

Learning to Rank for Geographic Information Retrieval

Bruno Martins and Pável Calado
{bruno.g.martins , pavel.calado}@ist.utl.pt
Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva
2744-016 Porto Salvo, Portugal

ABSTRACT

The task of Learning to Rank is currently getting increasing attention, providing a sound methodology for combining different sources of evidence. The goal is to design and apply machine learning methods to automatically learn a function from training data that can sort documents according to their relevance. Geographic information retrieval has also emerged as an active and growing research area, addressing the retrieval of textual documents according to geographic criteria of relevance. In this paper, we explore the usage of a learning to rank approach for geographic information retrieval, leveraging on the datasets made available in the context of the previous GeoCLEF evaluation campaigns. The idea is to combine different metrics of textual and geographic similarity into a single ranking function, through the use of the SVM^{map} framework. Experimental results show that the proposed approach can outperform baselines based on heuristic combinations of features.

Categories and Subject Descriptors

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

General Terms

Algorithms, experimentation

Keywords

Learning to Rank, Geographic Information Retrieval

1. INTRODUCTION

The task of Learning to Rank (L2R) is currently getting increasing attention both in Information Retrieval (IR) and Machine Learning (ML). The goal is to design and apply ML methods to automatically learn a function from training data capable of sorting documents according to their relevance to user queries. L2R approaches can naturally deal many different sources of evidence, without requiring human effort to decide explicitly how best to combine them.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GIR'10, 18-19th Feb. 2010, Zurich, Switzerland

Copyright © 2010 ACM ISBN 978-1-60558-826-1/10/02... \$10.00

Geographic information retrieval (GIR) has also emerged as an active and growing research area, addressing the retrieval of textual documents according to geographic criteria of relevance. Geographic search requires the combination of geospatial and thematic based relevance into one result. However, it is still an open research question how to best combine these thematic (e.g., term similarity) and geographic (e.g., proximity or area of overlap) sources of evidence. Applying L2R approaches to geographic information retrieval presents itself as a promising approach.

In this paper, we explore the usage of a learning to rank approach for GIR, leveraging on the datasets made available in the context of the previous GeoCLEF evaluation campaigns [13]. The idea is to combine different metrics of textual and geographic similarity into a single ranking function, through the use of the SVM^{map} L2R framework, an approach based on Support Vector Machines for optimizing the Mean Average Precision metric. Our experimental results show that L2R can outperform baseline approaches based on heuristic combinations of features.

The rest of this paper is organized as follows: Section 2 presents related research in both learning to rank and geographic information retrieval. Section 3 presents the proposed approach for learning to rank in geographic IR, detailing the proposed set of features and the application of the SVM^{map} framework. Section 4 presents the experimental evaluation. Finally, Section 5 summarizes the main conclusions and points directions for future work.

2. RELATED WORK

Previous research in geographic information retrieval has addressed problems such as the recognition and disambiguation of place references given over text [10, 15], the assignment of documents to encompassing geographic scopes [1], or the retrieval of documents considering geographic relevance [16, 21, 6]. The first two problems are normally seen as necessary pre-processing tasks, so that later one can use ranking formulas that leverage on the similarity between the geographic scopes of documents and of user queries.

Geographic text mining technology is nowadays mature and commercial services offering these functionalities are starting to appear. For instance, Yahoo! Placemaker¹ is a Web service for geotagging text, additionally using rules over the set of place references to determine the geographic scope of

¹<http://developer.yahoo.com/geo/placemaker/>

the document. The service delivers the bounding boxes and centroid coordinates for the named places and scopes. Each of the recognized place references is also assigned to a unique Where-on-Earth Identifier and, though this identifier, one can access hierarchical gazetteer information (i.e., parent regions in an administrative hierarchy for the region of the given identifier) using a separate Yahoo! Web service.

On what concerns document retrieval with basis on geographic relevance, different methods have been tested in the context of the GeoCLEF evaluation campaign [14, 13]. Many research and evaluation issues surrounding geographic mono- and bilingual search have been addressed in GeoCLEF. The most common approaches are based on heuristic combinations of the standard IR metrics used in text retrieval (e.g., TF/IDF), with similarity metrics for geographic scopes based on distance and/or containment [16].

Frontiera et al. compared different geographic similarity methods based on region overlaps [3]. Henrich and Lüdecke noticed that overlaps only provide a strict notion of similarity (e.g., two regions that are near each other but not overlapping are just as dissimilar as two regions that are hundreds of miles apart) and, for GIR, similarity metrics should also account with other perspectives besides overlap [5]. The same authors also noted that the exact similarity between the regions themselves is not the main focus in GIR, since the quality of the results merely depends of the ranking.

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 [16]. The normalization is done through a double sigmoid function with the center corresponding to the diagonal distance of the rectangular region corresponding to the query scope. The similarity is maximum when the distance is zero, and smoothly decays to zero as the distance increases.

