their crawl strategy. Concerning the applicability of focused crawling, a study on a medical topic [38] has proven the efficiency of focused crawling to build a topic-centered search service. [37] confirms their initial findings, but suggests using a broader scope to retrieve more high-quality, on-topic pages. Since location-relevant pages exhibit a structure different from the thematic topic regarding distribution and link features, these approaches have to be refined to design a geospatial link prediction.

The geospatial information on the Web is mostly hidden in unstructured Web pages. The identification of geographical entities and the use within a search engine is discussed in [24], [25] and [5] for broad regions down to precise mapping [28]. The use of geographic features to guide a Web crawler is discussed in [18]. It describes a load-balancing distributed system with one crawler per region of interest. [35] target the use of neighborhoods instead of precise coordinates in geo-IR and use commercial search engines for initial document retrieval. The derivation of time and place semantics from tagged media is discussed in [32]. The approach argued by our group [2] chooses an address or a precise coordinate as the desired granularity for location-based Web search. We use a geoparser that can reliably retrieve location information at this granularity from unstructured Web pages [3]. With a strong background in pedestrian applications, we are interested in positions that can actually serve as a navigation point rather than coarser allocations to regions or districts. Additional features extracted from the pages allow for multi-dimensional location-based search.

3. GEOSPATIALLY FOCUSED CRAWLING

Focused crawling is a preferred strategy to retrieve thematic topics. It exploits certain assumptions about the Web structure and distribution of relevant pages. Since the distribution and linkage of geospatial pages differs from known topics for focused crawling, we re-evaluate and refine these assumptions towards the geospatial domain.

We illustrate the specific challenges in designing an efficient spatially focused crawler with an exemplary part of the Web graph as depicted in Figure 1. This part of the graph is shown from the perspective of a Web crawler, as a crawl tree. The crawl tree consists of crawled Web pages and of only those links between them that were traversed by the crawler. Relevant pages would be identified by our geoparser and are drawn in a darker shade. Their respective location is referenced and their coordinate shown on a map. The full tree results from a broad crawl and contains a lot of

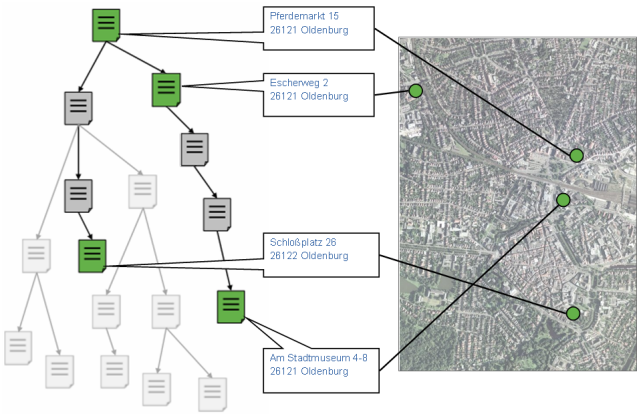

Figure 1: Exemplary selection of a crawl tree with mapped geospatially relevant pages

non-relevant pages. Of these, the lighter shaded pages are on a link path towards relevant pages and are thus necessary to reach the relevant pages. The greyed out pages lie on irrelevant link paths and could be fully removed without any impact on the resulting set of relevant pages.

The goal of the proposed system is two-fold. First, a majority of location-relevant pages should be found on a crawl while second, unpromising and irrelevant paths should be avoided. An optimal crawl would only comprise the solidly drawn pages and would exclude the greyed out ones, which in this example reduces the effort by half. Furthermore, such a system would help in assuring that relevant pages are encountered early on in a crawl [6]. To reach this goal, several issues need to be considered:

Location classification. The geospatial information we are seeking is hidden within the content of hyperlinked Web pages which are scattered across the Web graph. The location-relevant pages can only be reliably identified by examining their actual content. We use our geoparser to identify address-level location information within unstructured Web documents as described in [3] and receive a binary decision on location relevance. If multiple addresses are present in a page, they would map the page to different locations, but it would still register as relevant. A more differentiated treatment is future work.

Linkage locality. Within topic-focused resource discovery, usually a thematic topic is defined and assessed with a classifier. The topical focus depends on the thematic linkage locality hypothesis which states that random pages on the same topic are more likely to link to each other than purely random pages. Thus, the focused crawling approach depends heavily on the cohesion within the thematic topic. The geospatial focus, on the other hand, centers on a feature of the page's content that is not directly thematic but instead is defined by the presence of location information for a region of interest. The cohesion is therefore less distinctive. To our knowledge, no large-scale studies exist that would consider linkage phenomena within the geospatial domain. Several studies [21, 22, 25] state estimates on the number of location-aware Web pages at about 10–20%, but have not assessed link structure distribution.

