Concept representations in Geographic Information Retrieval as re-ranking strategies

ABSTRACT

Geographic Information Retrieval (GIR) is a specialized Information Retrieval (IR) branch that deals with information related to geographic locations. Traditional IR machines are perfectly able to retrieve the majority of relevant documents for most geographical queries, but they have severe difficulties generating a pertinent ranking of the retrieved results, which leads to poor performance. An important reason for this ranking problem has been a lack of information. Therefore some GIR research groups have tried to fill this gap using robust geographical resources (i.e. a geographical ontology), while other groups with the same aim have used relevance feedback techniques instead. This paper explores the use of random indexing (RI; a vector space methodology that generates semantic context vectors for words based on co-occurrence data) and holographic reduced representations (HRRs; a novel representation for textual structure) as re-ranking mechanisms for GIR. We show the feasibility of these techniques for re-ranking documents in GIR. Our results report an improvement of 7 % in mean average precision (MAP) when compared to the traditional vector space model.

Categories and Subject Descriptors

H.3.3 **[Information Storage and Retrieval**]: Information Search and Retrieval - *Search process*

General Terms

Algorithms, Measurement, Performance, Experimentation, Theory.

Keywords

Geographic Information Retrieval, Vector Model, Random Indexing, Context Vectors, Holographic Reduced Representation, Rank Merging.

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

1. INTRODUCTION

Information Retrieval (IR) is a discipline involved with the organization, storage, representation, and recovery of information items. IR systems are designed to provide, in response to a user query, references to documents which could contain the information desired by the user. In order to evaluate how well an IR system performs, the relevance of retrieved documents must be quantified. If a document according to a query is judged by the user to be interesting, it is relevant; otherwise, it is considered irrelevant. This relevance is measured by a similarity function which computes likenesses between a document and a query.

Geographic Information Retrieval (GIR) is a specialized IR branch, where the search of documents is based not only on conceptual keywords, but also on spatial information. Therefore, GIR deals with information related to geographic locations, such as the names of rivers, cities, lakes or countries. Information that is related to a geographic space is called geo-referenced information, which is often linked to locations expresed as place names or phrases that suggest a geographic location. There are several problems when considering name places as locations: *a)* different places could have the same name *b)* some place names change over time *c)* the geographic extension that the place name denotes can be extended, reduced or changed over time, *d)* some place names are temporal or cultural conventions rather than official names *e)* the borders of a location can be fuzzy, and finally, *f)* some place names denote an association linked to an area rather than a location.

There are also places that are geo-referenced by mentioning objects or phenomena that are not locations. For instance, consider the query: "Cities near active volcanoes". Traditional IR techniques will not be able to produce an effective response to this query, since the user's information need is not explicit. Therefore, GIR systems have to interpret implicit information contained in documents and queries to provide an appropriate response to a query. This additional information would be needed in the example to match the word "Cities" and "volcanoes" with real names.

Accordingly with the conditions stated above, GIR requires among others an understanding of time, cultural environment, historical events, and natural phenomena. They need to go beyond lexical analysis and then capture or use some semantic information. Because of this our main hypothesis is that by capturing some context information contained in geographical

queries and documents, it is possible to improve GIR systems performance.

We have observed that the GIR problem is usually solved through traditional or minor variations of common IR techniques. As a result: a) traditional IR machines are able to retrieve the majority of relevant documents for most geographical queries, but b) they have severe difficulties generating a pertinent ranking of the retrieved results, which leads to a poor performance.

An important source of the ranking problem has been the lack of information. Therefore some GIR research groups have tried to fill this gap using robust geographical resources (i.e. a geographical ontology), while other groups with the same aim have used relevance feedback techniques instead.

As an alternative, our method to reduce the lack of information suggests incorporating context information and syntactic structure to improve the document ranking, which is our main contribution. In particular, we consider the use of Random Indexing (RI) to produce context vectors using document occurrence representation (DOR) and Holographic reduced representation (HRR) to represent syntactic structure, which to the best of our knowledge, have neither been used in IR nor GIR.

