Using the Geographic Scopes of Web Documents for Contextual Advertising

Ivo Anastácio
ivo.anastacio@ist.utl.pt

ABSTRACT

Contextual advertisement is normally seen as an Information Retrieval problem, where the objective is to retrieve the most relevant ads given a target page. Geotargeting is a particularly interesting specialization of contextual advertising, where the objective is to target ads to audiences concentrated in well-defined areas. Currently, the most common approach involves targeting ads based on the physical location of the visitors, estimated through their IP address. However, there are many situations where it would be more interesting to target ads based on the geographic scope of the target pages, i.e., on the locations referred to in the textual contents of the pages. In this paper, we propose to apply techniques from the area of geographic information retrieval to the problem of geotargeting. We address the task through a pipeline of processing stages, which involves (i) determining the geographic scopes of target pages, (ii) classifying target pages according to locational relevance, and (iii) retrieving relevant ads to the target page, using both its textual content and its geographic scope. Experimental results attest for the adequacy of the proposed methods in each of the individual processing stages.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Abstracting methods; H.4.m [**Information Systems**]: [Miscellaneous]

General Terms

Algorithms, experimentation

Keywords

Contextual Advertisement, Geotargeting, Geographic Text Mining, Geographic Information Retrieval

1. INTRODUCTION

Online advertising platforms such as Google AdSense[1] or Yahoo! Content Match[2] are nowadays the financial backbone of the Web. The primary business model behind most non-transactional Web sites is currently based on contextual advertisement, where contextually relevant textual ads are displayed alongside the regular content of Web pages.

From a research standpoint, contextual advertising is normally seen as an Information Retrieval (IR) problem, where the objective is to retrieve the most relevant ads given a target Web page. Previous studies have shown that, in the contextual advertisement domain, relevance increases the probability of reaction and is therefore strongly associated with profitability (i.e., more relevant ads lead to improved user satisfaction and higher response rates) [25]. One problem that has been getting increasing attention is therefore the design of IR ranking functions to select advertisements that are highly relevant.

A particularly interesting specialization of contextual advertising is localized advertisement, also known as geotargeting, where the objective is to target ads to audiences concentrated in well-defined areas. This is particularly interesting to advertisers that have local businesses and are looking to generate shop traffic or calls for professional services. For example, a takeaway restaurant serving a particular region would like to target its advertisements to that region.

Nowadays, the most common geotargeting approach involves targeting ads based on the physical location of the Web site visitors, estimated using their IP addresses [26]. While this would work on the example of takeaway restaurants (i.e., potential clients are often interested in knowing what is near to their current location), the IP-targeting approach has several limitations. Besides the inacuracies involved in IP geocoding, there are many other situations where it would be more interesting to target ads based on the geographic scope described in the content of the target pages.

Consider the example of a user who is traveling to Lisbon and is browsing Web pages describing tourist attractions and events in the city. In these pages, it would be more interesting to place ads that are relevant to the geographic location that is described in the content of the pages. To handle these cases, advertisers often include region-specific keywords on the textual description of the ads, hoping that they match the placenames mentioned in the Web pages. However, this is by no means an optimal solution since it cannot account for other geographical factors, such as proximity or containment (e.g., the name Lisbon would not match with placenames that correspond to sub-regions, like Chiado).

[1] `http://www.google.com/adsense`
[2] `http://publisher.yahoo.com/sell/ContentMatch.php`

A recent trend in IR applications relates to extracting geographic context information from textual documents, in order to explore it for purposes of document retrieval [9, 11, 23]. This is usually referred to as Geographic Information Retrieval (GIR). In this paper, we explore the usage of GIR techniques for geotargeting advertisements. Similarly to traditional contextual advertisement, we model the problem as a task of retrieving the most *locationally relevant* ads, given a target Web page. By locationally relevant we mean advertisements whose target population matches the geographic scope of the target Web page. While techniques for geographic IR have been getting increased attention, their application to the contextual advertisement domain is, to the best of our knowledge, a novel contribution of this paper.

We propose to address the task through a pipeline of operations, in which we (i) extract place references from the target Web pages and assign them to geographic scopes, (ii) classify the target pages as either local or global, using features such as the text or the extracted place references, and (iii) find relevant ads to the target Web pages by using techniques from the area of geographic information retrieval, combining thematic and geographic similarities.

The main contributions of this paper are as follows:

1. We compare several strategies for assigning geographic scopes to target Web pages, including both well-known algorithms and baseline methods.
2. We propose and evaluate a supervised machine-learning approach for the task of classifying the target Web pages according to their implicit locational relevance, i.e., classifying them as either local or global.
3. We propose and evaluate different retrieval strategies for the task of displaying relevant advertisements in target Web pages, which leverage on the results obtained from the previous two tasks.

The rest of this paper is organized as follows. Section 2 presents related work, describing contextual advertisement approaches and techniques for mining geographic information from documents. Section 3 describes the approaches for assigning geographic scopes to Web documents. Section 4 describes the classification of documents according to locational relevance. Section 5 describes the proposed geographic information retrieval approach for selecting the most relevant ads. Section 6 presents the experimental validation of the proposed approaches. Finally, Section 7 presents our conclusions and directions for future work.