Yu and Cai proposed a dynamic document ranking scheme to combine the thematic and geographic relevance measures on a per-query basis [21]. The authors used query specificity (i.e., the geographic area covered by the query) to determine the relative weights of different sources of ranking evidence for each query (i.e., the weight of the geographic relevance measure is inversely proportional to the area of the query). In this paper, instead of relying on heuristics, we propose to use machine learning for finding the optimal combination of thematic and geographic similarity measures.

Learning to Rank (L2R) for information retrieval concerns with using machine learning methods, together with training data consisting of queries and relevance judgements, in order to construct ranking functions [11]. The main advantage of this approach is that many different sources of evidence can be naturally integrated into the ranking model. At the same time, system designers need not to decide explicitly how best to combine the different forms of evidence, instead relying on past data (e.g., query logs or explicitly provided relevance judgements) for fine tuning the retrieval process.

Like in any other machine learning application, learning to rank research is heavily affected by the availability of train-

ing data. The LETOR benchmark datasets [12] were released in the SIGIR 2007 workshop on Learning to Rank for Information Retrieval. Since then, they have been widely used in experiments, greatly speeding up the research in the area. Many different approaches have been proposed and evaluated using the LETOR dataset. The reader should refer to [11] for a survey on the area. In this paper, we show how the datasets made available in the context of the GeoCLEF evaluation campaigns can easily be used to build a benchmark collection that, similarly to LETOR, can speed up the research on ranking methods specific for geographic information retrieval.

One of the previously proposed L2R approaches is SVM^{map} , a Support Vector Machine (SVM) algorithm for predicting document rankings [22]. It performs supervised learning using binary labeled training examples (i.e., document labels in the training data are either relevant and non-relevant) with the goal of directly optimizing the Mean Average Precision (MAP), an important benchmark in the Information Retrieval community. SVM^{map} is generally superior to or competitive with other learning to rank methods. Section 3.2 presents a detailed description of this method, and Section 4 details the MAP evaluation metric.

3. L2R USING THE GEOCLEF DATASETS

To conduct learning to rank experiments, one first needs a document collection for which there are a set of known information needs (i.e., *topics*) together with *relevance judgements*. This paper proposes to use the English document collection from the previous GeoCLEF campaigns for testing GIR learning to rank approaches.

In the years of 2005, 2006, 2007 and 2008, there was a special track for geographic information retrieval, named GeoCLEF, at the Cross Language Evaluation Forum (CLEF) [14, 13]. The goal of this track was to investigate the retrieval behavior when the topics involved geospatial constraints. Specific sub-tracks of GeoCLEF involved different languages in either mono-lingual or multi-lingual scenarios.

In the monolingual English case, the track used a newspaper collection composed from documents taken from the American newspaper Los Angeles Times from the year of 1994, and the English newspaper Glasgow Herald from the year of 1995. These collections had been previously used in CLEF ad-hoc evaluations, which addressed a classic document retrieval scenario. There are in total 169,477 documents in the GeoCLEF English collection (i.e., in the Glasgow Herald plus Los Angeles Times collections). Documents have a headline field, separated from the main body of text.

The topics used in GeoCLEF were meant to express a natural information need which a user of the collection might have. The organizers aimed at creating a geographically challenging topic set, in which explicit geographic knowledge should be necessary to successfully retrieve relevant documents. The idea was that keyword-based approaches alone should not be enough.

From the four editions of GeoCLEF there are a total of 100 topics (i.e., there are 25 topics from each edition). Each topic contains a title (i.e., a short description of the infor-

mation need, similar to what a user would present to a geographic IR system), a description (i.e., a longer description of the information need) and a narrative (i.e., a longer explanation of what constitutes relevant and irrelevant documents for the topic). Some of the topics had place references clearly mentioned in the topic title and description (e.g., *car bombings in Madrid*), while others expressed more challenging geographic constraints (e.g., *cities near active volcanoes*). To some extent, the difficulty associated with the topics increased in each of the GeoCLEF editions, with several issues being explicitly included. The different types of topics are summarized next:

- Topics with a non-geographic subject restricted to a place (e.g., *music festivals in Germany*). This was the only kind of topic used in GeoCLEF 2005.
- Topics having a geographic subject with non-geographic restrictions (e.g., *rivers with vineyards*). This type was added since GeoCLEF 2006.
- Topics with a geographic subject and restricted to a place (e.g., *cities in Germany*).
- Topics with a non-geographic subject associated to a place (e.g., *independence of Quebec*).
- Topics with a non-geographic subject that is a complex function of place (e.g., *European football cup matches*).
- Topics with a vague and/or imprecise geographic region (e.g., *Sub-Saharan Africa*).
- Topics with geographical relations among places (e.g., *Oil and gas extraction found between the UK and the Continent*).
- Topics with geographical relations among events (e.g., *F1 circuits where Ayrton Senna competed in 1994*).
- Topics with relations between events which require their precise localization (e.g., *Casualties in fights in Nagorno-Karabakh*).

Documents in the GeoCLEF collections have binary relevance judgements for each of the topics, collected by the organizers through a pooling approach, where the set of documents retrieved by the participating systems were manually judged as either relevant or not. The documents marked as relevant match both the thematic and geographic constraints expressed in the topics.