The remainder of this paper is organized as follows. In Section 2 we briefly review some GIR related work. Section 3 presents Random Indexing word space technique and related work. Section 4 introduces the concept of Holographic Reduced Representations (HRRs) and presents how to use them to represent documents according to their spatial relations. Section 5 explains the experimental setup. Section 6 shows the results obtained with Geo-CLEF collection and queries from 2005 to 2008. Finally, Section 7 concludes the paper and gives some directions for further work.

2. GIR RELATED WORK

Geographical Information Retrieval (GIR) considers the search for documents based not only on conceptual keywords, but also on spatial information (i.e., geographical references) [21]. Formally, a geographic query (geo-query) is defined by a tuple <*what, relation, where*> [22]. The *what* part represents generic terms (non-geographical terms) employed by the user to specify its information need, it is also known as the thematic part. The *where* term is used to specify the geographical areas of interest. Finally, the *relation* term specifies the "spatial relation", which connects *what* and *where*. Complex geo-queries, which contain multiple spatial relations, are combinations of these tuples.

GIR was evaluated at the CLEF forum [17] from 2005 to 2008, under the name of the 'GeoCLEF' task [18]. Several approaches were focused on solving the ranking problem during these years. Two well known strategies to improve this problem were: *a)* query expansion through some feedback strategy, and *b)* re-ranking retrieved elements through some adapted similarity measures.

These strategies have two main research paths: first, research groups that have paid attention to construct and include robust geographical resources in the process of retrieving and/or ranking documents. And second, those groups that ensure that geographical queries can be treated and answered employing very little geographical knowledge. As an example of those in the first category, some research groups employ geographical resources in the process of relevance feedback [7]. Here, they first recognize the geographical entities (geo-terms) in the given geo-query by employing a GeoNER[1] system. Afterwards, they then employ a geographical ontology to search for these geo-terms, and retrieve some other related geographical terms. Then, the retrieved terms are given as feedback elements to the GIR machine. Some others groups that focus on the re-ranking problem propose algorithms that consider the existence of Geo-tags[2]; therefore, the ranking function measures levels of topological space proximity, or geographical closeness among the geo-tags of retrieved documents and geo-queries [8]. In order to achieve this, geographical resources (e.g., a geographical ontology) are needed. Although these strategies work well, in real world applications neither "geo-tags" nor robust geographical resources are always available; in addition, a major problem that these strategies are forced to solve is the geo-terms ambiguity.

In contrast, groups that do not depend on any robust geographical resource have proposed and applied variations of the relevance feedback process, where no special consideration for geographic elements is made [9], and they have achieved very good performance results. There are also groups focusing on the re-ranking problem; they consider the existence of several lists of retrieved documents (from one or many IR machines). Therefore, the ranking problem is seen as an information fusion problem, without any special processing for geo-terms contained in the retrieved documents. Some simple strategies only apply logical operators to the lists (e.g., AND) in order to generate one final re-ranked list [10], while some other work applies techniques based on information redundancy (e.g., CombMNZ)[11, 12].

Our study fits into the second research path, since we do not depend on the availability of robust geographical resources, but we contemplate the use of different lists of ranked retrieved documents by looking to improve the base ranker efficiency.

This work differs from previous efforts in that we consider, in the re-ranking process, the context information and syntactic structure contained in geo-queries and retrieved documents, this information is captured by RI, DOR and HRR, which to the best of our knowledge, have neither been used in IR nor GIR. Therefore, it is and important contribution. The following two sections describe RI, DOR and HRR; after that, section 5.3 explains how they are used in our experiments.

3. RANDOM INDEXING

The vector space model (VSM) [19] is probably the most widely used IR model, mainly because of its conceptual simplicity and acceptable results. The model creates a space in which both documents and queries are represented by vectors. This vector space is represented by a $n \times m$ matrix, known as term-document matrix, where n is the number of different terms, and m is the number of documents, in the collection. The VSM assumes that term vectors are pair-wise orthogonal. This assumption is very restrictive because using, for example, the cosine as the similarity function, the value assigned to each document/query pair is only

[1] Geographical Named Entity Recognizer

[2] A Geo-tags is a label that indicates the geographical focus of certain document or geographical query.

determined by the terms the query and the document have in common, not by the terms that are semantically similar in both.