2. RELATED WORK

In this section, we describe previous research on online textual advertisement, which is often formulated as an Information Retrieval (IR) problem. We also present previous research on geographic information retrieval, a specialization of IR that addresses issues related to the geographic relevance of documents.

2.1 Contextual Advertisement

To a large extent, contextual advertisement can be framed as a traditional document retrieval problem, where the ads are the "documents" to be retrieved given a query composed of a target page. Thus, one way of approaching it is to represent the target Web page as a set of keywords, in order to retrieve the ads that match those same keywords. From this perspective, Yih et al. proposed a system for keyword extraction from target pages [2]. Their system uses a variety of features (e.g. TF/IDF, HTML metadata, query logs) to determine the importance of phrases (i.e., sequences of up to 5 words) extracted from the target pages. Yih et al. suggests that finding the most important keywords is already a critical task in well-known contextual advertising systems.

A complementary approach was reported by Lacerda et al. in [14]. The authors focused on the selection of good ranking functions for matching ads to documents, proposing to use a genetic programming algorithm to generate a non-linear combination of traditional IR term weighting heuristics that maximizes the average precision on retrieving ads.

In our work, we perform the same keyword extraction and ad ranking operations. However, since these are not the main focus of this paper, we used a state-of-the-art commercial keyword extraction service, namely the Yahoo! Term Extractor[3], and performed the ad ranking using a linear combination of textual and geographic similarity measures, using heuristics to fine-tune the combination weights.

In general, the effectiveness of an ad is strongly affected by the similarity between the ad and the context where it appears. Assuming access to the text of the target page, to the keywords declared by an advertiser (i.e., the bid terms), and to the textual description of the ad, Ribeiro-Neto et al. examined a number of strategies for matching pages to ads based on keywords, concluding that one of the main problems is that ads contain very little text [15]. The authors used the standard vector space model to represent both the ads and the pages, examining different strategies to improve the matching process. Some of the proposed strategies matched pages and ads based on the cosine of the angle between their respective vectors, using different ad sections (e.g., bid terms, title and body) as the basis for the ad vector. The winning strategy reported by Ribeiro-Neto et al. required that at least one of the bid terms appeared on the target page, and then ranked ads by the cosine of the union of all the ad sections and the page vectors.

The same authors also noted that while both target pages and ads are mapped to the same space, there is a discrepancy between the vocabulary used in the ads and in the pages. For example, the standard vector space model cannot easily account for synonyms, i.e., it cannot easily match target pages and ads that describe related topics using different vocabularies. The authors improved the matching precision by expanding the page vocabulary with terms from similar Web pages, weighted through a Bayesian model based on their similarity to the original page.

Noting that, when no ads are relevant to the visitor's interests, showing ads would produce no economic benefit, Broder et al. approached the decision problem of *whether to swing*, i.e., whether or not to show any ads for the incoming request [21]. The authors experimented with two different

[3] `http://developer.yahoo.com/search/content/V1/termExtraction.html`

approaches for addressing the problem, first through simple thresholding and afterwards through a machine learning approach. In both cases, the idea was to classify target pages as either relevant for advertisement or not. Jin et al. studied similar ideas for classifying the target pages, with the specific intent of finding sensitive Web pages where advertisements should not be placed [10]. In this paper, we also use a similar machine learning approach. In our case, however, we use it to decide whether or not it would be interesting to place local advertisements on a target Web page.

2.2 Geographic Information Retrieval

Geographic information is pervasive on the Web and building IR systems capable of exploring this information is a research problem that is getting increasing attention [24]. Previous works in the area of Geographic Information Retrieval (GIR) have addressed issues such as the recognition and disambiguation of place references in text, the assignment of geographic scopes to documents, or the retrieval of documents taking geographic relevance into account.

Leidner presented a variety of approaches for handling place references in textual documents [9]. The problem is usually seen as an extension of the named entity recognition task (NER), as proposed by the natural language processing community [9, 17]. More than recognizing mentions of places over text, which is the subject of NER, the task also requires for the place references to be disambiguated into the corresponding locations on the surface of the Earth, assigning geospatial coordinates to the place references. Disambiguation is often dealt using heuristics such as default senses (i.e., disambiguation should be made to the most important referent, estimated using population counts) or spatial minimality (i.e., disambiguation should minimize the bounding polygon that contains all candidate referents) [9]. Metacarta, a company that sells state-of-the-art geographic information retrieval technology, provides a freely-available Web service that can recognize and disambiguate place references in text. An early version of the Metacarta geotagger[4] was described by Rauch et al. in [20].

Reference disambiguation usually relies on finding the placename in a dictionary of known locations, a.k.a. a *gazetteer*. Geonames is an example of a modern gazetteer, having been used in many previous experiments concerning the handling of place references over textual documents [19].

The automatic assignment of geographic scopes to Web documents, based on the place references that are present in the text, is an example of a complex GIR problem that has been getting increasing attention. Given a set of diverse geographic regions, corresponding to the placenames mentioned in a given Web page, the problem concerns finding the geographic region that best summarizes and describes them all. While several different strategies have been proposed in the past, until this work there was no clear information about the trade-offs involved in choosing a particular algorithm. Each different algorithm makes specific assumptions, therefore resulting in different approximations for the geographic scope of the documents.

In our experiments, we used the methods proposed by Amitay et al. [23], Woodruff and Plaunt [6], and Martins and Silva [3], as well as some baselines, to assign geographic scopes to the target Web pages.