Loose cohesion. To gain a better understanding, we examined the Web link structure in our initial studies [4] and have shown that a strict cohesion of the geospatial topic is only rarely present. In the example in Figure 1, only one link on the right branch directly links relevant pages, while other pages are less directly linked. This is in part due to the sparse population of the graph with spatially relevant pages, making a broad crawl unfeasible. However, the spatially relevant pages are connected by intermediate, non-relevant pages. These so-called bridge pages maintain the cohesion of the geospatial topic. Many geospatial pages tend to be linked with only a small number of bridge pages between them. Therefore, a distance heuristic can be defined around relevant pages that can include the following relevant pages. One disadvantage, as can be seen in the example, is that it is not easy to distinguish a bridge page from a non-relevant page that does not lead to any more relevant pages, a so-called tail. Therefore, such an approach would still include most of the pages in the example. To reach the goal, we have to identify which of the non-relevant pages are actually bridge pages and will later on lead to more spatially relevant pages. We therefore aim at a probability classification of links leading to relevant pages as a way for a broader definition of linkage locality. The crawler has to balance the immediate reward of quickly harvested relevant pages to the delayed reward that comes from following bridge pages [7, 33]. This delayed harvest depends on the classification of bridges and tails to efficiently guide the crawler towards the location-relevant pages sparsely distributed in the Web graph.

Link prediction. For a location-relevance prediction, we need to design a classifier to separate bridges from tails. This is equivalent to a link prediction that assigns priorities to outgoing links according to their judged relevance to the geospatial topic. The link prediction needs to take the presence of bridge pages into account and should use a lookahead over several links. This means that the relevance of a link needs to be assessed in several successive link distances to properly predict the link's relevance. For this, we examine relevant features from Web pages and their outgoing links and use machine learning techniques to uncover similarities and relations between those features that hint at geospatial relevance.

4. APPROACH

Based on the requirements discussed before, we develop several components for the adaptive geospatial focused crawler. An adaptation of a breadth-first crawling strategy includes bridge pages by defining link distances and a propagation of relevance. This is input to a link prioritization according to a relevance score. Finally, a link prediction assesses the probability that a link leads to relevant pages, both for the directly linked page and, since we have to allow for bridge pages, for arbitrary link depths.

4.1 Location Relevance Propagation

We use information from the currently crawled pages as feedback to dynamically adapt the crawl strategy. The most obvious information gained from a crawl is the location relevance of the downloaded pages. We exploit the weak geospatial cohesion and assume the heuristics that the location assessment of a page determines to some extent the assessment prediction of its outgoing links.

We define a radius of bridge pages around each relevant page to dynamically define the extension of further crawl-

ing. Thereby the geospatial relevance of a page is defined as extendable to its outlinks with a falloff function. This models a relevance propagation tied to the bridge page distance. Figure 2 shows a diagram of bridge pages on a crawl branch. The upper part shows a crawl branch with two bridge pages between relevant pages, the lower depicts the definition of a distance around a relevant page. Note that this metric is different from using seed distance. Using the distance of a page to its seed, the crawl tree would grow steadily from its root. On the other hand, by using information gained during the crawl, the growth of the crawl tree can dynamically be adapted to grow towards more promising pages and reduce the download of non-relevant crawl branches.

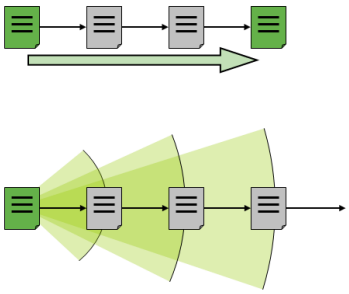

Figure 2: Bridge pages and distances

4.2 Prioritization

The propagated relevance score is the basis for an adaptation of the breadth-first crawling strategy which we turn into a best-first approach. The crawling strategy is basically a selection strategy. Each URL is given an evaluation value and based on the order of the values, the next URL to be crawled is selected.

In our crawler, pages are given an exponentially decaying score value that is derived from the bridge page distance. They are then processed in score value order.

The prioritization of pages leads to a crawler that performs well for both short and long-term operation. In the early stages of a crawl where many relevant pages are close to the seeds, these would be preferred. For longer crawl durations, it would maintain the distance-based heuristic over arbitrary link distances, giving precedence to those pages that have a higher probability according to the topical locality assumption.

This allows the crawl to run over arbitrary link distances without having to make assumptions about bridge page distances beforehand. The crawl will continue to find new relevant pages, even if it will reach a state of lower marginal utility. Thus the crawl runs in a breadth-first mode at the beginning as all links are rather similar in priority. After a while, when the crawler turns to less-prioritized pages, newly discovered relevant pages trigger a restricted in-depth traversal from that page on, leading to a continuous best-first crawling. Hence the crawl steadily advances through all low-priority pages but can concentrate efforts and performance on newly discovered relevant pages to quickly explore their link neighborhood.

We design the relevant functions for the crawl strategy as follows. We distinguish between the actual evaluated relevance *eval* of a page as determined by a classifier and the derived predicted relevance of a page designated $score_{page}$.