There have been various extensions to the VSM. One example is Latent Semantic Indexing (LSI) [20], a method of word co-occurrence analysis to compute semantic vectors (context vectors) for words. LSI applies singular-value decomposition (SVD) to the term-document matrix in order to construct context vectors. As a result the dimension of the produced vector space will be significantly smaller; consequently the vectors that represent terms cannot be orthogonal. However, dimension reduction techniques such as SVD are expensive in terms of memory and processing time. As an alternative, there is a vector space methodology called Random Indexing [3], which presents an efficient, scalable, and incremental method for building context vectors.

Random Indexing (RI) accumulates context vectors for words based on co-occurrence data. The technique can be described as: a) A unique random representation known as index vector is assigned to each context (either document or word), consisting of a vector with a small number of non-zero elements, which are either +1 or -1, with equal amounts of both. For example, if the index vectors have twenty non-zero elements in a 1024-dimensional vector space, they have ten +1s and ten -1s. Index vectors serve as indices or labels for words or documents; b) Index vectors are used to produce context vectors by scanning through the text. Every time a target word occurs in a context, the index vector of the context is added to the context vector of the target word. Thus, with each appearance of the target word *t* with a context *c* the context vector of *t* is updated as follows:

$$ct + = ic$$

where *ct* is the context vector of *t* and *ic* is the index vector of *c*. In this way, the context vector of a word keeps track of the contexts in which it occurred. Similarity between words is then measured by comparing their context vectors, e.g., measuring their cosine.

Particularly, we apply RI to generate context vectors using 'document occurrence representation' (DOR). This representation is based on the work of Lavelli et al. [16]. They compare DOR with another representation, named 'term co-occurrence representation' (TCOR). In DOR the meaning of a term is considered as the bag of documents in which it occurs, whereas in TCOR the meaning of a term *t* is viewed as the bag of terms with which it co-occurs, given a window centered in *t*. In the application of these representations, Lavelli et al. use both DOR and TCOR for term categorization and term clustering tasks with WordNetDomains (42) as an evaluation resource. Their results show that TCOR outperform the DOR representation in both tests. They argue that TCOR better identifies perfect synonyms. However, Sahlgren [40] argues that DOR or large window TCOR representations better capture associative relations between words, what the words are about, while small window TCOR representations better capture semantic relations between words, what the word means. This suggests that DOR and large window TCOR representations are better for topical information tasks, such as retrieval information. Our experiments confirm this, since DOR shows itself to be the better mechanism for re-ranking the retrieved documents for GIR: however, in this paper we only present the DOR results because TCOR results were considerably lower.

In DOR, the term t_j is represented as a vector $\vec{t}_j = (w_{1j}, w_{2j},\ldots,w_{mj})$ of context weights, where m is the cardinality of the document collection, and w_{kj} represents the contribution of context k to the specification of the semantics of term t_j. We use DOR to represent the content of a document as a bag of concepts (BoC), which is a recent representation scheme introduced by Sahlgren & Cöster[3]. BoC representation is based on the idea that the meaning of a document can be considered as the union of the meanings of its terms. This is accomplished by generating term context vectors for each term within the document, and generating a document vector as the weighted sum of the term context vectors contained within that document.