The original document collections were not geographically tagged and contained no semantic location-specific information. For our experiments, we processed the English document collection with the Yahoo! Placemaker Web service, in order to extract place references from the text and assign each document to a corresponding geographic scope. We used the service option that weights the title text (in the case of GeoCLEF English documents, the headline) as more important. The set of topics was also processed with the Yahoo! Placemaker service, in order to extract place references from the title, the description or the narrative, and also assign each topic to a corresponding geographic scope.

There are two flavors of document scopes in Placemaker, namely the *geographic scopes* and the *administrative scope*. The geographic scope is the place that best describes the document. The administrative scope is the place that best describes the document and has an administrative type (i.e., Continent, Country, State, County, Local Administrative Area or Town). In our experiments, we used the geographic scopes returned by Placemaker, since some of the topics contained non-administrative geographical references. For each scope, Placemaker returns the corresponding bounding box and we use these spatial representations to compute several geospatial similarity functions.

The set of English documents, augmented with the semantic location-specific information provided by Placemaker, was used to generate a feature vector representation for each topic-document pair. The following sub-section details the set of features that were considered for learning.

3.1 The Considered Features

The considered set of features can be divided into three groups, namely textual, geographic and averaged features. The textual features are similar to those used in standard text retrieval systems and also in previous learning to rank experiments (e.g., TF-IDF scores). We considered two document streams, namely (i) using just the headlines or (ii) using the headlines plus the main body of the documents. Separate textual features were computed for each of the streams. The geographic features correspond to similarity metrics proposed in the GIR literature [3, 16, 9], computed between topic and document scopes. Finally, the averaged features correspond to heuristic combinations of the textual and geographic features.

For each topic-document pair, the considered set of textual features is as follows:

- Two separate term frequency (TF) features: one for the headline of the document and another for the concatenation of the headline with the document body. The value of these features is the sum of the term frequencies for each individual term in the topic title.

Eq. (1) shows the term frequency formula used, where $n_{i,j}$ is the number of occurrences of term i in document D_j and $|D_j|$ is the number of terms in document D_j .

$$tf_{i,j} = \frac{n_{i,j}}{|D_j|} \quad (1)$$

- Similarly, two inverse document frequency (IDF) features: one for the headline of the document and another for the concatenation of the headline with the document body.

The value of these features is the sum of the IDF for each individual term in the topic title, where each document frequency corresponds to the number of documents containing the term. The formula is shown in Eq. (2), where n_i is the number of documents containing term i and N is the total number of documents.

$$idf_i = \sum_i \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \quad (2)$$

Since the IDF is document independent, all the documents under a topic have the same IDF values.

- Document length features for the headline and for the concatenation of the headline with the document body. The value of these features is the stream size in terms of the number of words.
- Term frequency times inverse document frequency (TF-IDF) features between documents and topics, considering the headline and considering the concatenation of the headline with the document body.

The value for these features corresponds to the cosine similarity between the term vector for the topic title and the term vector for the stream, each using TF-IDF weights, as shown in Eq. (3):

$$\cos(D_j, T_k) = \frac{\sum_i w_{i,j} \times w_{i,k}}{\sqrt{\sum_i w_{i,j}^2} \times \sqrt{\sum_i w_{i,k}^2}} \quad (3)$$

where $w_{a,b} = tf_{a,b} \times idf_a$ is the TF-IDF weight of term a in document (or topic) b .

- BM25 features between documents and topics, for the headline and for the concatenation of the headline with the document content. The value for these features is the sum of the BM25 score for each individual term in the topic title, given by Eq. (4):

$$\text{BM25}(D_j, T_k) = \sum_{i \in T_k} idf_i \times \frac{(k_1 + 1) \times tf_{i,j}}{tf_{i,j} + k_1 \times (1 - b + b \times \frac{|D_j|}{\mathcal{A}})} \quad (4)$$

where $|D_j|$ is the the number of terms in document D_j and \mathcal{A} is the average length of the documents in the collection. In the computation of BM25, we used the standard parameters of $k_1 = 2.5$ and $b = 0.8$.

Textual features like BM25 are by themselves complete ranking models, involving parameters tuned through heuristics. Other textual features, like term frequency or document length, are simple statistical measures. Using both types of features allows our model to take advantage of relevance estimates that are known to be effective, together with simple features to discover new relevance estimates. A similar approach was taken in the LETOR datasets [12].

The geographic features consist of similarity metrics computed between the geographic scopes of the topic and the document. The considered set of features is as follows:

- The area of the geographic scope of the topic, or zero if the topic does not have a geographic scope.
- The area of the geographic scope of the document, or zero if the topic does not have a geographic scope.
- The hierarchical distance measure originally proposed by Jones et al. [9], computed over the taxonomy given by the ancestors of the geographic scopes of the topic and the document. If either the document or the topic do not have geographic scopes, this feature is assigned a value of minus one.