Another of our goals was to classify target pages as either locationally relevant or not. To this effect, we use a technique similar to the one proposed by Gravano et al. [22], in which search engine queries are classified as either local or global, using the distributional characteristics of location names occurring in the search results. As classification features we used both place references and text terms.

Existing approaches for retrieving documents according to geographic relevance are mostly based on combinations of the standard IR metrics used in text retrieval (e.g. cosine similarity of TF/IDF vectors) with similarity metrics for geographic scopes, based on distance and/or containment [11]. Larson and Frontiera compared the performance of different methods for computing spatial similarity scores for query-document pairs [13]. A simple method proposed originally by Greg Janée [5] performed almost as well as a highly-tuned method based on logistic regression. In Janée's method, the similarity between two regions S_1 and S_2 is given by the formula bellow:

$$sim(S_1, S_2) = \frac{area(S_1 \cap S_2)}{area(S_1 \cup S_2)} \quad (1)$$

Martins et al. proposed a similarity function for GIR that, instead of using area overlaps, uses a non-linear normalization of the distance between the document and query scopes [11]. The normalization is done through a double sigmoid function with the center corresponding to the diagonal distance of the rectangular region for to the query scope. The similarity is maximum when the distance is zero, and smoothly decays to 0 as the distance increases. The formula proposed by Martins et al. is given bellow.

$$sim(S_1, S_2) = \begin{cases} 1 \text{ if } S_1 \text{ is contained in } S_2 \\ 1 - \frac{1+sign(D)\times(1-e^{-(\frac{D}{d\times 0.5})^2})}{2} \text{ otherwise} \end{cases} \quad (2)$$

where d refers to the diagonal distance of the rectangular region corresponding to the query's geographic scope S_2 and $D = centroidDistance(S_1, S_2) - d$.

Cai proposed the GeoVSM framework for retrieving geographic information [18]. The author argues that thematic and geographic similarity should be computed independently, and afterwords combined into a single similarity value.

3. ASSIGNING GEOGRAPHIC SCOPES

The first stage in the proposed pipeline of processing operations involves assigning geographic scopes to target Web pages. In turn, this stage involves two separate sub-tasks, namely i) recognizing and disambiguating place references in the text, and ii) assigning geographic scopes to documents based on the disambiguated place references. This section presents both sub-tasks in more detail.

3.1 Handling Place References in Text

For handling place references in text, we relied on the Placemaker[6] text mining Web service provided by Yahoo!. This

[4] `http://ondemand.metacarta.com/`

[5] `http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html`

[6] `http://developer.yahoo.com/geo/placemaker`

service provides functionalities for recognizing and disambiguating place references over text, returning an unique identifier and a confidence score for each reference recognized in a document. Using this identifier it is possible to query the Yahoo! GeoPlanet[7] gazetteer service, and obtain further information on the location. This way, each of the resolved place references is associated with the corresponding city, state, country and continent, as well as with the bounding rectangle that covers its area.

Since Yahoo! Placemaker uses natural language contextual clues, the service can often disambiguate whether a word like "Reading" refers to the location in England or to the verb sense of to read. It also covers many colloquial location names (e.g. nyc for "New York City"), as well as geo-referenced interest points (e.g. Eiffel Tower) that may appear in the text of the target pages.

3.2 Assigning Geographic Scopes to Pages

We tested seven different approaches for assigning geographic scopes to target pages, namely the methods proposed in the context of the Web-a-Where [23], GIPSY [6] and GREASE [3] projects, as well as four baseline methods. In all cases, we used the results of the Yahoo! Placemaker as the source of disambiguated place references.

The Web-a-Where technique leverages on part-of relations among the recognized place references, provided by a hierarchical gazetteer. The basic idea is that, for instance, if several cities from the same country are mentioned, this might mean that this country is the scope, i.e. the algorithm tries to generalize from the disambiguated place references. More specific places are scored higher if they are the only places mentioned, or if they are mentioned many times.

The algorithm starts by building a geographical hierarchy from the disambiguated place references. By looping over these references, it aggregates the confidence scores from lower levels in the hierarchy. The references are then sorted by score and the highest is chosen as the scope.

Differently, the GIPSY algorithm works by disambiguating the place references into the bounding boxes that correspond to their area over the surface of the Earth. The geographic scope of the document is afterwards computed using the overlapping area for all the boxes, thus trying to find the most specific place that is related to all the place references made in the document.

For the GIPSY algorithm, the bounding boxes are seen as thick polygons, with a base positioned at an (x, y) plane, but extending upwards a distance of z, to a higher parallel plane. One by one, in decreasing order of size, the bounding boxes corresponding to the place references are analysed by the GIPSY algorithm, in order to build a skyline of bounding boxes. Finally, all the bounding boxes would be sorted according to their z order and the highest ranking bounding box is selected as the scope. In our implementation, each bounding box has a thickness z equal to the number of times its respective place name occurs, weighted according to the confidence score given by the disambiguation task. For computing the 2D spatial operations, we used the open-source Java Topology Suite [8].

In the context of the GREASE project, Martins and Silva proposed a scope assignment method based on a graph-ranking approach [3]. The idea was to represent the gazetteer used for place reference disambiguation as a graph, where the nodes correspond to different places and the edges correspond to semantic relationships (*part-of*, *containment* or *adjacency*) between places.

