Using the Geographic Scopes of Web Documents for Contextual Advertising

Ivo Anastácio, Bruno Martins, Pável Calado

INESC-ID Lisboa / Instituto Superior Técnico, PT

{ivo.anastacio, bruno.g.martins, pavel.calado}@ist.utl.pt

ABSTRACT

Geotargeting is a specialization of contextual advertising where the objective is to target ads to Website visitors concentrated in well-defined areas. Current approaches involve targeting ads based on the physical location of the visitors, estimated through their IP addresses. 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 general area implied by the locations mentioned in the textual contents of the pages. Our proposal applies 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 scope of target pages, (ii) classifying target pages according to locational relevance, and (iii) retrieving ads relevant to the target page, using both textual contents and geographic scopes. 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

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

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 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 tied with profitability (i.e., more relevant ads lead to improved user satisfaction and higher response rates) [19]. 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 visitors, estimated using their IP addresses [7]. 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 also many situations where it would be more interesting to target ads based on the geographic scope described in the content of the target pages.

Consider a user who is traveling to Lisbon and browsing Web pages describing tourist attractions and events in the city. Here 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 text 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 geographical proximity or containment (e.g., the name Lisbon would not match places that correspond to sub-regions, like Chiado).

A recent trend in IR applications relates to extracting geo-

graphic context information from textual documents, in order to explore it for purposes of document retrieval [10, 14, 1]. 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 pages and assign them to geographic scopes, (ii) classify the target pages as either local or global, using features like the text or the extracted place references, and (iii) find ads relevant to the target pages by using GIR ranking techniques that combine 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 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, leveraging on the results obtained from the previous two tasks.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 describes the approaches for assigning geographic scopes to Web pages. Section 4 describes the classification of target pages according to locational relevance. Section 5 describes GIR approaches for finding the most relevant ads. Section 6 presents the experimental validation of the proposed approaches. Finally, Section 7 presents conclusions and directions for future work.

2. RELATED WORK

This section describes previous research on contextual advertisement, which is often formulated as a retrieval problem. We also present previous research on Geographic Information Retrieval (GIR), a specialization of IR that addresses issues related to the geographic relevance of documents.

2.1 Contextual Advertisement

Contextual advertisement can be framed as a 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 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 [23], arguing that this is already a critical task in well-known contextual advertising systems. The authors use 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.

Complementary approaches have been reported by Lacerda et al. [9] and by Ribeiro-Neto et al. [16]. Lacerda et al. focused on the selection of good ranking functions for matching ads to pages, using genetic programming for generating a non-linear combination of term weighting heuristics that maximizes the average precision on retrieving ads. Ribeiro-Neto et al. examined different strategies for matching pages to ads, based on keywords. The winning strategy required that at least one of the keywords declared by an advertiser appeared on the target page, and ranked ads by the cosine of the angle between the ad and page vectors. These authors also concluded that one of the main problems in contextual advertisement is that ads contain little text and often use a different vocabulary than that of the pages.

In our work, we perform the same keyword extraction and ad ranking operations. However, since these are not our main focus, 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.

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 an incoming request [3]. The authors experimented with a simple thresholding approach and with machine learning. In both cases, the idea was to classify target pages as either relevant for advertisement or not. Here we use a similar approach, but in our case to decide whether or not a target Web page is apropriate for placing local advertisements.

2.2 Geographic Information Retrieval

Geographic information is pervasive on the Web and exploring this information is a research problem that is getting increasing attention [8, 12]. Previous works in the area of Geographic Information Retrieval (GIR) have addressed issues such as the recognition and disambiguation of place references in text [13], the assignment of geographic scopes to documents [1], or the retrieval of documents taking geographic relevance into account [14, 12].

Leidner presented a variety of approaches for handling place references in text [10]. The problem is usually seen as an extension of the named entity recognition task (NER), as proposed by the natural language processing community. 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 (i.e., assigning them to unambiguous geospatial coordinates or gazetteer identifiers). This

---

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

disambiguation often uses 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) [10].

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 also been addressed. 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, there is still 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 this paper, we experimented with the methods proposed by Amitay et al. [1], Woodruff and Plaunt [21], and Martins and Silva [15], as well as with some simple baseline methods.