The formula for the hierarchical distance is shown in Eq. (5), where the function $TaxonomyLevel()$ returns the distance from a given node in the geographic taxonomy to the root node (i.e., in a world gazetteer, the level for */World/Europe/Portugal* is 2) and the function $Ancestors()$ returns, for a given node, its set of ancestor nodes in the taxonomy.

$$\begin{aligned} \text{sim}(S_d, S_t) = & \frac{1}{TaxonomyLevel(S_d)} + \frac{1}{TaxonomyLevel(S_t)} + \\ & \sum_{x \in Ancestors(S_d) - Ancestors(S_t)} \frac{1}{TaxonomyLevel(x)} + \\ & \sum_{y \in Ancestors(S_t) - Ancestors(S_d)} \frac{1}{TaxonomyLevel(y)} \end{aligned} \quad (5)$$

In our experiments, the taxonomical information is provided by the Yahoo! GeoPlanet Web service.

- The area of overlap between the geographic scope of the topic and the geographic scope of the document. If either the document or the topic do not have scopes, than this feature is assigned a value of zero.
- The distance between the centroid point for the geographic scope of the document and the centroid point for the geographic scope of the topic. If either the document or the topic do not have geographic scopes, this feature is assigned a value of minus one.
- The normalized distance between the centroid point for the geographic scope of the document and the geographic scope of the topic, computed through the procedure originally proposed by Martins et al. [16]. If either the document or the topic do not have geographic scopes, this feature is assigned a value of minus one.

The formula for the normalized distance metric is shown in Eq. (6), where S_d is the scope of the document, S_t is the scope of the topic, d is the diagonal distance of the rectangular area corresponding to S_t and D is the distance between the centroids of S_d and S_t .

$$\text{sim}(S_d, S_t) = \begin{cases} 1 & \text{if } S_d \subseteq S_t \\ 1 - \frac{1 + \text{sign}(D-d) \times (1 - e^{-\left(\frac{D-d}{2}\right)^2})}{2} & \text{otherwise} \end{cases} \quad (6)$$

- The degree of overlap between the area for the scope of the document and the area for the scope of the topic, given by the formula originally proposed by Hill [6]:

$$\text{sim}(S_d, S_t) = \frac{2 \times \text{area}(S_d \cap S_t)}{\text{area}(S_d) + \text{area}(S_t)} \quad (7)$$

- The degree of overlap between the area for the scope of the document and the area for the scope of the topic, given by the formula originally proposed by Walker et al. [19], shown in Eq. (8):

$$\text{sim}(S_d, S_t) = \min \left(\frac{\text{area}(S_d \cap S_t)}{\text{area}(S_d)}, \frac{\text{area}(S_d \cap S_t)}{\text{area}(S_t)} \right) \quad (8)$$

- The degree of overlap between the area for the scope of the document and the area for the scope of the topic, given by the formula originally proposed by Beard and Sharma [2], shown in Eq. (9):

$$sim(S_d, S_t) = \begin{cases} \frac{\text{area}(S_d)}{\text{area}(S_t)} & \text{if } S_d \subset S_t \\ \frac{\text{area}(S_t)}{\text{area}(S_d)} & \text{if } S_d \supset S_t \\ \frac{\text{area}(S_d \cap S_t)}{\text{area}(S_t)} & \text{if } S_d \cap S_t \neq \emptyset \\ \frac{\text{area}(S_d \cap S_t)}{\text{area}(S_d)} & \text{if } S_d \cap S_t \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

- The degree of overlap between the area for scope of the document and the the area for the scope of the topic, given by the formula originally proposed by Janée [7]:

$$sim(S_d, S_t) = \frac{\text{area}(S_d \cap S_t)}{\text{area}(S_d \cup S_t)} \quad (10)$$

- The degree of overlap between the the area for scope of the document and the area for the scope of the topic, given by a formula inspired by the method proposed by Frontiera et. al [3], shown in Eq. (11):

$$sim(S_d, S_t) = \frac{\frac{\text{area}(S_d \cap S_t)}{\text{area}(S_t)} + \frac{\text{area}(S_d \cap S_t)}{\text{area}(S_d)}}{2} \quad (11)$$

- The Hausdorff distance [20] between the scope of the document and the scope of the topic, given by the formula shown in Eq. (12):

$$sim(S_d, S_t) = \max\left\{ \begin{aligned} & \max\{\forall p_1 \in S_d | \min\{\forall p_2 \in S_t | \text{dist}(p_1, p_2)\}\}, \\ & \max\{\forall p_1 \in S_t | \min\{\forall p_2 \in S_d | \text{dist}(p_1, p_2)\}\} \end{aligned} \right\} \quad (12)$$

If the document or the topic do not have a geographic scope, this feature is assigned a value of minus one.

The averaged features combine geographic and thematic aspects into individual estimates. Two of the considered features correspond to mean values computer over textual and geographic features, and we also considered the combination approach originally proposed by Yu and Cai [21].

- The mean value between the normalized distance proposed by Martins et al. and the BM25 score computed over the concatenation of the headline with the document content.
- The mean value between the overlap measure proposed by Janée and the BM25 score computed over the concatenation of the headline with the document content.
- A combination of BM25 and the similarity metric by Janée, following the scheme originally proposed by Yu and Cai [21] where the geographic similarity score has a weight that is proportional to the area of the geographic scope of the topic.