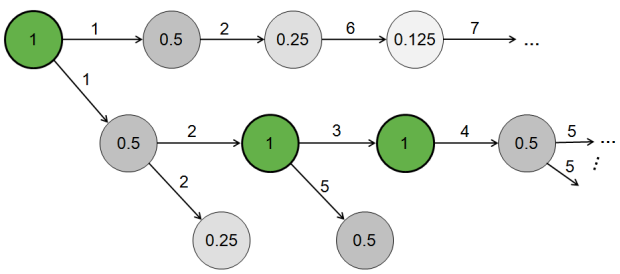

Figure 3: Prioritized crawl tree with crawl order

Each seed is assigned an initial $score_{seed} = initscore = 1$. We denote $eval_{page}$ as an indicator function based on our geoparser *geo* to determine whether a *page* contains relevant location information at a high granularity and is part of our given region. We currently use a binary classification of geo-relevance.

$$eval_{\text{page}} = geo_{\text{page}} = \begin{cases} 1 & \text{if } page \text{ is location-relevant,} \\ 0 & \text{if } page \text{ is not location-relevant.} \end{cases}$$

The relevance score of a location-relevant page is propagated along links with an exponential decay. We denote $page.dist \geq 1$ as the link distance of a *page* to the last relevant page P, $P \xrightarrow{dist} page$. We choose a dampening factor to model the diminishing relevance of more distanced pages by a factor of *decay*, similar to the PageRank algorithm [10]. For pages $u \rightarrow v$ the score would be calculated as $score_v = decay \times score_u$, leading to a general term $score_{page} = decay^{page.dist} \cdot initscore$. Since we keep *initscore* at 1, we can remove the factor. With these definitions, we can fully define the relevance score of a page as follows:

$$score_{page} = \begin{cases} initscore & \text{if } page \in seeds, \\ initscore & \text{if } eval_{page} = 1, \\ decay^{page.dist} & \text{if } P \xrightarrow{dist} page \wedge eval_P = 1 \\ & \quad \wedge eval_{page} = 0. \end{cases}$$

To ensure a prioritized crawl, the selection strategy for pages is to select the next page *next* to be crawled from all currently assessed candidate page URLs p_i in the queue *queue*.

$$next = p \in queue | score_p = \max_{p_i \in queue} (score_{p_i})$$

When multiple pages fulfil the selection criteria, pages are chosen randomly from the pool, while of course observing politeness rules for Web crawlers.

Figure 3 visualizes the algorithm, with prioritization employed with scores according to the exponential dampening. The value for decay is set at $decay = \frac{1}{2}$. The starting score of $initscore = 1$ is multiplied by *decay* for each link encountered. For this example, links are annotated with the order in which they will be crawled. The first page dampens its outlink by the *decay*, thus their priority is computed as 0.5 and they are similarly likely to be crawled. In the next step, the three outgoing links are again dampened to 0.25 and crawled accordingly. Then, the page on the lower branch

gets evaluated as location-relevant, which means that its score is replenished to 1 and subsequently, its children receive a higher priority score of 0.5. These are therefore processed first. Only when the lower branch runs out of pages or its remaining links have a low priority, can the upper branch continue to be crawled.

Note that for more complex examples, the exact order in which pages are crawled also depends on the processing speed of the crawler for individual pages and might therefore be subject to small changes in the crawl order. Still, the prioritization allows for fine-grained control over the order in which pages are crawled, providing an important step towards best-first crawling.

4.3 Predictive Geospatial Focus

The location relevance propagation already constitutes a rather simple best-first crawling approach. However, the definition of the 'best' link to follow relies solely on the link distance measure. Furthermore, the amount of non-relevant pages is rather high since all outlinks from a given page are treated equally.

We refine the notion of what links should be considered more relevant for prioritization to also include the actual features of individual hyperlinks. From these features, we compute a relevance prediction for each link.

The common rationale is that many links contain a description of the content of the linked page. However, they do not necessarily do so with keywords detectable by a geoparser because, as mentioned before, the link description might rather focus on the thematic aspects of the resource. Still, links pointing to geospatially related Web content are considered to contain detectable patterns. As an example, the crawl tree in Figure 4 shows how a prediction on the outgoing links from the initial pages can raise the priority of a successful branch so that the relevant pages would be found faster, provided that the link actually can be classified as belonging to a successful crawl branch.

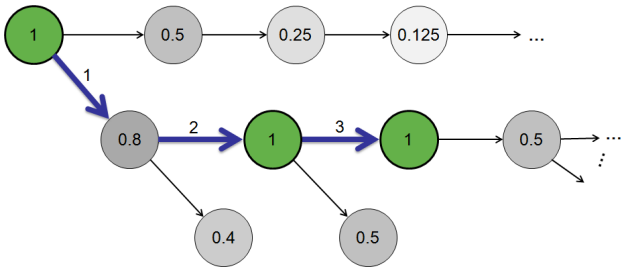

Figure 4: Crawl tree with adaptive link prediction

Note that the approach has to be aware of bridge pages which could span multiple non-relevant pages. We extend previous approaches [17, 19, 27] at link prediction to not only arrive at a prediction on the page immediately linked, but to also give an estimate on the location relevance of pages at a larger distance. This allows a lookahead of multiple pages on a single link and further improves the focused crawling strategy. In the following, we describe how the prediction lookahead is designed based on the used features and the prediction on individual links.

4.4 Adaptive Link Prediction

The link prediction makes assumptions about pages still to be crawled utilizing only information from the linking page and already established knowledge. The context information on a link is vital to the reliable prediction of the target page's content. Reviewing the relevant literature, e.g., [7, 20, 26, 27], we select several features of a hyperlink within an HTML page to be considered for the link prediction:

– *URL host* is used to take clues from the hostname of a URL. This in some cases already contains a location reference if the main topic of the pages deals with a geospatial topic or with an entity related to a location but can also be used to discover the main topic of discourse.