There are works that have validated the use of RI in text processing tasks: for example, Kanerva et al. in [13] used Random Indexing to solve the part of the TOEFL, in which given a word, the subject is asked to choose its synonym from a list of four alternatives. The results obtained are 48-51% correct using DOR. Karlgren and Sahlgren [4, 14] used TCOR to enhance the performance of Random Indexing in the same task to 64.5% – 67% correct answers, which is comparable to results reported for foreign applicants to U.S. colleges (64.5%). Sahlgren & Karlgren [15] demonstrated that Random Indexing can be applied to parallel texts for automatic bilingual lexicon acquisition. In their experiments, they demonstrated how bilingual lexica can be extracted using Random Indexing working with parallel data for Swedish–Spanish and English–German data. They computed the overlap between the automatically extracted lexicon and the goldstandard (Lexin's online Swedish–Spanish lexicon, and TU Chemnitz' online English–German dictionary), getting around 60% when only terms with frequency above 100 occurrences in the source languages were included. Sahlgren & Cöster [3] used Random Indexing to carry out text categorization. They use Reuters-21578 collection and BoC to represent the documents. These representations are the input to a support vector machine classifier. Their results using BoC were comparable to standard text representations at around 82 %.

4. HRR DOCUMENT REPRESENTATION

In addition to random indexing, which produces context vectors using DOR, we explore the use of syntactic structures (prepositional phrases as '*in Asia'*) to represent spatial relations and re-rank the retrieved documents. The traditional IR methods that include compound terms extract them, and then include these compound terms as new VSM terms [5, 6]. We explore a different representation of such structures, which uses a special kind of vector binding (called holographic reduced representations (HRRs) [2]) to reflect text structure and distribute syntactic information across the document representation. HRRs use circular convolution to associate items, which are represented by vectors. A processing text task where HRRs have been used together with Random Indexing is text classification, where they have shown improvement under certain circumstances, having BoC as its baseline [1]. (Up to now, we are not aware of other work that uses RI together with HRRs for clasification.)

The Holographic Reduced Representation, HRR, was introduced by Plate [2] as a method to represent information in a computer that could be suitable for modeling how the brain processes

information. It had been thought that structure of language could not be encoded, in a practical way, by distributed representations, but HRRs provide an alternative to this situation. HRRs are vectors whose entries follow a normal distribution N(0,1/n). They allow us to express structure using a circular convolution operator to bind terms, without increasing vector dimensionality. This circular convolution operator ($\otimes$) binds two vectors $\bar{x} = (x_0, x_1,\ldots,x_{n-1})$ and $\bar{y} = (y_0,y_1,\ldots,y_{n-1})$ to produce $\bar{z} = (z_0, z_1, \ldots, z_{n-1})$ where $\bar{z} = \bar{x} \otimes \bar{y}$ is defined as:

$$z_i = \sum_{k=0}^{n-1} x_k y_{i-k} \quad i = 0 \text{ to } n\text{-}1 \text{ (subscripts are module } n)$$

A finite-dimensional vector space over the real numbers with circular convolution and the usual definition of scalar multiplication and vector addition form a commutative linear algebra system, so all the rules that apply to scalar algebra also apply to this algebra [2].

We adopt HRRs to build a text representation scheme in which spatial relations could be captured. Therefore, to define an HRR document representation the following steps are done: a) We determine the index vectors for the vocabulary by adopting the random indexing method, as described earlier; b) The documents are labelled using a Name Entity Recognition System; c)For each location entity, its index vector tf.idf-weighted is bound to its location role. This location role is the preposition (i.e. in, near, around, across, etc.) extracted from the text considering the previous preposition to the location entity, which is represented as an HRR; d) The resulting HRRs with the spatial relations encoded, are multiplied by an attenuating factor α, and then added to obtain a single HRR vector representing the document, which is then normalized. For example, when given a spatial relation: *R = in Asia.* Therefore, *R* will be represented using the index vectors r_1 for *Asia*, where r_1 will be joined to its location role, an HRR, $role_1$ which represents the relation *in.* Then the *in Asia* vector will be:

$$\vec{R} = (role_1 \otimes \vec{r}_1).$$

Thus, given a document D, with spatial relations *in:* t_{x1}, t_{y1}, its vector will be built as:

$$\vec{D} = \left\langle \alpha((role_1 \otimes \vec{t}_{x1}) + (role_1 \otimes \vec{t}_{y1})) \right\rangle$$

where $\langle\ \rangle$ denotes a normalized vector and α is a factor less than one intended to lower the impact of the coded relations. Queries are represented in a similar way.