As for the existing approaches for retrieving documents according to geographic relevance, they 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 or containment [14]. Larson and Frontiera compared the performance of different methods for computing spatial similarity scores for query-document pairs, using area overlaps [5]. Martins et al. proposed a similarity function that, instead of area overlaps, uses a non-linear normalization of the distance between the document and query scopes [14].

Cai proposed the GeoVSM framework for geographic document retrieval, combining traditional IR heuristics with geographic similarity scores [4]. The author argues that thematic and geographic similarities should be computed independently, and afterwords combined into a single retrieval score. Markowetz et al. proposed an efficient strategy also based on a linear combination of scores [12]. This paper explores similar ideas for the task of retrieving ads that are both thematically and geographically relevant.

3. ASSIGNING GEOGRAPHIC SCOPES
The first stage in the proposed processing pipeline involves assigning geographic scopes to target pages. In turn, this stage involves two separate sub-tasks, namely (i) handling place references in the text, and (ii) assigning geographic scopes to documents based on the disambiguated place references. This section details both sub-tasks.

3.1 Handling Place References in Text
For handling place references in text, we relied on the Placemaker[4] text mining Web service provided by Yahoo!. This 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[5] gazetteer service, and obtain

---

[4]http://developer.yahoo.com/geo/placemaker
[5]http://developer.yahoo.com/geo/geoplanet

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 a word like *Reading* between the location in England or the verb sense of to read. It also covers many colloquial location names (e.g. nyc for *New York City*), as well as interest points (e.g. *Eiffel Tower*) that may appear in the text of the target pages. In a separate publication we measured the performance of the this geotagger for the tasks of recognizing and disambiguating place references [13]. Over the English dataset from the CoNLL-03 experiment on named entity recognition, we obtained F_1 measures of approximately 59% and 57% for recognition and disambiguation, respectively.

3.2 Assigning Geographic Scopes to Pages
We tested seven different approaches for assigning geographic scopes to target pages, namely the method from the Placemaker service, the methods proposed in the context of the Web-a-Where [1], GIPSY [21] and GREASE [15] projects, and three simple baseline methods. In all cases, we used the results of the Yahoo! Placemaker Web service as the source of disambiguated place references.

Yahoo! Placemaker, besides recognizing and disambiguating place references, also assigns scopes to documents. We use this Web service as a black box, in order to understand what is the current performance of commercial applications.

The Web-a-Where technique leverages on *part-of* relations among the recognized place references, provided by a hierarchical gazetteer [1]. 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 geographic 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.

The GIPSY algorithm uses the bounding boxes that correspond to the place references in the text, after they have been disambiguated [21]. The geographic scope is 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. In 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, in order to build a skyline of bounding boxes. Finally, the bounding boxes are 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 from the disambiguation task.

The approach from the GREASE project is based on graph-ranking [15]. The idea is 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 are weighted according to the occurrence frequency of place references in a document, and edges are weighted according to the relative importance of the different types of relationships. A graph-ranking algorithm, namely *PageRank*, is applied to this graph and 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.

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

1. 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 as the scope.

2. 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.

3. Only the place references that are somewhat interrelated should be considered, this way filtering the errors made while recognizing and disambiguating place references, as well as filtering 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 are remaining, we choose the one closest to the centroid. This baseline is inspired on a technique proposed by Smith and Crane for place reference disambiguation [18].

4. CLASSIFYING TARGET PAGES

The second stage of the proposed processing pipeline involves classifying target pages according to their locational relevance, i.e. classifying them as either local or global, so that locally relevant ads are 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 events in a specific city could be regarded as local, likely to be more interesting to a regional audience.

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 of specific features better suited to reflect the locational characteristics. A work by Gravano et al. [6] already described similar ideas, although for classifying search engine queries instead of documents.

Our features attempt to capture two different but complementary aspects of locational relevance, namely (i) 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 (ii) 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 (e.g., seperate counts for cities or states), and (iii) high level locative features, e.g., the spatial area of the geographic scopes obtained through different algorithms. Due to space restrictions we omit the complete description of the considered features here, but the reader can refer to a separate publication [2].

We chose to use a classifier based on Support Vector Machines (SVMs) since SVMs represent the state-of-the-art text classification technology [17]. Moreover, they offer the possibility to assign a value in the interval [0,1] that estimates the likeliness of the document being either in the local or global class. In particular, we used a gaussian-based SVM with parameters C and γ optimized through the grid-search functionality offered by Weka [20]. The final classifier scores correspond to probabilities for the documents being either local or not, according to the normalization procedure described by Lin et al. [11].