– *URL path* is used separately since use of indicator terms is different on hostname and path. This can be used to indicate individual pages dealing with a location or common indicator of contact information.

– *Anchor text* and *Link title* is used to describe a target page on the linking page and therefore can bear hints towards the target's content. This can be an indicator of the information on the linked page, but needs not necessarily use the same keywords present on that page.

– *Surrounding text* around a link gives information about the context in which the link was used. The context can carry more information than only the anchor itself [30].

– *Domain switch* is used to distinguish between links within the same domain and those pointing outwards.

The features of each link are evaluated to derive the probability that the link will point towards geospatially relevant pages. Also based on the literature [7, 12, 27], we utilize a naive Bayes classifier on the link's textual features. We leave a comparison of various classification techniques as was done in [31] to future work. We use a training dataset for which all relevant pages and the links towards them are known. This data is generated by the crawler during a training phase. Since we can assess the relevance of a Web page, we can implement the supervised learning without human intervention. The teacher component is our geoparser which decides whether pages are relevant to the geospatial topic. Therefore, it is known which links actually linked to geospatially relevant information and we can feed back this information accordingly.

As per Bayes' theorem, we denote the probability of a link L with features $\vec{f_L}$ leading to geospatially relevant pages:

$$p(geoRel_1|\vec{f_L}) = \frac{p(geoRel_1) \cdot p(\vec{f_L}|geoRel_1)}{p(\vec{f_L})}$$

where $p(geoRel_1)$ is the prior probability that a linked page would contain geographic references. $p(\vec{f_L}|geoRel_1)$ is the conditional probability that hyperlinks with the features $\vec{f_L}$ would lead to pages containing geographic references, and finally $p(\vec{f_L})$ is the prior probability that the given link features are present in a hyperlink. We currently assume conditional independence of the features. Given these, a new link's features can be evaluated to arrive at its probability to point to a relevant page.

To also support online learning where we might not have the full information available, we use a maximum likelihood estimate which derives an estimate about the prior and conditional probabilities from the training data. Furthermore, as the full feature vectors $\vec{f_L}$ are unlikely to repeat them-

selves between different links, we break up the feature vector into its constituent terms t_i and derive the probability for the individual terms instead, which would actually count from the probability tables. $p(\vec{f_L}|geoRel_1)$ can then be given as the estimate $\widehat{p}(\vec{f_L}|geoRel_1)$, We denote T as the list of terms t_i in $\vec{f_L}$ with the number of terms $|T| = n$ and compute the estimate as

$$\widehat{p}(\vec{f_L}|geoRel_1) = \prod_{i=1}^{n} \widehat{p}(t_i|geoRel_1)$$

We further estimate the probability of the term t_i being present in geospatially relevant links L_{georel_1}. We use Laplace smoothing to cover cases where terms are not present in the training data, since the whole estimate would otherwise drop to zero.

$$\widehat{p}(t_i|geoRel_1) = \frac{|t_i \in L_{georel_1}| + 1}{\sum_{t_k \in T} |t_k \in L_{georel_1}| + 1}$$

This defines the probability of a link leading to relevant pages based on individual terms of the examined features of a link with respect to the training data.

4.5 Bridge Page-Aware Link Prediction

The described classifier only maintains a lookahead of one for the link prediction. Since bridge pages occur frequently within the geospatial topic, they have to be considered. We therefore design the prediction to decide whether a link leads to unsuccessful crawl branches or to bridge pages which demands a multi-hop prediction.

The example in Figure 5 shows the prediction graph of one Web page with one of its outlinks currently being evaluated for the prediction. The prediction considers the pages reachable within several depth levels from the current page. The challenge for the multi-depth prediction is of course that the amount of reachable pages increases with any further step. The trivial prediction depth of one only evaluates one page. For a prediction depth of two, this is the number of all pages linked to by the first page etc. Note that the prediction only applies to the first single outlink. It incorporates a prediction about further links, but these are not accessed at prediction time. The prediction about successive pages can be re-evaluated in successive steps when pages are actually downloaded and more information becomes available during a crawl.

Generalizing the previous equation for the geospatial relevance probability of a link, we define the probability over arbitrary distances. The depth prediction score depends on the examined link and the prediction depth.

$$\begin{aligned} depthPrediction_L(depth) &= p(geoRel_{depth}|\vec{f_L}) \\ &= \frac{p(geoRel_{depth}) \cdot p(\vec{f_L}|geoRel_{depth})}{p(\vec{f_L})} \end{aligned}$$

We combine the prediction scores of successive depths into one measure. Since the reliability of the prediction decreases with increasing prediction depth, we adjust by adding a dampening factor to predictions of larger depth as $d(depth) = decay^{depth}, decay \in [0, 1]$. We normalize this against the sum of all dampening and arrive at the overall link prediction score:

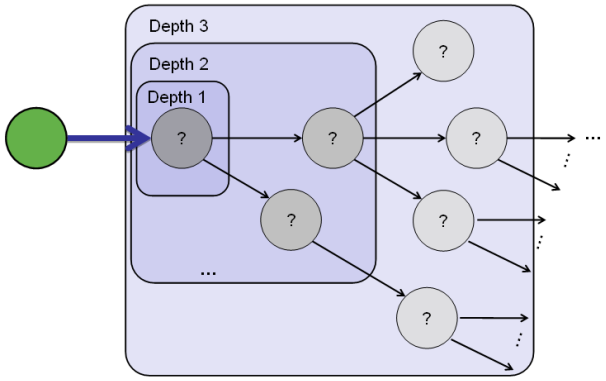

Figure 5: Prediction graph of a Web page

$$prediction(L) = \frac{\sum_{depth=1}^{n} d(depth) \cdot depthPrediction_L(depth)}{\sum_{depth=1}^{n} d(depth)}$$

The geospatial relevance can only roughly be predicted since the lack of corresponding features within the link prediction does not imply a lack of geospatial information in the linked pages. We therefore use the prediction score only as a measure to boost pages to a higher priority, but not to reduce it below the link-distance decay discussed in Section 4.1. This delivers satisfactory results.

The overall priority $priority_L$ of a link is defined by the distance-based decay score $score_L$ which models an exponential decay, refined by the link prediction $prediction_L$ which can increase the priority of a link. The priority is normalized to the interval $[0, 1]$.

The prediction is normalized to fill the interval between the link's decay score value and the maximum priority. This models the intuitive assumption that a link predicted for certain to lead to a relevant page should be most highly prioritized.

$$priority(L) = score(L) + prediction(L) \cdot (1 - score(L))$$

Using the prioritization selection function as defined in Section 4.2, the crawler implements a best-first strategy based on the fusion of distance-based and prediction-based link evaluation. It is then able to efficiently retrieve even distanced relevant pages by targeting promising bridge pages.

4.6 Architecture

The integration of the described components into the architecture of the adaptive geospatially focused crawler are laid out in Figure 6. Other aspects of a geospatial search engine such as indexing, storage, and query processing are left out for clarity.

A frontier component is employed for data management. It manages all link queues and handles link assignment to multiple work threads, where URLs are processed and downloaded. All discovered links enter the frontier through the link evaluation. The frontier selects the next link with the highest priority to crawl from the prioritized link queue and

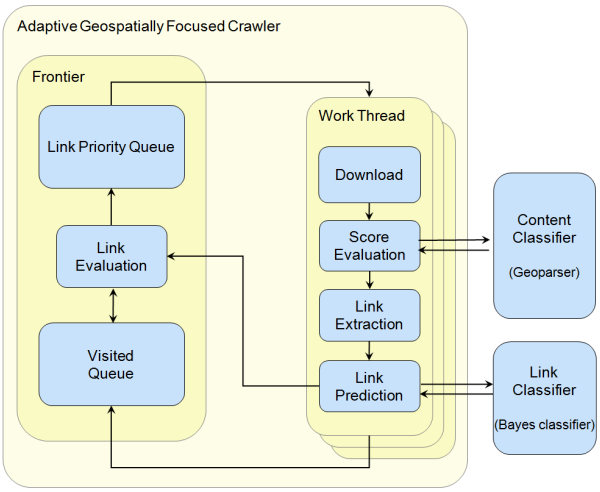

Figure 6: Architecture of the crawler

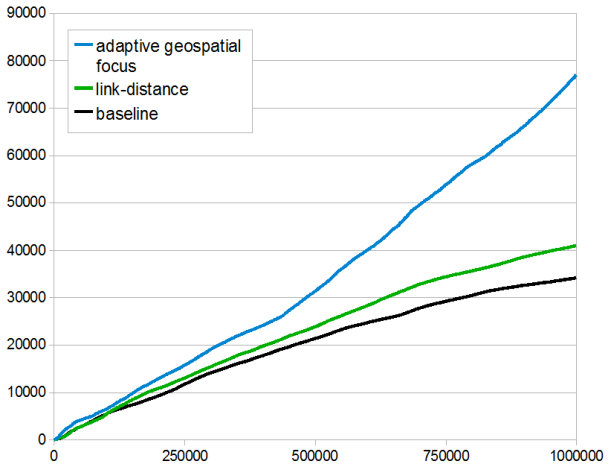

Figure 7: Harvestrate of retrieved relevant pages against all retrieved pages

hands it to a work thread. This downloads the URL and establishes the relevance score of the page using the geoparser component. Following, the links of the page are extracted. The link extraction also handles the propagation of score values to child pages. Using the link context on a page, a prediction is computed for each link by using the trained classifier. The original link is added to the visited pages queue, where all already downloaded links are kept. The extracted links are then processed by the link evaluation. Each extracted link is compared against the visited pages queue so no page can be downloaded twice. Finally, the link evaluation combines the score values of the link and computes the priority. The link is then inserted into the prioritized queues according to its priority.

5. CRAWL STRATEGY EVALUATION

The goal of the evaluation is two-fold. First, we want to measure the performance of the developed components and the focused crawler, and second, we want to arrive at a better understanding of the Web structure relating to our task.