5. EXPERIMENTAL SETUP

In any experimental science area, as is the case of IR, it is crucial to have systems that allow large-scale experiments to optimally test new methods. We used in our experiments Lemur[3], an open-source system designed to facilitate research in information retrieval. The results produced by the VSM, configured in Lemur, were taken as our baseline.

[3] http://www.lemurproject.org/

5.1 Data

Our experiments were conducted in English. The collection for this specific language is composed of news articles taken from the Glasgow Herald (British) 1995 and LA Times (American) 1994. The news covers both national and international events; accordingly they include several geographic references. Table 1 shows some collection data.

Table 1. GeoCLEF English document collection

Collection Name	*Origin*	*Number of documents*
GH95	The Glasgow Herald	56,472
LAT94	The Los Angeles Times	113,005
		Total: 169,477

5.2 Topics

We worked with the queries from GeoCLEF 2005 to GeoCLEF 2008. A total of 25 topics or queries were emitted for each year to total at the last conference in 2008 a set of 100 queries. Figure 1 shows the structure of each topic. The main query or title is between labels <EN-title> and </EN-title>. A brief description (<EN-desc>,</EN-desc>) and a narrative (<EN-narr>,</EN-narr>) are given too. These last two fields usually increase the requirement specificity of the original query.

Fig. 1. Topic GC030: Car bombings near Madrid

<top>
<num> GC030</num>
<EN-title>Car bombings near Madrid</EN-title>
<EN-desc> Documents about car bombings occurring near Madrid</EN-desc>
<EN-narr> Relevant documents treat cases of car bombings occurring in the capital of Spain and its outskirts</EN-narr>
</top>

Participant research groups at GeoCLEF have the freedom to employ any or all of the three fields in their experiments. We took the title and description for all our experiments, except for the query representations with HRR, where we considered the narrative statement in order to have better relations for representation. It is worth mentioning that Lemur results are lower when the narrative is included than when only title and description are.

5.3 Representations

To prove our hypothesis we consider two phases. The aim of the first was to retrieve as many relevant documents as possible for a given query, whereas the function of the second was to improve the final ranking of the retrieved documents by applying DOR and HRR representations.

Lemur was used to process the 169,477 documents, first with the queries for 2005 and after with the queries of the other years. Thereafter, only the top 1000 documents ranked by the VSM, were selected for each query. With this process a sub-collection of at most 25,000 documents was produced for each year.

This annual sub-collection was processed to generate the BoC representations of its documents and queries. BoC representations were generated by first stemming all words in the corpus, using the Porter stemmer to reduce the words to their root form. We then used Random Indexing to produce context vectors for the given sub-collection. The dimensionality of the context vectors was fixed at 4092 dimensions. The index vectors were generated with +1s and -1s distributed over the 4092 dimensions, representing about 0.49% of the dimensionality of the vector. These context vectors were then tf × idf-weighted and added up for each document and query, as described earlier to produce DOR representations.

On the other hand, HRRs were generated by firstly tagging all sub-collections with the Named Entity Recognition of Stanford University[4]. Afterwards, the single word locations preceeded by the preposition *in* were extracted. This restriction was taken after analyzing the queries for each year and realizing that only about 12% of them had a different spatial relation. HRRs for documents and queries were then produced by generating a 4096- HRR to represent the *in* relation. The *in* HRR vector was then bound to the index vector of the identified location words by a Fast Fourier Transform implementation of circular convolution, tf × idf-weighted and added to each document, as described earlier to generate SRR representations.

5.4 Re-ranking

We present the results of three processes for re-ranking documents. The first is named Geo-DOR, which is created by combining the Lemur similarity value with its corresponding value from DOR. The second, which was given the name of Geo-SRR, follows the same process as described above, but now with the similarity lists generated by Lemur and SRR. Finally the three similarity lists (Lemur; DOR and SRR) are combined to form Geo-DOR-SRR. Thus the similarity value was calculated by the cosine function in all cases.