The geospatial operations involved in the computation of the geographic features were all implemented through the use of the Java Topology Suite (JTS) API². For the textual similarity metrics, we provided our own implementations.

²<http://tsusiatsoftware.net/jts/main.html>

3.2 The SVM^{map} L2R Framework

In our experiments, we used the SVM^{map} [22, 8] implementation³ by Yisong Yue and Thomas Finley, which relies on the SVM^{struct} framework⁴ for learning structured (i.e., where the output labels are multivariate) prediction tasks [17]. SVM^{map} is essentially a listwise learning to rank approach that incorporates average precision into the optimization constraints of SVM^{struct} .

Support vector machines (SVMs) are set of related supervised learning methods commonly used for classification and regression. Structural SVMs generalize traditional SVMs to problems where the label space is structured (i.e., we no longer have a single class label y_i for each training instance, but instead have a set of labels $y_{i,j}$ with dependencies among themselves) and of possibly infinite size. Unlike in the case of multi-class classification problems, it is not easy to break down the structured label spaces into a set of binary classification problems (e.g., mapping each possible combination of labels $y_{i,j}$ into a class results in an exponential number of classes). The training of a structural SVM maximizes the margin between the sets of correct and incorrect labels, using a loss function that corresponds to an upper bound on the number of label errors (i.e., the number of times at least one $y_{i,j}$ in the set of labels was wrong). The number of labels that have to be considered is very large, but SVM^{struct} tackles this problem by performing optimization only on a working set of labels which, at each step, is extended with the most violated one. The resulting algorithm works in polynomial time.

In SVM^{map} , the loss function is defined in terms of the Mean Average Precision (MAP) derived from the predicted list and the ground truth list. The number of labels in the problem is exponential in the number of rankings, but SVM^{struct} efficiently handles the optimization task.

For detailed information about the theory behind SVM^{map} , the reader should refer to [22, 8, 17].

4. EXPERIMENTAL EVALUATION

The considered learning to rank approach (i.e., SVM^{map}) optimizes results in terms of the Mean Average Precision (MAP). Average Precision (AP) uses two classes in the relevance judgements (i.e. relevant versus non-relevant) and, given a ranked list of documents retrieved for a topic, computes a measure of the quality of the results using the formula that is presented next:

$$MAP(D, T) = \frac{1}{|D_T +|} \sum_{j=1}^{|D_T|} \frac{1}{j} \sum_{k=1}^j isRelevant(k) \quad (13)$$

In the formula, $|D_T +|$ denotes the number of relevant documents with respect to the topic and $|D_T|$ denotes the number of retrieved documents. The MAP is defined as the mean of the average precision over a set of queries.

³<http://projects.yisongyue.com/svmmmap/>

⁴http://www.cs.cornell.edu/People/tj/svm_light/svm_struct.html

Previous GeoCLEF results have shown that the considered set of topics is indeed challenging and the performance of the GIR systems, in terms of metrics such as the MAP, lags behind that of systems designed for typical ad-hoc retrieval without geographical parameters [14, 13]. The specific objectives of our experiments were therefore to see if (i) a combination of thematic and geographic similarity metrics could outperform a traditional textual-based IR baseline like TF-IDF or BM25, and to see if (ii) the combination of multiple sources of thematic and geographic evidence, made through the SVM^{map} learning to rank approach, could outperform heuristic combinations of features.

Taking the topics and relevance judgements from the four past editions of GeoCLEF, we started by splitting the collection into four different subsets. Each of these subsets considered the topics from one edition for validation, and the topics from the remaining editions for training. With this setting, one can also compare the obtained results against the best performing systems in each edition of GeoCLEF. Table 4 presents a statistical characterization of the four different validation subsets, also showing the best results reported in the GeoCLEF evaluations, in terms of the Mean Average Precision. The table also distinguishes topics for which the Placemaker Web service assigned scopes corresponding to small (i.e., less than $100Km^2$), medium (i.e., between $100Km^2$ and $1000Km^2$) or large (i.e., more than $1000Km^2$) geographic areas. It is interesting to note that the participating systems achieved better results in GeoCLEF 2005, reflecting the fact that the topics from that edition were somewhat simpler.

Table 1: Data from past GeoCLEF editions.