5.1 Methodology

To evaluate the geospatially focused crawler we evaluate it against two other crawl strategies. The first is a common broad crawl that traverses the Web in an unbounded breadth-first manner. A second comparison for the crawler is the link distance heuristic that assigns priority values based on distance from the last seen relevant page on a crawl branch. In a previous paper [4], we already established that a strict geospatial focus with a fixed cut-off distance will exhaust the crawl rather quickly. The results were similar on the present testset, we therefore leave them out for clarity.

There are two main reasons for resources to be included in a crawl for geospatial information. One is the identification and extraction of geospatial information itself, the other is the extraction of links leading to further resources. So only resources that either carry location references or hyperlinks need to be considered in the crawl, all others can be excluded without losing any relevant information. To reduce the amount of downloaded resources, we set up a set of fixed filters that exclude resources based on content type, URL patterns or content language. The address data we are seeking is only present in textual Web documents and further hyperlinks could only be extracted from HTML files. We therefore remove all binary content from the crawl and of the textual content, only keep HTML documents. Further filters remove country code top level domains (TLD) that cannot carry the German location references we are seeking. An additional set of filters tries to identify and dismiss crawler traps.

We created identical conditions for all crawls. The system was initialized with a region of interest, the city of Oldenburg in Germany and the geoparser was trained to this region. We selected a seed set from the DMOZ Open Directory, where we chose those pages under the geographical hierarchy that deal with the city of Oldenburg. The crawler was started with this seed set, the filters as described above were in place.

The crawls each ran for a duration of about 24 hours. The baseline crawl received no further changes. The link-distance prioritization was parameterized as described in Section 4.2. Finally, the adaptive geospatially focused crawler was parameterized with values for $decay = \frac{1}{2}$; the prediction depth was parameterized from 1 to 3. For the depth of 3, we also varied the size of the set of training documents.

5.2 Results

The results of the evaluation clearly indicate that there exists a topical cohesion of location-related information on the Web and that it is sufficient to support the focused crawling strategy discussed in this paper. We further can show that the described approach outperforms other crawl strategies.

5.2.1 *Crawler performance*

Focused crawling is a resource-constrained trade-off approach aimed at retrieving relevant pages much faster than a common crawler, yet it might miss some relevant pages. Considering the size of the current Web, it is not possible to retrieve all pages. Since we do a crawl on the live Web, we cannot give reliable values on recall as it remains unclear what amount of potentially relevant pages are missed. However, as an estimate on precision and recall [36], we evaluate

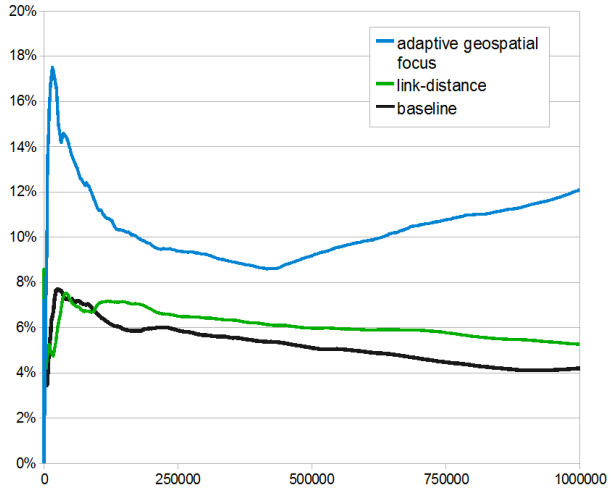

Figure 8: Relative harvestrate of retrieved relevant pages against all retrieved pages

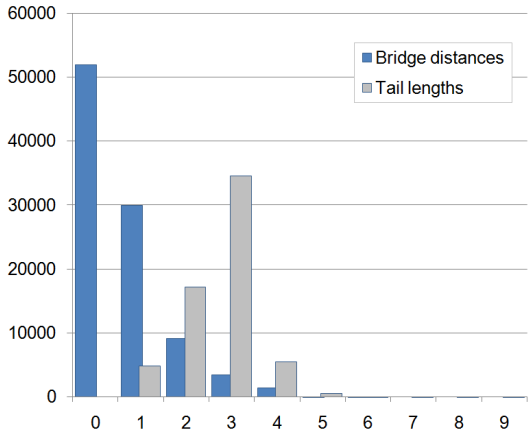

Figure 9: Bridge page distances and tail lengths

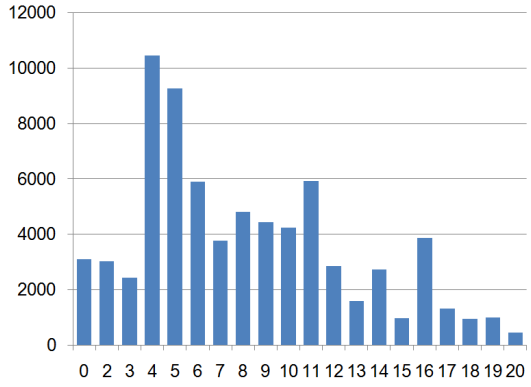

Figure 10: Seed distances of relevant pages

the crawler by the amount of relevant pages it retrieves over the time of a crawl and its harvest rate.