5.5 Evaluation

The evaluation of the results after re-ranking the documents was carried out with two metrics that have demonstrated their stability to compare IR systems:

- Mean Average Precision (MAP), which is defined as the average of all the *AveP* obtained for each query. *AvgP* is defined as:

$$AvgP = \frac{\sum_{r=1}^{m} P(r)\ x\ rel(r)}{n}$$

  Where *P(r)* is the precision at *r* considered documents, *rel(r)* is a binary function which indicates if document *r* is relevant or not for a given query *q*; *n* is the number of relevant documents for *q; m* is the number of relevant documents retrieved for *q*.

- R-Prec, which is defined as the precision reached after R documents have been retrieved, where R is the number of relevant documents for the current query.

6. RESULTS

We consider two experiments:

a) The aim of the first was to prove that incorporating context information and syntactic structure for re-ranking documents in GIR could improve precision, i.e. to explore the DOR and HRR effectiveness as re-ranking strategies for GIR.

b) The objective of the second was to compare our strategies against a traditional mechanism for re-ranking documents: Pseudo Relevance Feedback (PRF).

6.1 First Experiment

Table 2 compares Lemur results, with the ones produced after adding to it, DOR, SRR and DOR-SRR similarity lists. Notice how Geo-DOR increments MAP in a constant form, always above 5%. The increment with Geo-SRR is very slightly at 1%, however. But when added together with DOR, the difference is raised by a further 2% to a total of 7% in 2008 and by a lesser degree in 2005-2007. Table 2 illustrates these favorable percentages to our proposals (numbers in bold).

Table 3 shows the same comparison, but now in terms of R-Prec. Geo-DOR results are higher than Lemur only for years 2006 and 2007. On the other hand, Geo-SRR results have almost the same behavior as MAP results; Geo-DOR-SRR were only lower in 2005, but showed a significant improvement in 2006.

Because the results for 2006 in terms of MAP were unfavorable, we tried to find a justification: Table 4 demonstrates statistics for each sub-collection where vocabulary size; number of different documents per sub-collection; number of words, spatial relations, and relevant document per query are all shown. Finally, the total number of relevant documents per year is displayed in Table 4. All the four sub-collections are uniform considering the vocabulary size, the number of documents and the number of words per query. In contrast, the number of spatial relations per query is notably higher for 2008. Also, the number of relevant documents per query and total number are considerably lower for 2006.

We also found that the SRR representation contributes to improve precision when there are enough relations that represent the query. This representation actually increased the precision in 2008, where the queries had on average 7 spatial relations, yet had little effect in the years where the number of relations is less than 2.

It is known that the behavior of retrieval methods depends on the number of relevant documents. For example, a blind feedback method works well for broad topics that have many relevant documents but may harm topics with few relevant documents; we believe that this is also true for our proposed representations. To support this idea we examined the queries with the highest number of relevant documents having also the highest number of relevant retrieved documents. Table 5 shows queries that meet these conditions for each year. There it is clear that even for 2006

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

Table 2: MAP results for Geo-CLEF collection (2005 – 2008)

	2005	*2006*	*2007*	*2008*
Lemur	0.3191	0.2618	0.1612	0.2347
Geo-DOR	0.338	0.2495	0.1695	0.2475
% Diff	**5.92**	-4.7	**5.15**	**5.45**
Geo-SRR	0.3193	0.2619	0.1623	0.2357
% Diff	**0.06**	**0.04**	**0.68**	**0.43**
Geo-DOR-SRR	0.3381	0.2495	0.1699	0.2512
% Diff	**5.95**	-4,7	**5.40**	**7.03**

Table 3: R-Prec results for Geo-CLEF collection (2005 – 2008)