	2005	2006	2007	2008
Number of English Topics	25	25	25	25
Avg. Terms per Topic Title	6.64	5.76	6.08	5.48
Topics with correct scope	21	13	12	16
Topics with incorrect scope	5	6	8	5
Topics with no scope	0	1	3	3
Topics with small scope	9	14	11	9
Topics with medium scope	3	4	4	3
Topics with large scope	13	6	7	10
Relevant Topic-Doc. Pairs	1,028	378	650	747
Judged Topic-Doc. Pairs	14,546	17,964	15,637	14,528
Considered Topic-Doc. Pairs	18,000	18,000	18,000	18,000
Best Mean Avg. Precision	0.3936	0.3034	0.2850	0.3037

Notice that for some of the topics, the Placemaker Web service was not able to assign a corresponding geographic scope (e.g., for the topic “Malaria in the Tropics,” the Placemaker service did not recognize the reference to the region of the Tropics). In some other topics, the assigned scope did not correspond to the correct one (e.g., for the topic “Diamond trade in Angola and South Africa,” the Placemaker service assigned a geographic scope corresponding to the African continent). In the cases where no scope was assigned, one cannot expect improvements from using geographic similarity metrics, although the ranking approach should nonetheless be able to rely on the thematic similarity metrics for retrieving relevant documents. Errors in the scope assignment process will also propagate into the ranking of documents. It should nonetheless be noted that the experiments reported here correspond to a realistic scenario of having a GIR system interpreting user queries.

Instead of using the entire document collection (i.e., assuming that the non-judged documents are not relevant and then building document-topic pairs for all the documents versus all the topics), we mostly used the documents for which we had relevance judgements. The rationale for this is that learning from the entire collection can become computationally unfeasible (i.e., a collection of 169,477 documents and a total of 100 topics results in 16,947,700 document-topic pairs). In the LETOR datasets, the set of documents associated with each topic was also a selection and not the whole original corpus. Since, in GeoCLEF, the relevance judgements were made from pools of documents obtained from a wide variety of participating systems, there should not be any systematic bias towards a particular set of features. Nonetheless, and since a wider range of non-relevant documents is available in the full collection (i.e., judged documents were always retrieved by some system as being relevant, therefore matching some version of the topic), we also considered some random pairs taken from the entire collection, for which no relevance judgements were available. These random documents were assumed to be non relevant. For each of the four GeoCLEF editions, we considered a total of 18,000 document-topic pairs.

Afterwards, and for each of the four sub-sets of the collection, we tested eight different ranking approaches:

1. Using SVM^{map} with only the textual features.
2. Using SVM^{map} with only the geographic and the averaged features.
3. Using SVM^{map} with both the textual, geographical and averaged features.
4. Using TF-IDF without geographic similarity.
5. Using BM-25 without geographic similarity.
6. Using a linear combination of BM25 and the normalized distance proposed by Martins et. al [16], with equal weights for each.
7. Using a linear combination of BM25 and the similarity metric by Janée [7], with equal weights for each.
8. A geographic retrieval approach based on a linear combination of BM25 and the similarity metric by Janée [7]. Following the scheme proposed by Yu and Cai [21], the geographic similarity score has a weight proportional to the area of the geographic scope of the topic.

In all the considered approaches, we did a topic-based normalization on the features, in order to normalize the positive values (i.e., notice that for some of the features, a value of minus one is given if they cannot be computed) into values in the range [0..1]. For a topic t_i , the normalized value of feature $f_k(t_i, d_j)$ is calculated by Eq. (14), where $\max\{f_k(t_i, d_j)\}$ and $\min\{f_k(t_i, d_j)\}$ are the maximum and minimum value of $f_k(t_i, d_j)$ respectively for all the documents in the collection.

$$f_k(t_i, d_j) = \frac{f_k(t_i, d_j) - \min\{f_k(t_i, d_j)\}}{\max\{f_k(t_i, d_j)\} - \min\{f_k(t_i, d_j)\}} \quad (14)$$

In approaches one to three, the SVM^{map} C parameter controlling the tradeoff between regularization and training loss was set to the default value of 0.01.

Table 4 presents the obtained results in terms of standard IR metrics such as MAP and precision at position 10 (P@10), showing that the leaning to rank approach combining both the textual and geographic similarity metrics outperforms all the baseline techniques.

Other interesting conclusions drawn from the results presented in Table 4 are given next:

- Using geographic similarity features alone results in a poor performance. However, using textual features alone provides a very competitive baseline. Both these facts correspond to our expectations, since the geographic scopes do not reflect thematic characteristics of the documents and topics, although the textual contents also reflect geographic constraints (e.g., place names mentioned in the text).
- The BM25 baselines outperform the ones using TF-IDF ranking. This also corresponds to our expectation, as these results are consistent with previous information retrieval experiments.
- The baseline involving the linear combination of textual similarity with the normalized distance metric, using equal weights, outperforms all other baselines involving linear combinations of textual and geographical similarity. This includes the scheme similar to the one proposed by Yu and Cai.

Regarding the baseline approaches that involve a linear combination of textual and geographic similarity scores, it is interesting to note that they only slightly outperform the baseline approaches involving textual similarity alone. This raises the question of whether linear combinations are a good approach to combine both scores. In the past, Vogt and Cottrell argued that linear combinations should only be used when at least one of the metrics involved has a high performance, and when the individual metrics produce sets of relevant documents that have a large overlap and sets of non-relevant documents with a small overlap [18].