Figure 7 shows the harvest rate defined as the amount of retrieved relevant pages drawn against the number of overall downloaded resources for the three crawl strategies. The adaptive geospatially focused crawling outperforms the broad crawl and the link-distance strategy and displays a steady growth. While at the beginning all crawls manage to retrieve relevant pages near the seeds, the other crawls' harvestrate begins to flatten after a while, while the adaptive system manages to uphold and even increase its performance. This is clearly shown in Figure 8, which plots the accumulated relative harvest rate against the overall downloaded resources. It demonstrates the good initial performance of the system, which outperforms the other crawls in the beginning by rapidly targeting promising pages. However, the initial peak is not sustainable and it drops as the pages in the immediate vicinity of the seeds have been processed. However, the crawler then again manages to efficiently target further relevant pages during the remainder of the crawl. At the end of the 24-hour crawl, which we have plotted over the roughly 1.000.000 documents downloaded, the adaptive approach reached a harvest rate of about 12% while the baseline retrieved 4% and the link-distance-heuristic managed to retrieve 5.5% of relevant documents.

We have further evaluated the adaptive geospatially focused crawling with various parameters for the size of the training set. From 30.000 up to 100.000 documents there was no significant change. Regarding the prediction depth, we found that a value of 3 performs significantly better than a value of 2 and 1. Higher prediction depths had no more significant influence on the results.

5.2.2 Location Distribution on the Web

While the results suggest the presence of linkage locality, its characteristics need to be better examined. The graph in Figure 9 shows the link distances as the number of bridge pages between location-relevant pages. A value of 0 means directly successive relevant pages. Additionally, unsuccessful crawl branches are denoted as tails. These branches started from a relevant page but then failed to retrieve further relevant pages. The graph shows only few longer tails. Combined with the analysis of the bridge pages, this shows that location-relevant pages exhibit a loose topical cohesion. Namely, there are pages on the crawl which were only reached through long bridge page streaks. In more than half the cases, the cohesion is only achieved by intermediate bridge pages that link the relevant pages.

To complement the bridge page analysis, Figure 10 shows the distance of relevant pages from the seeds. While relevant pages occur near each other, as evidenced in Figure 9, their distance from the seed reaches rather high values and furthermore shows the necessity of the prediction and prioritization in the adaptive crawler to reach these pages with a focused crawl.

5.3 Discussion

Geospatially focused crawling can benefit from the detectable and exploitable cohesion of its topic, although the cohesion is often weak. Through the inclusion of bridge pages and the assumption of the loose cohesion, geospatially focused crawling becomes feasible. While the inclusion of

more pages and therefore a larger distance from the seeds produces an exponentially growing queue, the techniques described in this paper can help to keep a crawl focused for a long duration. Naturally for a crawl with these characteristics, the inherent uncertainty about uncrawled pages leads to a high ratio of non-relevant pages for a longer crawl. This cannot be completely avoided, but is kept within manageable bounds. Finally, the distribution of location-bearing Web pages – similar to crawling other topics – cannot be fully predicted. There remain some pages whose content cannot be properly predicted that would only be discovered by an exhaustive crawl. Some of the relevant information might only be found in rather obscure parts of the Web graph, a long distance from any other relevant page. Thus, the aim cannot be completeness for the topic but rather the retrieval of substantial amounts of geospatial information. Therefore, when a substantial amount of reachable relevant pages have been found, the overall performance will decrease and approach that of the broad crawl. However, durations as described here are still far from the level of marginal utility. Regarding long-term performance, our crawler manages to keep up the advance on the baseline crawls even on longer crawls. As the evaluation shows, the crawler manages to uphold the geospatial focus even for larger crawl durations within the demanding geospatial domain and can be used to quickly generate a geospatial index.

6. CONCLUSION

In this paper, we proposed the adaptation of focused crawling for geospatial resource discovery and presented an efficient design for a geospatial crawler. Utilizing link cohesion between relevant pages, the crawler adapts itself to the content and link features encountered during the crawl. A prediction of location-relevance for uncrawled pages allows a precise and fast crawl with limited resources. The bridge-page-aware link prediction allows for a larger lookahead and is suitable for the loose cohesion of the geospatial topic. The designed adaptive geospatially focused crawler is a reliable technique to retrieve a majority of location-relevant pages much faster than ordinary crawlers. The analysis of the retrieved data fosters understanding of the distribution of location-related content on the Web. Building upon these promising results, further steps will include an improvement of the link prediction, the further analysis of dependency on the seed sets and a more thorough analysis of the geospatial Web graph to advance our understanding and use it to more efficiently retrieve and index geospatial Web pages.

Acknowledgements

We thank our student C. Krumm for valuable contributions to the system described in this paper. Further thanks go to our students A. Waldenburger and T. Scheffler for help with some of the analyses. Part of this work has been supported by the state of Lower Saxony, Germany as a subproject of the Niccimon and C3World projects.

7. REFERENCES

[1] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu. Intelligent crawling on the World Wide Web with arbitrary predicates. In *WWW '01*. ACM, 2001.

[2] D. Ahlers and S. Boll. Location-based Web search. In A. Scharl and K. Tochterman, editors, *The Geospatial Web*. Springer, London, 2007.

[3] D. Ahlers and S. Boll. Retrieving Address-based Locations from the Web. In *GIR '08: 5th Workshop on Geographic Information Retrieval*, 2008.