Nodes on this graph can be weighted according to the occurrence frequency of place references in a document, and edges can be weighted according to the relative importance of the different types of relationships. A graph-ranking algorithm, *PageRank*, is then applied to this graph, and finally the highest ranked node is selected as the scope. In case of ties, the node connected to the highest number of edges is selected. By propagating scores across the graph, this algorithm tries, at the same time, to generalize and to specify from the available information, in order to find the region that best reflects the scope of the document. For computing the PageRank score, we used the open-source weighted PageRank implementation made available by the Laboratory for Web Algorithmics of the University of Milan [4].

The three previously described methods make non-trivial assumptions about how place references should be combined to discover the geographic scope of a document. In order to assess what are the gains introduced by these assumptions, we implemented three simple baseline methods, which we describe as follows.

The number of times a place is referenced in a document reflects the importance of that place to the document's subject. We therefore experimented with a simple scope assignment method that chooses the most frequently occurring place reference as the scope. In case of ties, the place reference corresponding to the largest area is chosen.

The different place references made in the document should all contribute to the document's scope. We therefore experimented with a simple scope assignment method that computes the bounding box that covers all the place references made in the document.

Only the place references that are somewhat interrelated should be considered. The idea is to be able to filter the errors made while recognizing and disambiguating place references, as well as filtering out the place references that are only tangential to the content of the document. We first compute the average centroid point for all the place references made in the document, as well as the average distance between the place references and this centroid. Then, we filter out those place references whose centroid is at a distance that is greater than twice the average distance value. Finally, we assign a scope corresponding to the bounding box that covers all the remaining place references, if none the closest is chosen. This baseline is inspired on a technique proposed by Smith and Crane for place name disambiguation [16].

The last considered method was the Yahoo! Placemaker Web service, which can also assign scopes to the Web documents. We use this Web service as a black box, only to

[7] `http://developer.yahoo.com/geo/geoplanet`

understand the current performance of commercial applications.

4. CLASSIFYING TARGET PAGES

The second stage of the proposed pipeline of processing operations involves classifying target pages according to their locational relevance, i.e. classifying them as either local or global, so that locally relevant ads are mostly placed on the target pages that are more interesting for them. For example, a target page on the subject of computer programming can be considered global, as it is likely to be of interest to a geographically broad audience. In contrast, a document listing pharmacies or take-away restaurants in a specific city could be regarded as a local, likely to be of interest only to an audience in a relatively narrow region.

In the context of this work, locational relevance is therefore a score that reflects the probability of a given document being either global, meaning that users interested in the document are likely to have broad geographic interests, or local, meaning that users interested in the document are likely to have a single narrow geographic interest.

Assigning documents to global and local classes, according to their implicit locational relevance, can be naturally formulated as a binary classification problem. However, instead of applying the standard classification approach, based on a bag-of-words representation of the documents, we argue for the use specific features, better suited to reflect the locational characteristics. The features we propose attempt to capture two different but complementary aspects of locational relevance, namely the geographic information inferred from the distribution of place references in the text and geographic scopes, since we can assume that local documents are more likely to contain a cohesive set of place references, and the thematic relatedness to subjects that are typically regarded as local, since words like "restaurant" or "hotel" are more often associated to local pages than words like "tutorial" or "mp3".

We group the considered features into three sets, namely (i) textual features, e.g. TF-IDF weights for term stems, (ii) simple locative features, e.g. counts for the different types of geographical references made in the document, and (iii) high level locative features, e.g., spatial area of the geographic scopes obtained through different algorithms.

We chose a classifier based on Support Vector Machines (SVMs) because SVMs represent state-of-the-art classification technology and have been shown to be highly effective for a variety of other text classification tasks [12]. Moreover, they offer the possibility to assign a value in the interval [0,1] that estimates the likeliness of the document being either local or global. In particular, we used a gaussian-based SVM with parameters C and γ optimized using the grid-search functionality offered by Weka [5]. The final classifier scores correspond to probabilities for the documents being either local or not, according to the procedure described by Lin et al. [7].

5. GEOGRAPHIC RETRIEVAL OF ADS

The final stage of the proposed pipeline of processing operations involves retrieving and ranking ads based on a combination of both thematic and geographic relevances.

As previously stated, contextual advertising can be interpreted as a search problem over the corpus of ads. Ads are, in our case, represented as a bag of words, where the words come from ad fields like a title, a small textual description, and a set of descriptive keywords. Additionally, ads are also associated with a geographic scope. The query triggering the search is derived from the context of the target Web page (the text and the geographic scope) where the ads are to be displayed. Since users are unlikely to click on irrelevant ads, systems should attempt to maximize ad relevance.