Figure 1 plots the average precision in each of the 100 different topics, for the experiment corresponding to the usage of the SVM^{map} framework with all the available features. The horizontal dashed line corresponds to the mean average precision obtained in the same experiment, averaged across the four test document collections corresponding to the different GeoCLEF editions. The results show that for topics not containing geographic scopes, results were generally worse. Results are also worse for topics from the latter GeoCLEF editions, reflecting the fact that the difficulty associated with the topics increased in each edition. Comparing our results with those reported by the participating systems in the previous GeoCLEF editions, one can see that the learning to rank approach provides inferior results to those achieved by the best scoring systems reported for each GeoCLEF edition. However, many of the participating systems reported

the usage of techniques such as stop-word removal, stemming, pseudo-relevance feedback or topic expansion. Many systems also returned a smaller set of documents than that which considered in our learning to rank experiments, or did not return documents for some topics. All the systems that participated on GeoCLEF had nonetheless relied on empirically tuned heuristics. Learning to rank offers a principled approach to combine different sources of evidence, which can lead to improvements in terms of the accuracy of the retrieval results.

5. CONCLUSIONS AND FUTURE WORK

This paper describes the usage of a learning to rank approach for geographic information retrieval, leveraging on the datasets made available in the context of the previous GeoCLEF evaluation campaigns. Different metrics of textual and geographic similarity were combined into a single ranking function, through the use of the SVM^{map} L2R framework. Experimental results show that the proposed L2R method outperforms previous approaches based on heuristic combinations of features.

Despite the promising results, there are many ideas for future work. For instance, the considered features are not all equally effective and an interesting challenge relates to the usage of feature selection mechanisms. To the best of our knowledge, feature selection for ranking is still an unsolved problem which is particularly in need of more research efforts. The considered geographic features are also based on geographic scopes computed for the documents. It would be interesting to experiment with features computed from the individual placenames mentioned in the documents [15].

Our L2R method, although using some features that are dependent of the queries, is also mostly query independent (i.e., a single best-fits-all model is learned and used for the entire set of queries). However, from the GIR point of view, it seems better to employ query-dependent models [21]. It would be interesting to see whether it is possible to partition the space of topics and train different ranking models for different subspaces (i.e., models for different types of topics, for instance one for geographic topics and another for non-geographic ones, or even different models for different types of geographic scopes in the topics). Previous GIR research has already tackled the issue of classifying queries as either global or local [4], and it would be interesting to see if similar approaches could be used in the L2R context.

6. ACKNOWLEDGMENTS

This work was partially supported by the Fundação para a Ciência e a Tecnologia (FCT), through project grant PTDC/EIA/73614/2006 (GREASE-II).

7. REFERENCES

- [1] I. Anastácio, B. Martins, and P. Calado (2009) A Comparison of Different Approaches for Assigning Documents to Geographic Scopes. In Proceedings of InForum 2009, the 1st Portuguese Symposium on Informatics.
- [2] K. Beard and V. Sharma (1997) Multidimensional ranking for data in digital spatial libraries. International Journal of Digital Libraries, 1.
- [3] P. Frontiera, R. Larson, and J. Radke (2008) A comparison of geometric approaches to assessing spatial similarity for GIR. International Journal of Geographical Information Science, 22(3).

Table 2: Results obtained with the different ranking approaches.

Approach	Mean Average Precision					Precision at 10				
	2005	2006	2007	2008	Avg.	2005	2006	2007	2008	Avg.
SVM^{map} text	0.2841	0.2052	0.1637	0.1767	0.2074	0.3880	0.1822	0.2240	0.2360	0.2575
SVM^{map} geo	0.1215	0.0377	0.0767	0.0827	0.0790	0.1120	0.0040	0.0760	0.0640	0.0640
SVM^{map} text+geo	0.3116	0.2131	0.1859	0.2024	0.2282	0.3960	0.1840	0.2160	0.2400	0.2590
TF-IDF	0.2293	0.1500	0.1541	0.1595	0.1732	0.2400	0.1160	0.1320	0.1760	0.1660
BM25	0.2834	0.2052	0.1629	0.1761	0.2069	0.3800	0.1760	0.2160	0.2320	0.2510
(BM25 + distance) / 2	0.2849	0.2063	0.1637	0.1786	0.2083	0.3960	0.1840	0.2160	0.2360	0.2580
(BM25 + overlap) / 2	0.2838	0.2055	0.1636	0.1762	0.2072	0.3920	0.1800	0.2240	0.2360	0.2580
BM25 + ($w_t \times$ overlap)	0.2856	0.2064	0.1636	0.1773	0.2082	0.3560	0.1800	0.2320	0.2360	0.2510