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 thematic and geographic relevance.

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, advertisers

also associate their ads with the intended geographic scope. The query triggering the search is derived from the context of the target page (the text and the geographic scope) where the ads are to be displayed. Since users are unlikely to click on irrelevant ads, the retrieval 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 [4]. As shown in Equation 1, GeoVSM independently computes a geographic similarity gs and a thematic similarity ts, later combining them through some function f.

$$Rel(doc, ad) = f(ts_{\{doc,ad\}}, gs_{\{doc,ad\}}) \qquad (1)$$

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 2, is an adequate function f. In the context of multimedia information retrieval, Wu et al. [22] 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 usage of static weights, the locational relevance score provides a weighting scheme that dynamically adapts itself to each document.

$$f(ts, gs) = (1 - w) \cdot ts + w \cdot gs \qquad (2)$$

We experimented with two different approaches for measuring the thematic relevance ts, namely the similarity between document key terms (extracted with Yahoo's service) 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[6]. However, since the PostgreSQL full-text engine does not provide a thematic similarity value v in the range [0,1], we applied the min-max normalization procedure shown in Equation 3.

$$ts = \frac{ts' - min_{ts'}}{max_{ts'} - min_{ts'}} \qquad (3)$$

For measuring geographic relevance gs we also experimented with two different strategies, namely the normalized distance metric by Martins et al. [14] and the relative area of overlap metric proposed by Greg Janée[7].

The method by Martins et al. uses a double sigmoid function with the center corresponding to the diagonal distance of the rectangular region for the query scope (i.e., the geographic scope of the target page). The similarity is maximum when the distance is zero, and smoothly decays to zero as the distance increases. The method is presented in Equation 4, where d refers to the diagonal distance of the rectangular region corresponding to the geographic scope of the ad S_{ad} and $D = centroidDistance(S_{page}, S_{ad}) - d$.

$$sim(S_{page}, S_{ad}) =$$
$$\begin{cases} 1 \text{ if } S_{page} \text{ is contained in } S_{ad} \\ 1 - \frac{1 + sign(D) \times (1 - e^{-(\frac{D}{d \times 0.5})^2})}{2} \text{ otherwise} \end{cases} \qquad (4)$$

---

[6] http://www.postgresql.org
[7] http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html

Frontiera et al. [5] showed that the method proposed by Janée performs almost as well as a highly-tuned method based on logistic regression. In Janée's approach, the similarity between two regions S_{page} and S_{ad} is given as follows:

$$sim(S_{page}, S_{ad}) = \frac{area(S_{page} \cap S_{ad})}{area(S_{page} \cup S_{ad})} \qquad (5)$$

In our implementations of the geographic similarity functions, PostGIS[8] was used to compute distances and overlaps.

6. EXPERIMENTAL EVALUATION

In this section, we describe the details of our empirical evaluation. This includes our experimental design for evaluating the effectiveness of the proposed approaches, as well as the obtained results in the different experiments.

6.1 Assigning Scopes to Target Pages

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

We ran all seven 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 the ODP. This test jointly evaluates place reference resolution and scope assignment, since any errors made by the geotagger influence the scope assignment. The overlap was measured using the scheme proposed by Janée, which we described in Section 5. We also measured the accuracy for both exact matches and approximate matches. Table 1 summarizes our results.

The Web-a-Where and GraphRank algorithms obtained the best overall performances, with errors equally distributed across countries, states, and cities. Both approaches were particularly good in pages with country scopes, which was already expected due to their generalization behaviour (i.e., propagate scores towards encompassing regions). 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 using the covering area produces the worst results on most metrics. Removing the outliers substantially improves the results, but this is also a very weak baseline in terms of overall performance, specially when dealing with narrow scopes. The baseline that simply assigns the scope as the most frequent location proved to be a very competitive approach, regularly outperforming other methods on pages with scopes corresponding to states or cities.

---

[8] http://www.postgis.org

Table 1: Comparative evaluation of geographic scope assignment methods.

Algorithm	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

The Placemaker scope assignment method has good results on pages with country scopes, but modest to weak results on pages with state and city scopes. This suggests that this service has a tendency to overgeneralise from place references towards their encompassing regions.