For combining multiple sources of relevance into a single ranking, we follow the GeoVSM framework proposed by Cai [18]. As shown in Equation 3, GeoVSM independently computes a geographic similarity gs and a thematic similarity ts, which are then combined through a function f.

$$Rel(doc, ad) = f(ts_{\{doc,ad\}}, gs_{\{doc,ad\}}) \tag{3}$$

We argue that, for the contextual advertising problem, a linear combination of relevance scores that uses the proposed locational relevance as the weight, as shown in Equation 4, is an adequate function f. In the context of multimedia information retrieval, Wu et al. [27] demonstrated that a linear combination might be sufficient when fusing a small number of relevance rankings from different domains, as in this case. Also, contrary to the traditional static weights, the locational relevance score provides a weighting scheme that dynamically adapts itself to each document.

$$f(ts, gs) = (1 - w) \cdot ts + w \cdot gs \tag{4}$$

We experimented with two different approaches for measuring thematic relevance, namely the similarity between document key terms and terms from the ads, and the similarity between the full-text of the document and the terms from the ads. Our implementation relies on the full-text search capabilities provided by the PostgreSQL database management system[8]. However, since PostgreSQL does not provide a thematic similarity value v in the range [0,1], we applied the min-max normalization presented in Equation 5.

$$v' = \frac{v - min}{max - min} \tag{5}$$

For geographic relevance, we also experimented with two different strategies, namely the normalized distance originally proposed by Martins et al. [11], and the relative area of overlap proposed by Janée. Both equations were described in Section 2.2. For the implementation, we used the PostGIS extension[9] for PostgreSQL, in order to compute distances and area overlaps.

6. EXPERIMENTAL EVALUATION

In this section, we describe the details of our empirical evaluation. This includes our experimental design for evaluating

[8] `http://www.postgresql.org`
[9] `http://www.postgis.org`

the effectiveness of the proposed approaches, as well the obtained results in the different experiments.

6.1 Assigning Scopes to Target Pages

We evaluated the algorithms described in Section 3 for assigning scopes to the target pages by comparing their assignments to those of the human editors from the Open Directory Project[10] (ODP). Specifically, we took a sample of 6,000 Web-pages classified under the ODP's *Regional/North_America/United_States* section. The pages were written in English, had more than 2 KBytes, and contained at least one geographic reference. The human-assigned scopes for the pages in this collection were equally distributed across scopes corresponding to the entire country, states, and cities. There were a total of 1,100 unique scopes.

We ran all algorithms over the test collection, measuring the distance and the relative overlap between the scopes that were assigned by the algorithms and the scopes that were assigned by the human editors of ODP. The relative overlap was measured using the scheme proposed by Greg Janée, which we described in Section 2.2. We also measured the accuracy for both exact matches and approximate matches. Table 1 summarizes the obtained results.

When interpreting the results, it should be noted that the scope assignment algorithms use the information provided by the Yahoo! Placemaker about the individual places mentioned in the text. Any errors made by the geotagger influence the outcome of the scope assignment methods (i.e., this test only evaluates geotagging plus scope assignment as a whole). Nevertheless, this does not invalidate the goal of the experiments, which is to compare two scope assignment methods under the same conditions.

The Web-a-Where and GraphRank algorithms obtained the best overall performances, with errors more equally distributed across countries, states, and cities. Nonetheless, both were particularly good with country scopes, which was already expected due to their tendency to propagate place reference scores towards their encompassing region. The GIPSY method performed well in both average distance and accuracy for approximations bellow 100Km, although it had a weak performance in terms of exact matches. The algorithm privileges narrow regions and often fails in generalizing from the available place references.

Regarding the baselines, the results show that the one based on the covering area is the weakest approach to the task of geographic scope resolution, producing the worst results on most metrics. Removing the outliers does substantially improve its results, but this is also a very weak baseline in terms of overall performance, specially when dealing with narrow scopes. On the contrary, the baseline that simply assigns the most frequent location as the scope proved to be a very competitive approach, regularly outperforming other methods on pages with scopes corresponding to states or cities.

The last considered method was the Yahoo! Placemaker, which obtained the best results on most metrics for pages referring to country scopes, but modest to weak results on pages with state and city scopes. This suggests that the Placemaker service has a tendency to overgeneralise multiple locations towards their encompassing regions.

Having discussed the evaluation results of all methods, it is necessary to decide which one is the most suitable for the contextual advertising problem. One of the assumptions behind this work is that the more specific the geographic scope is, the more interesting it will be to the user. As such, pages about countries or continents will not often be considered good candidates for displaying geotargeted advertisements based on the geographic scope. It is therefore important for the chosen method to be especially good in determining the scopes in pages with narrow geographic scopes. Thus, the high performance across most metrics by the approach assigning the most frequent location as the scope suggests that this baseline is the best choice.

6.2 Locational Relevance Classification

In order to evaluate the locational relevance classifier, a dataset consisting of 8,000 Web pages crawled from the ODP was developed. The dataset was formed by 4,000 pages classified as *local* and 4,000 pages classified as *global*. Pages under small locations in the "Regional" portion of the directory were regarded as local (i.e., US cities and US states), while pages outside the "Regional" category or under a large region (i.e., USA) were regarded as global. All pages were written in English, has at least on place reference, and were randomly selected from the various ODP thematic categories.

The experiments considered three classifier configurations, i.e., (i) based only on textual features, (ii) based only on simple locative features, (iii) based only on high level locative features, and (iv) a combination of both textual and the best locative features. Table 2 overviews the results for each of the different configurations. All results correspond to values obtained after a 10-fold cross validation.

An analysis of the results shows that the combination of textual and simple locative features has the best performance, achieving an accuracy of 90.7%, although both the locative or the textual features alone are enough for achieving good results. The classifier based on the high level locative features had worse results than anticipated, most certainly caused by the low overall effectiveness of geographic scope algorithms, as discussed in the previous section.