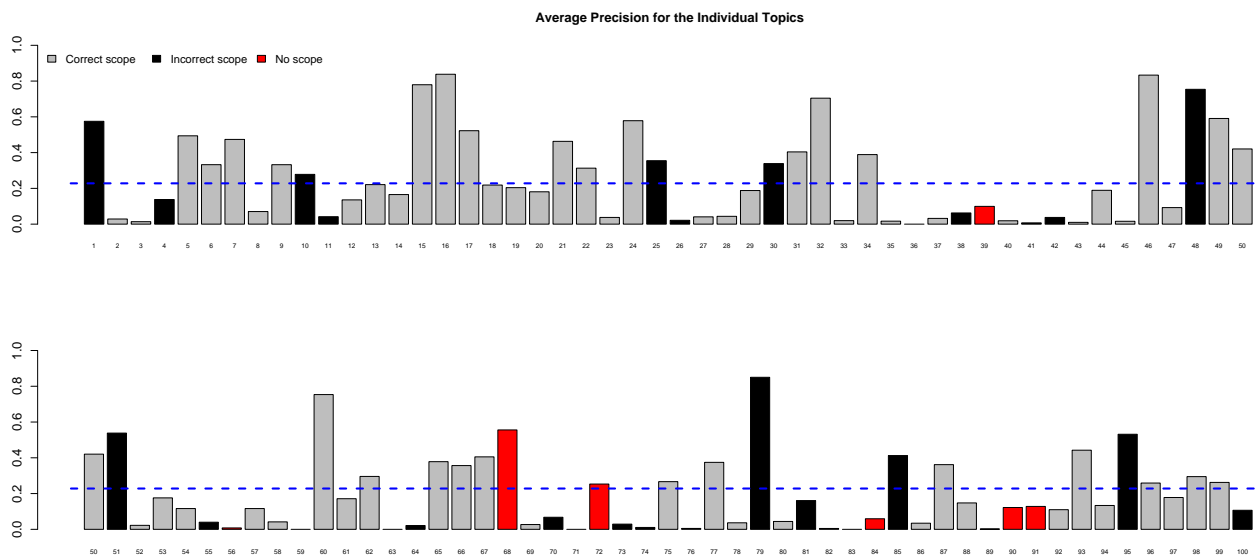


Figure 1: Average precision over the different topics for the SVM^{map} text+geo approach.

- [4] 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.
- [5] A. Henrich, and V. Lüdecke (2009) Measuring Similarity of Geographic Regions for Geographic Information Retrieval. In Proceedings of the 31st European Conference on Information Retrieval.
- [6] L. L. Hill (1990), Access to geographic concepts in online bibliographic files: effectiveness of current practices and the potential of a graphic interface. PhD thesis, University of Pittsburgh.
- [7] G. Janée (2003) Spatial similarity functions, Available online at: www.alexandria.ucsb.edu/~gjane/jane/archives/2003/similarity.html
- [8] T. Joachims (2002) Optimizing search engines using clickthrough data. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- [9] C. B. Jones, H. Alani, and D. Tudhope (2001) Geographical Information Retrieval with Ontologies of Place. In Proceedings of the 5th International Conference on Spatial Information Theory.
- [10] J. L. Leidner (2007) Toponym Resolution in Text. PhD thesis, University of Edinburgh.
- [11] T.-Y. Liu (2009) Learning to Rank for Information Retrieval, Foundations and Trends in Information Retrieval, 3(3).
- [12] T.-Y. Liu, T. Qin, J. Xu, W. Y. Xiong, and H. Li (2007) LETOR: Benchmark dataset for research on learning to rank for information retrieval. In Proceedings of the 1st SIGIR workshop on Learning to Rank for Information Retrieval.
- [13] T. Mandl, P. Carvalho, F. Gey, R. Larson, D. Santos, and C. Womser-Hacker (2008) GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In Working Notes for the Cross Language Evaluation Forum 2008 Workshop
- [14] T. Mandl, F. Gey, G. di Nunzio, N. Ferro, M. Sanderson, D. Santos, and C. Womser-Hacker (2008) An evaluation resource for Geographical Information Retrieval. In Proceedings of the 6th Conference on Language Resources and Evaluation.
- [15] B. Martins, I. Anastácio, and P. Calado (2010) A Machine Learning Approach for Resolving Place References in Text. In Proceedings of AGILE-2010, the 13th AGILE International Conference on Geographical Information Science.
- [16] B. Martins, N. Cardoso, M. S. Chaves, L. Andrade, and M. J. Silva (2007) The University of Lisbon at GeoCLEF 2006. In Proceedings of the 6th Cross-Language Evaluation Forum.
- [17] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun (2005) Large Margin Methods for Structured and Interdependent Output Variables, Journal of Machine Learning Research, 6.
- [18] C. Vogt, and G. Cottrell (1999) Fusion via a linear combination of scores. Information Retrieval, 1(3).
- [19] D. Walker, I. Newman, D. Medyckyj-Scot, and C. Ruggles (1992) A system for identifying datasets for GIS users. International Journal of Geographical Information Systems, 6.
- [20] M. D. Wills (2007) Hausdorff Distance and Convex Sets. Journal of Convex Analysis, 14(1).
- [21] B. Yu and G. Cai (2007) A query-aware document ranking method for geographic information retrieval. In Proceedings of the 4th ACM Workshop on Geographical Information Retrieval.
- [22] Y. Yue, T. Finley, F. Radlinski, and T. Joachims (2007) A Support Vector Method for Optimizing Average Precision, In Proceedings of the 30th ACM SIGIR international Conference on Research and Development in Information Retrieval.