	2005	*2006*	*2007*	*2008*
Lemur	0.3529	0.2423	0.1782	0.2610
Geo-DOR	0.3520	0.2599	0.1843	0.2605
% Diff	-0.26	**7.26**	**3.42**	-0.19
Geo-SRR	0.3544	0.2423	0.1784	0.2631
% Diff	**0.43**	0	**0.11**	**0.80**
Geo-DOR-SRR	0.3472	0.2599	0.1843	0.2688
% Diff	-1.62	**7.26**	**3.42**	**2.99**

Table 4: Statistics for the sub-collections used to evaluate the proposed representations

	2005	*2006*	*2007*	*2008*
vocabulary	89446	93887	91929	90557
documents	20267	20851	21372	20224
words/ query	9	9	10	8
relations/ query	0.72	2	2	7
relevant docs./ query	41	15	29	31
total of relevant docs.	1028	378	650	747

(when the query has a reasonable number of relevant documents) the results are favorable for our method. The improvement goes from 6.43 % for query 31 in 2006 to 24.78% for query 87 in 2008.

6.2 Second Experiment

Finally, we compare the Geo-DOR-SRR results with a traditional re-ranking method known as Pseudo Relevance Feedback (PRF). This was initially proposed to improve the retrieval accuracy of Lemur. PRF treats the *n* top ranked documents as true relevant documents for a given query. In order to apply this approach, we used the VSM, represented queries and documents as tf-idf vectors, and similarity was computed by the cosine function.

Table 5. Queries with the highest number of relevant documents and relevant retrieved documents

Year	*Id. Qry*	*Rel*	*Rel. Ret.*	*Lemur*	*Geo-DOR-SRR*	*% Dif.*
2005	15	110	110	0.6691	0.7363	10.04
2006	31	59	59	0.2844	0.3027	6.43
2007	51	112	106	0.4864	0.5714	17.48
2008	87	106	104	0.2115	0.2639	24.78

Table 6. Difference between PRF MAP and MAP of Geo-DOR-SRR in 2008, which was 0.2512

# selected terms	PRF with # top documents		% diff with # top docs.	
	5	10	5	10
5	0.2214	0.2064	**13.45**	**21.70**
10	0.2252	0.2025	**11.54**	**24.04**
15	0.2017	0.2209	**24.54**	**13.71**

Queries were expanded by adding the *k* words selected from the *n* top documents, and then a second IR process was made with the expanded query. Table 6 presents results when the top 5 and 10 documents are taken to extract 5, 10, and 15 words. Query texts are built from title and description fields. The difference in MAP between this traditional technique and our Geo-DOR-SRR proposal is about 11% or higher in favour of our method.

7. CONCLUSION AND FUTURE WORK

In this paper we have presented two document and query representations for re-ranking documents and improving precision for Geographic Information Retrieval. RI builds context vectors with DOR, which capture semantic information that has allowed precision improvements at above 5% for the years 2005, 2007 and 2008. HRR representation works well with 2008 data where queries have enough spatial relations for representation. Our results have been compared with one of the most accurate IR models: theVSM.

The results of our study showed that: i) traditional IR methods are able to retrieve relevant documents for geographic queries; ii) a lack of relevant documents in a collection causes bad ranking and produces low precision; iii) both DOR and HRR representations are able to improve the base ranker; iv) when more relations are added to the HRRs, a better ranking is achieved v) comparing our method against a traditional re-ranking strategy PRF, results in higher scores for this new method. Therefore, overall the results demonstrate that our approach performs better.

We will continue working with other collections that provide us more specific contexts and conceptual relations, allowing us to explore in-depth the usefulness of the proposed representations as a mechanism for re-ranking documents to improve precision.|

9. REFERENCES

[1] Fishbein, J.M., Eliasmith, C: Integrating structure and meaning: A new method for encoding structure for text classification. In: Advances in Information Retrieval: Procs. of the 30th European Conference on IR Research, vol. 4956 of LNCS, ed. C. Macdonald, et al., pp. 514–521, Springer (2008).

[2] Plate, T.A.: Holographic Reduced Representation: Distributed representation for cognitive structures. CSLI Publications, (2003).