[4] D. Ahlers and S. Boll. Urban Web Crawling. In S. Boll and E. Wilde, editors, *First International Workshop on Location and the Web (LocWeb2008)*. ACM, 2008.

[5] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web Content. In *SIGIR '04*, pages 273–280, New York, NY, USA, 2004. ACM.

[6] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Crawling a Country: Better Strategies than Breadth-First for Web Page Ordering. In *WWW '05*, pages 864–872. ACM Press, 2005.

[7] L. Barbosa and J. Freire. An Adaptive Crawler for Locating Hidden-Web Entry Points. In *WWW '07*. ACM, 2007.

[8] D. Bergmark, C. Lagoze, and A. Sbityakov. Focused Crawls, Tunneling, and Digital Libraries. In *ECDL '02*, pages 91–106, London, UK, 2002. Springer.

[9] S. Boll and D. Ahlers. A Web more Geospatial: Insights into the Location Inside. In D. De Roure and W. Hall, editors, *Workshop on Understanding Web Evolution: A Prerequisite for Web Science (WebEvolve2008)*, 2008.

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, 33(1):309–320, June 2000.

[12] S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated Focused Crawling through Online Relevance Feedback. In *WWW '02*. ACM, 2002.

[13] S. Chakrabarti, M. van den Berg, and B. Dom. Distributed Hypertext Resource Discovery Through Examples. In *VLDB Journal*, pages 375–386, 1999.

[14] S. Chakrabarti, M. van den Berg, and B. Dom. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. *Computer Networks*, 31(11-16):1623–1640, 1999.

[15] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. In *WWW7*, 1998.

[16] B. D. Davison. Topical locality in the Web. In *SIGIR 2000*, pages 272–279. ACM Press, 2000.

[17] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori. Focused Crawling using Context Graphs. In *26th Intl. Conf. on Very Large Databases (VLDB 2000)*, pages 527–534, Cairo, Egypt, 2000.

[18] W. Gao, H. C. Lee, and Y. Miao. Geographically focused collaborative crawling. In *WWW '06*, 2006.

[19] A. M. Grigoriadis and G. Paliouras. Focused crawling using temporal difference-learning. In *SETN*, pages 142–153. Springer, 2004.

[20] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur. The shark-search algorithm. An application: tailored Web site mapping. In *WWW7*, 1998.

[21] M. Himmelstein. Local Search: The Internet Is the Yellow Pages. *IEEE Computer*, 38(2):26–34, 2005.

[22] M. Jakob, M. Großmann, D. Nicklas, and B. Mitschang. DCbot: Finding Spatial Information on the Web. In L. Zhou, B. C. Ooi, and X. Meng, editors, *DASFAA 2005*, volume 3453 of *LNCS*. Springer, 2005.

[23] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *CHI '06*, pages 701–709. ACM, 2006.

[24] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger. Design and Implementation of a Geographic Search Engine. In *WebDB 2005*, 2005.

[25] K. S. McCurley. Geospatial Mapping and Navigation of the Web. In *WWW '01*. ACM, 2001.

[26] F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Inter. Tech.*, 4(4):378–419, 2004.

[27] A. Micarelli and F. Gasparetti. Adaptive Focused Crawling. In *The Adaptive Web*, volume 4321 of *LNCS*, pages 231–262. Springer, 2007.

[28] Y. Morimoto, M. Aono, M. E. Houle, and K. S. McCurley. Extracting Spatial Knowledge from the Web. In *SAINT '03*. IEEE, 2003.

[29] M. Najork and J. L. Wiener. Breadth-First Crawling Yields High-Quality Pages. In *WWW10*, pages 114–118, Hong Kong, May 2001. Elsevier Science.

[30] G. Pant. Deriving link-context from HTML tag tree. In *DMKD '03*. ACM, 2003.

[31] G. Pant and P. Srinivasan. Learning to Crawl: Comparing Classification Schemes. *ACM Trans. Inf. Syst.*, 23(4):430–462, 2005.

[32] T. Rattenbury, N. Good, and M. Naaman. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. In *SIGIR '07*. ACM, 2007.

[33] J. Rennie and A. McCallum. Using Reinforcement Learning to Spider the Web Efficiently. In *ICML '99*, 1999.

[34] M. Sanderson and J. Kohler. Analyzing geographic queries. In *ACM SIGIR Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004.

[35] S. Schockaert and M. D. Cock. Neighborhood Restrictions in Geographic IR. In *SIGIR '07*, 2007.

[36] P. Srinivasan, F. Menczer, and G. Pant. A General Evaluation Framework for Topical Crawlers. *Information Retrieval*, 8:417–447, 2004.

[37] T. T. Tang, D. Hawking, N. Craswell, and K. Griffiths. Focused Crawling for both Topical Relevance and Wuality of Medical Information. In *CIKM '05*, pages 147–154, New York, NY, USA, 2005. ACM.

[38] T. T. Tang, D. Hawking, N. Craswell, and R. S. Sankaranarayana. Focused Crawling in Depression Portal Search: A Feasibility Study. In *ADCS 2004*, pages 2–9, Melbourne, Australia, 2004.