One of the assumptions behind this work is that the more specific the geographic scope of a page is, the more interesting the page will be for advertisement. 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 for the envisioned application in contextual advertising.

6.2 Locational Relevance Classification

In order to evaluate the locational relevance classifier, we used a dataset consisting of 8,000 Web pages crawled from the ODP. The dataset consists of 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, had at least one place reference, and were randomly selected from the various ODP thematic categories.

The experiments considered four classifier configurations, i.e., (i) using only textual features, (ii) using only simple locative features, (iii) using only high level locative features, and (iv) using a combination of textual features and the best locative features. Table 2 overviews the results. All values where 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, probably due to 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*.

Besides the binary classification approach described in Section 4 (i.e., local vs. global), we also experimented with a classifier that considered four classes: *city-local*, *state-local*, *country-local*, and *non-local*. The same set of ODP pages was used in this experiment and the considered feature set combined the textual with 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

We experimented with different combinations of thematic and geographic similarity measures. Specifically, we considered the following three groups of experimental variables:

- Thematic similarity, based on the full text of the pages (T1) or using only the extarcted keywords (T2);

- Geographic similarity, based on the normalized distance (G1) or the relative area of overlap (G2);

- Combined weighting schemes, using a baseline which assigns a weight of 0.5 to each similarity (W1) or using the locational relevance score (W2).

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

	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**

Tests were made using two collections of target pages, namely local and global sets of pages from the ODP. The local dataset had 20 pages taken from regional sub-sections, at city or state level, 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`. These pages had advertisements placed on them, and had at least one geographic reference. We also developed an advertisements collection, using the title, description, and geographic scope assigned by the ODP editors for 133.200 pages associated with `Business` categories. The ad's keywords were obtained by retrieving the most important words from the corresponding 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 ads to be relevant to the pages on the local dataset when they belong to a related thematic category and have 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 correspondence between thematic categories on ads and target pages 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 considered experimental combinations, measuring the precision at different cut-off points, as well as the mean average precision (MAP) 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 although these target pages have a low locational relevance, it is enough to retrieve ads with a high thematic relevance. In this case, using geographic similarity 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

Table 3: Comparison of retrieval performance.

	Experiment	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

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 ads relevant 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 main findings of this research are as follows:

- The method from 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 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 a normalized geospatial distance, where individual scores are weighted according to the locational relevance of the document, achieves the best results for selecting ads to place in 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 the text-only approaches perform better. 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. Currently ongoing research aims as measuring how the quality of place reference disambiguation influences the results of the scope assignment algorithms and locational relevance classifiers.

Our experimental results also showed significant differences in the geographic scope algorithms. It would therefore be interesting to see if a combination of the best algorithms could lead to better results. If we simply took the best algorithm for each document in the test collection, it would be possible to obtain an accuracy of 70%, showing that this is indeed a promising alternative

Putting a threshold on the locational relevance, in order to ensure that the geographic similarity is only considered on geographically relevant Web pages, is a possible alternative for improving the retrieval results over the global pages. Another possible alternative for improving retrieval scores concerns with the usage of machine learning approaches for optimizing the ranking formula that combines the multiple sources of evidence (i.e., thematic and geographic).

8. ACKNOWLEDGMENTS

9. REFERENCES

[1] 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.

[2] I. Anastácio, B. Martins, P. Calado (2009) Classifying documents according to locational relevance. In Proceedings of EPIA '09: 14th Portuguese Conference on Artificial Intelligence.

[3] 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.

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

[5] 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).

[6] 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.

[7] 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.

[8] 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).

[9] 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.

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

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

[12] A. Markowetz, Y.Y. Chen, T. Suel, X. Long, and B. Seeger (2005) Design and implementation of a geographic search engine. In Proceedings of the 8th International Workshop on the Web and Databases.

[13] B. Martins, I. Anastácio, P. Calado (2010) A Machine Learning Approach for Resolving Place References in Text. In Proceedings of the 13th AGILE International Conference on Geographic Information Science.

[14] 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.

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

[16] 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.

[17] F. Sebastiani (2002) Machine learning in automated text categorization. ACM Computing Surveys, 34(1).

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

[19] 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.

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

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

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

[23] W. 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.