By performing an information gain analysis, we observed that the top most descriminative features included the locative features plus word stems like *park*, *local* or *hotel*.

Instead of a binary classification approach (i.e., local vs. global), we also experimented with building a classifier that considered four classes: *city-local*, *state-local*, *country-local*, and *non-local*. The same set of ODP web pages was used in this experiment and the feature set considered was the textual features plus the simple locative features. An accuracy of 78.6% was achieved, indicating that it is feasible to rank pages as more or less locative.

6.3 Geographic Retrieval of Ads

To measure the impact of introducing geographic similarities when retrieving ads, we experimented with different combinations of thematic and geographic similarity measures.

[10] `http://www.dmoz.org`

	Level	Average Distance (Km)	Average Overlap	Accuracy Distance=0	Accuracy Distance<100Km	Accuracy Overlap>0.75
GIPSY	Country	2986	0.07	0.07	0.07	0.07
	State	**442**	0.22	0.19	0.41	0.21
	City	**398**	0.37	0.16	**0.81**	0.32
	All	1275	0.22	0.14	0.43	0.2
Web-a-Where	Country	1336	0.59	0.54	0.54	0.54
	State	855	0.51	0.5	0.55	0.5
	City	704	0.42	**0.39**	0.58	0.4
	All	**959**	**0.51**	**0.48**	0.56	**0.48**
GraphRank	Country	1048	0.64	**0.61**	**0.61**	**0.61**
	State	925	0.52	0.51	0.55	0.51
	City	1281	0.34	0.33	0.47	0.34
	All	1085	0.5	**0.48**	0.54	**0.48**
Most Frequent	Country	2250	0.36	0.35	0.35	0.35
	State	501	**0.54**	**0.52**	**0.63**	**0.53**
	City	549	**0.47**	0.24	0.74	**0.45**
	All	1100	0.46	0.37	**0.57**	0.45
Covering Area	Country	2190	0.47	0	0.3	0.31
	State	3158	0.23	0	0.21	0.18
	City	2632	0.05	0	0.13	0.05
	All	2660	0.25	0	0.21	0.18
Non-outliers	Country	1523	0.57	0.45	0.5	0.55
	State	1838	0.38	0.24	0.39	0.36
	City	1872	0.12	0.02	0.28	0.1
	All	1744	0.35	0.24	0.39	0.34
Placemaker Admin.	Country	**774**	**0.71**	**0.61**	**0.61**	**0.61**
	State	1173	0.44	0.42	0.46	0.43
	City	1125	0.12	0.05	0.28	0.1
	All	1033	0.42	0.36	0.45	0.38

Table 1: Comparison of scope assignment methods.

	Recall		Precision		F-Measure			
	Local	Global	Local	Global	Local	Global	Error	Accuracy
Text	0.81	0.83	0.82	0.81	0.82	0.82	18.4	81.6
Simple Locative	0.92	0.73	0.78	0.9	0.85	0.81	17.1	82.9
High Level Locative	0.75	0.67	0.7	0.73	0.72	0.7	28.8	71.2
All Locative	0.82	0.79	0.8	0.81	0.81	0.8	19.5	80.5
Text + Best Locative	0.92	0.89	0.9	0.92	0.91	0.91	9.3	**90.7**

Table 2: Obtained results for the page classification algorithm, using different combinations of features.

Specifically, we combined the following three groups:

- Thematic similarity, based on the full text of the web pages (T1), and only on the most relevant terms (T2)
- Geographic similarity, based on the normalized distance (G1), and on the relative area of overlap (G2)
- Combined weighting schemes, using a baseline which assigns a weight of 0.5 to each similarity (W1), and using the locational relevance score (W2).

Tests were made using two collections of target pages, namely local and global sets of twenty pages from the ODP. The local dataset had 20 pages taken from regional sub-sections with topics like `Real Estate Guides` or `Travel Guides`. The global dataset had 20 pages taken from outside the regional section, and belonging to topics like `Computer Guides` or `Investing Guides`. The pages had advertisements, and at least one geographic reference. We also developed an advertisements collection, using the title, description, and geographic scope assigned by the ODP editors for the pages associated with `Business` categories. The ad's keywords were obtained by retrieving the most important words from its Web page, using the Yahoo! Term Extraction service.

In order to evaluate how the different retrieval strategies deal with the two scenarios, i.e., pages where the geography is important and pages where it is not, we considered an ad to be relevant to the pages on the local dataset when it belongs to a related thematic category and has the same geographic scope. One disadvantage of this approach is that we are ignoring potentially interesting ads from nearby locations. As for pages on the global dataset, the relevant ads are the ones from a related thematic category and with an inexistent or country-wide geographic scope. The choice of related thematic categories was checked by hand. For instance, a related thematic category for target pages under `Literature` would be `Shopping/Books`.

Table 3 overviews the results for each of the different datasets and combinations, measuring the precision at different cut-off points, as well as the mean average precision for the top five results.