[3] Sahlgren, M., Cöster R.: Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. In: Procs. of the 20th International Conference on Computational Linguistics, pp. 487– 493 (2004).

[4] Sahlgren, M.: Vector-based semantic analysis: Representing word meanings based on random labels. In Procs. of the ESSLLI Workshop on Semantic Knowledge Acquisition and Categorization, Helsinki, Finland, (2001).

[5] Mitra M., Buckley C., Singhal A., Cardie C.: An Analysis of Statistical and Syntactic Phrases. In: Procs. of RIAO-97, 5th International Conference, pp. 200-214 (1997)

[6] Evans D., Zhai C.: Noun-phrase Analysis in Unrestricted Text for Information Retrieval. In: Procs. of the 34th Annual Meeting on Association for Computational Linguistics, pp. 17-24 (1996)

[7] Wang R., Neumann G.: Ontology-based query construction for Geoclef. In: Working notes for the CLEF 2008 Workshop, Aarhus, Denmark (2008).

[8] Martinis B., Cardoso N., Chavez M. S., Andrade L., and Silva M. J.: The University of Lisbon at Geoclef 2006. In: Working notes for the CLEF 2006 Workshop, Alicante, Spain. (2006).

[9] Larson R. R.: Cheshire at Geoclef 2008: Text and fusion approaches for GIR. In: Working notes for the CLEF 2008 Workshop, Aarhus, Denmark. (2008).

[10] Ferrs D., Rodríguez H.: Talp at GEOCLEF 2007: Using terrier with geographical knowledge filtering. In Working notes for the CLEF 2007 Workshop, Budapest, Hungary. (2007).

[11] Larson R. R.: Cheshire II at GEOCLEF 2005: Fusion and query expansion for GIR. In: Working notes for the CLEF 2005 Workshop, Wien, Austria (2005).

[12] Villatoro-Tello E., Montes-y-Gómez M, Villaseñor-Pineda L. INAOE at GEOCLEF 2008: A ranking approach based on sample documents. In: Working notes for the CLEF 2008 Workshop, Aarhus, Denmark. (2008).

[13] Kanerva, P., Kristofersson J., Holst A.: Random Indexing of text samples for Latent Semantic Analysis. Procs. of the 22nd annual conference of the cognitive science society. New Jersey: Erlbaum. (2000).

[14] Karlgren, J., Sahlgren M.: From words to understanding. In: Uesaka, Y., Kanerva P., Asoh H.: Foundations of real-world intelligence. Stanford: CSLI Publications, 2001.

[15] Sahlgren. M., Karlgren J.: Automatic bilingual lexicon acquisition using Random Indexing of parallel corpora. Journal of Natural Language Engineering Special Issue on Parallel Texts, (2005).

[16] Lavelli A., Sebastián F., Zanoli R.: Distributional term representations: an experimental comparison. In: CIKM ’04: Procs. of the thirteenth ACM conference on Information and knowledge management, pp. 615–624, New York, NY, USA, ACM Press. (2004).

[17] Cross-lingual evaluation forum. http://www.clef-campaign.org/, May 2009.

[18] Mandl T., Carvalho P., Gey F., Larson R., Santos D., Womser-Hacker C., Di Nunzio G., Ferro N.: Geoclef 2008: the CLEF 2008 Cross Language Geographic Information Retrieval Track Overview. In: Working notes for the CLEF 2008 Workshop, Aarhus, Denmark, (2008).

[19] Salton, G.; Wong, A., Yang, C. S.: A vector space model for automatic indexing. Communications of the ACM, v.18 n.11, pp.613-620, (1975).

[20] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, pp.391–407. (1990)

[21] Henrich A., Lüdecke V.: Characteristics of Geographic Information needs. In Procs. of Workshop on Geographic Information Retrieval, GIR'07, Lisbon, Portugal, ACM Press. (2007)

[22] Andrade. L, Silva. M. J.: Relevance ranking for geographic IR. In: Procs. of 3rd Workshop on Geographic Information Retrieval, SIGIR'06, Seattle, USA, ACM Press. (2006).