An analysis of the results shows that the combination of thematic and geographic similarities using the locational relevance as weight has the best performance over the local dataset. Over the global dataset the locational relevance scheme only looses to the keywords-only approach. This may be explained by the fact that despite these target pages have a low locational relevance, it is enough to retrieve ads with a high thematic relevance. In this case, using geographic sim-

		P@1	P@3	P@5	MAP
Local Dataset	T1	0.2	0.15	0.15	0.21
	T2	0.2	0.22	0.21	0.29
	G1	0.1	0.07	0.06	0.1
	G2	0.05	0.07	0.06	0.1
	G1T1W1	0.45	0.35	0.35	0.42
	G1T1W2	0.45	0.35	0.35	0.42
	G2T1W1	0.4	0.48	0.43	0.52
	G2T1W2	**0.45**	**0.48**	**0.45**	**0.54**
	G1T2W1	0.35	0.33	0.32	0.36
	G1T2W2	0.35	0.33	0.32	0.36
	G2T2W1	0.4	0.43	0.41	0.48
	G2T2W2	0.4	0.43	0.41	0.48
Global Dataset	T1	0.25	0.37	0.28	0.36
	T2	**0.6**	**0.43**	**0.4**	**0.58**
	G1	0	0	0	0
	G2	0	0	0	0
	G1T1W1	0.15	0.08	0.07	0.16
	G1T1W2	0.2	0.15	0.13	0.2
	G2T1W1	0.15	0.16	0.14	0.2
	G2T1W2	0.25	0.2	0.2	0.24
	G1T2W1	0.3	0.23	0.19	0.35
	G1T2W2	0.3	0.22	0.2	0.31
	G2T2W1	0.35	0.23	0.24	0.32
	G2T2W2	0.35	0.28	0.28	0.37

Table 3: Comparison of retrieval performance.

ilarity introduces noise. For future work, we plan on testing a thresholding approach in order to only consider geographic similarity for pages with a high locational relevance. Overall, results seem to indicate that the locational relevance does successfully adapt according to the geographical interest of the page. Also, computing the thematic similarity based on the keywords instead of the full-text of the pages produces better results on both sets. As for the geographic similarity, using the normalized distance instead of the relative overlap produces the best results over the local dataset. According to this, using the GIPSY method for scope resolution might be a better option for this problem, since it produces the best average distances.

7. CONCLUSIONS AND FUTURE WORK

The contextual advertisement task introduces new challenging technical problems and raises interesting questions to IR practitioners. In this work, we studied the application of techniques from the area of geographical information retrieval to the problem of geotargeting Web advertisements.

We address the task through a pipeline of processing stages, which involves (i) determining the geographic scope of the target pages, (ii) classifying target pages according to locational relevance, and (iii) retrieving relevant ads to the target page, based on both the textual content and the geographic scope. An experimental evaluation for the methods proposed in each of the individual sub-tasks was made by leveraging on Web pages from the Open Directory Project, using specific parts of the directory to simulate both the advertisements and the target Web pages.

The method originally proposed in the Web-a-Where system achieved the best overall results for the task of assigning geographic scopes to documents. However, a method that simply assigns the most frequent location as the scope produced higher results for pages with narrow scopes.

An SVM classifier combining features related to textual terms with features related to the geographic dispersion achieves an accuracy of 90.7 on the task of classifying target pages as locationally relevant or not.

A linear combination of textual similarity and normalized geospatial distance, where individual scores are weighted according to the locational relevance of the document, achieves the best results over the local documents, yielding a MAP of 0.54. On global pages, the weighting scheme based on the locational relevance classifier outperformed the other combination approaches, although it is outperformed by the text-only approaches. Applying a threshold to the locational relevance, forcing only geographically relevant results to be considered, might be a simple option for improving the results over the global pages.

Despite the promising results, there are also many challenges for future work, the remaining of this work will discuss the most important ones.

The experimental results showed significant differences in the geographic scope algorithms and it would be interesting to see if a combination of the best algorithms could lead to better results. Moreover, by taking the best algorithm for each document in the test collection, it was possible to obtain an accuracy of 70%, showing that this is indeed a promising alternative

Experiment with link-based features on the locational relevance classifier. Previous research on Web document classification has shown that better performance can be achieved trough combinations of content-based features with additional features derived from the neighbouring documents using the link structure of the Web graph [30].

Devise experiments for finding a locational relevance threshold, capable of ensuring that the geographic similarity is only considered on geographically relevant Web pages. Such technique might be a simple option for improving the retrieval results of the proposed weighting scheme over the global pages.

Using more sophisticated *learning to rank* approaches, in order to fine-tune the ranking method, is also a promising direction for future work. Learning to rank techniques could be used to explore a larger set of features in a non-linear way, in order to retrieve the most relevant ads [1].

Experiment with a retrieval model that works with semantic information from multiple domains (e.g., thematic, geographic, and temporal domain). The development of a Web-scale retrieval approach that works with concepts and relationships, rather than individual words, is gaining increasing attention by the scientific community [28, 29]. For the particular purpose of this work, instead of combining the results from different retrieval engines, and determining an individual locational relevance score for each document, a retrieval approach with basis on concepts and relationships can implicitly determine the most relevant ads with basis on more than just thematic relevance. Such concept-based approach is the focus of ongoing work.

8. REFERENCES

[1] T. Joachims (2002) Unbiased evaluation of retrieval quality using clickthrough data. Technical report,

Cornell University, Department of Computer Science.

[2] W. tau Yih, J. Goodman, and V. R. Carvalho (2006) Finding advertising keywords on web pages. In Proceedings of the 15th international conference on World Wide Web, pages 213–222.

[3] Martins, B., and Silva, M. J. (2005) A Graph-Ranking Algorithm for Geo-Referencing Documents, In Proceedings of the 5th IEEE International Conference on Data Mining.

[4] Boldi, P., Santini, M., and Vigna, S. (2005) PageRank as a function of the damping factor. In Proceedings of the 14th International World Wide Web Conference.

[5] I. H. Witten, and E. Frank (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann.

[6] A. G. Woodruff, and C. Plaunt (1994) GIPSY: automated geographic indexing of text documents. Journal of the American Society of Information Sciences, 45(9), pages 645-655.

[7] H. Lin, C. Lin, and R. Weng (2007) A note on Platt's probabilistic outputs for support vector machines, Machine Learning, 68(3), pages 267-276

[8] M. Johansson, and L. Harrie (2002) Using Java Topology Suite for real-time data generalization and integration. In proceedings of the 2002 workshop of the International Society for Photogrammetry and Remote Sensing.

[9] J. L. Leidner (2007). Toponym Resolution: a Comparison and Taxonomy of Heuristics and Methods. PhD Thesis, University of Edinburgh.

[10] X. Jin, Y. Li, T. Mah, and J. Tong (2007) Sensitive webpage classification for content advertising. In Proceedings of the 1st international Workshop on Data Mining and Audience intelligence For Advertising, pages 28-33.

[11] B. Martins, N. Cardoso, M. S. Chaves, L. Andrade, and M. J. Silva (2007) The University of Lisbon at GeoCLEF 2006. Evaluation of Multilingual and Multi-modal Information Retrieval, pages 986-994.

[12] T. Joachims (1998) Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In Proceedings of the 10th European Conference on Machine Learning, pages 137-142.

[13] P. Frontiera, R. Larson, and J. Radke (2008) A comparison of geometric approaches to assessing spatial similarity for GIR. International Journal of Geographic Information Sciences, 22(3), pages 337-360.

[14] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto (2006) Learning to advertise. In Proceedings of the 29th ACM SIGIR Conference on Research and Development in information Retrieval, pages 549-556.

[15] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. Silva de Moura (2005) Impedance coupling in content-targeted advertising. In Proceedings of the 28th ACM SIGIR Conference on Research and Development in information Retrieval, pages 496-503.

[16] Smith, D. A. and Crane, G. (2001) Disambiguating Geographic Names in a Historical Digital Library. In Proceedings of the 5th European Conference on Research and Advanced Technology For Digital Libraries.

[17] A. Kornai (2003 eds.) Proceedings of the HLT-NAACL 2003 workshop on the analysis of geographic references.

[18] G. Cai (2002) GeoVSM: An Integrated Retrieval Model For Geographical Information. In Proceedings of the 2nd International Conference on Geographic Information Science, pages 65–79.

[19] M. Wick, and T. Becker (2007) Enhancing RSS Feeds with Extracted Geospatial Information for Further Processing and Visualization. in A. Scharl, and K. Tochtermann eds. The Geospatial Web - How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society, Springer, pages 105-115.

[20] E. Rauch, M. Bukatin, and K. Baker (2003) A confidence-based framework for disambiguating geographic terms. In Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, pages 50-54.

[21] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras (2008) To Swing or not to Swing: Learning when (not) to Advertise. In Proceeding of the 17th ACM Conference on Information and Knowledge Management, pages 1003-1012.

[22] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein (2003) Categorizing web queries according to geographical locality. In Proceedings of the 12th international Conference on information and Knowledge Management, pages 325-333.

[23] E. Amitay, N. Har'El, R. Sivan, and A. Soffer (2004) Web-a-where: geotagging web content. In Proceedings of the 27th ACM SIGIR Conference on Research and Development in information Retrieval, pages 273-280.

[24] R. Jones, W. V. Zhang, B. Rey, P. Jhala, and E. Stipp (2009) Geographic intention and modification in Web search. International Journal of Geographical Information Science, 22(3), pages 229-246.

[25] C. Wang, P. Zhang, R. Choi, and M. D. Eredita (2002) Understanding consumers attitude toward advertising. In Proceedings of the 8th Americas Conference on Information Systems, pages 1143–1148.

[26] C. Guo, Y. Liu, W. Shen, H. Wang, Q. Yu, and Y. Zhang (2009) Mining the Web and the Internet for Accurate IP Address Geolocations. In Proceedings of the 28th IEEE Conference on Computer Communications.

[27] Wu, Y., Chang, E. Y., Chang, K. C., and Smith, J. R. (2004) Optimal multimodal fusion for multimedia data analysis. In Proceedings of the 12th annual ACM international conference on Multimedia, pages 572-579.

[28] Dalvi, N., Kumar, R., Pang, B., Ramakrishnan, R., Tomkins, A., and Bohannon, P. (2009) A Web of Concepts. In Proceedings of the 28th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pages 1-12.

[29] Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., and Attardi, G. (2007) Ranking very many typed entities on wikipedia. In Proceedings of the 16th ACM Conference on Conference on information and Knowledge Management, pages 1015-1018.

[30] Qi, X., and Davison, B. D. (2006). Knowing a web page by the company it keeps. In Proceedings of the 15th ACM international conference on Information and knowledge management, pages 228-237.