Concept Based Representations for Ranking in Geographic Information Retrieval*

Maya Carrillo[1,2], Esaú Villatoro-Tello[1], A. López-López[1], Chris Eliasmith[3], Luis Villaseñor-Pineda[1], Manuel Montes-y-Gómez[1]

[1]Coordinación de Ciencias Computacionales, INAOE
Luis Enrique Erro 1, Santa Maria Tonantzintla, Puebla, México, C.P.72840
{cmaya,villatoroe,allopez,villasen,mmontesg}@inaoep.mx
[2]Facultad de Ciencias de la Computación, BUAP
Av. San Claudio y 14 Sur Ciudad Universitaria, 72570 Puebla, México
[3]Department of Philosophy, Department of Systems Design Engineering, Centre for Theoretical Neuroscience, University of Waterloo
200 University Avenue West Waterloo, Canada
celiasmith@uwaterloo.ca

Abstract. Geographic Information Retrieval (GIR) is a specialized Information Retrieval (IR) branch that deals with information related to geographical locations. Traditional IR engines are perfectly able to retrieve the majority of the relevant documents for most geographical queries, but they have severe difficulties generating a pertinent ranking of the retrieved results, which leads to poor performance. A key reason for this ranking problem has been a lack of information. Therefore, previous GIR research has tried to fill this gap using robust geographical resources (i.e. a geographical ontology), while other research with the same aim has used relevant feedback techniques instead. This paper explores the use of Bag of Concepts (BoC; a representation where documents are considered as the union of the meanings of its terms) and Holographic Reduced Representation (HRR; a novel representation for textual structure) as re-ranking mechanisms for GIR. Our results reveal an improvement in mean average precision (MAP) when compared to the traditional vector space model, even if Pseudo Relevance Feedback is employed.

Key words: Geographic Information Retrieval, Vector Model, Random Indexing, Context Vectors, Holographic Reduced Representation.

1 Introduction

Geographic Information Retrieval (GIR) deals with information related to geographic locations, such as the names of rivers, cities, lakes or countries [18].

* The first and second author were supported by scholarship 208265 and 165545 respectively, granted by CONACYT, while the third, fifth and sixth author were partially supported by SNI, Mexico. In addition this work has been supported by Conacyt Project Grant 61335.

Information that is related to a geographic space is called geo-referenced information, which is often linked to locations expressed as place names or phrases that suggest a geographic location. For instance, consider the query: "ETA in France". Traditional IR techniques will not be able to produce an effective response to this query, since the user information need is very general. 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 "France" with other French cities as Paris, Marseille, Lyon, etc.

Recent developments in GIR systems have demonstrated that the GIR problem is partially solved through traditional or minor variations of common IR techniques. It is possible to observe that traditional IR engines are 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 source of the ranking problem has been the lack of information. Therefore, previous research in GIR has tried to fill this gap using robust geographical resources (i.e. a geographical ontology), whilst other research has used relevance feedback techniques instead.

As an alternative, our method suggests representing additional information incorporating concept-based representations. We think that concept-based schemes provide important information, and that they can be used as a complement to the Bag of Words representations. Our goal is therefore to investigate whether combining word-based and concept-based representations can be used to improve GIR. In particular, we consider the use of two document representations: a)Bag of Concepts (BoC), as proposed by Sahlgren and Cöster [3], to represent a document as the union of the meanings of its terms; b)Holographic Reduced Representation (HRR) defined by Plate [2] to include syntactic structure. The purpose is to represent relations that give different ideas of location like: *in Paris, near Paris, across Paris*. This representation can help to state specific information for GIR.

The proposed BoC and HRR representations are vector representations constructed through the aid of Random Indexing (RI), a vector space methodology proposed by Kanerva et al [20].

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. Section 4 describes the Bag of Concepts representation. Section 5 introduces the concept of Holographic Reduced Representations (HRRs) and presents how to use them to represent documents according to their spatial relations. Section 6 explains the experimental setup. Section 7 shows the results obtained with Geo-CLEF collections and queries from 2007 to 2008. Finally, Section 8 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) [18]. Formally, a geographic query (geo-query) is defined by a tuple ⟨*what, relation, where*⟩ [19]. The *what* part represents generic terms (non-geographical terms) employed by the user to specify its information need, which 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*. For example in query: *Child labor in Asia*, the *what* part would be: *Child labor*, the *relation* term would be *in* and the *where* part *Asia*.

GIR was evaluated at the CLEF forum [14] from 2005 to 2008, under the name of the 'GeoCLEF' task [15]. Several approaches were focused on solving the ranking problem during these years. Common employed strategies are: a) query expansion through feedback relevance [6], [9], [10]; b) re-ranking retrieved elements through adapted similarity measures [7]; and c) re-ranking through information fusion techniques [9], [10], [11].

These strategies have been implemented following two main paths: first, techniques that have paid attention to constructing and including robust geographical resources in the process of retrieving and/or ranking documents. And second, techniques that ensure that geographical queries can be treated and answered by employing very little geographical knowledge.

As an example of those in the first category, previous research employed geographical resources in the process of query expansion. Here, they first recognize the geographical named 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. The retrieved terms are then used as feedback elements to the GIR engine. However, a major drawback with these approaches is the huge amount of work needed in order to create such ontologies: for instance, Wang et al. in [6] employ two different geographical taxonomies (Geonames[2] and WorldGazetter[3]) to construct a geographical ontology with only two spatial relations: *"part-of"* and *"equal"*. This leads to the fact that the amount of geographical information included in a general ontology is usually very small, which limits it as an effective geographical resource. Some other approaches that focus on the re-ranking problem propose algorithms that consider the existence of Geo-tags [4]; therefore, the ranking function measures levels of topological space proximity, or geographical closeness among the geo-tags of retrieved documents and geo-queries [7]. In order to achieve this, geographical resources are needed. Although these strategies

[1] Geographical Named Entity Recognizer
[2] Geonames geo coding web service: http://www.geonames.org/
[3] WorldGazetteer: http://www.world-gazetteer.com
[4] A Geo-tags is a label that indicates the geographical focus of certain document or geographical query.

work well for certain type of queries, in real world applications neither "geo-tags" nor robust geographical resources are always available.

In contrast, approaches that do not depend on any geographical resource, have proposed and applied variations of the query expansion process via relevance feedback without special consideration for geographic elements [8], [9]. Despite this, they have achieved acceptable performance results, sometimes even better than those obtained employing resource-based strategies. There is also work focusing on the re-ranking problem; it considers the existence of several lists of retrieved documents from one or more IR engines. For instance, one IR engine can be configured to manage a thematic index (i.e., non geographical terms), while another IR engine is configured to manage only geographical indexes [8], [9], [10], [11], [18]. Therefore, the ranking problem is seen as an information fusion problem; where simple strategies only apply logical operators to the lists (e.g., AND) in order to generate one final re-ranked list [10], while others apply techniques based on information redundancy (e.g., CombMNZ, Round-Robin or Fuzzy Borda)[8], [10], [11], [18].

Recent evaluation results indicate that there is not a notable advantage of resource-based strategies over methods that do not depend on any geographical resource [11]. Motivated by these results, our method does not depend on the availability of geographical resources, but we contemplate the use of different lists of ranked retrieved documents (VSM, BoC and HRR) looking for improvement of the base ranker efficiency by the combination.

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 additional information is captured by BoC and HRR representations, which need special vectors, built by Random Indexing (RI).

3 Random Indexing

The vector space model (VSM) [16] is probably the most widely known 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 V a n x 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 the similarity between each document/query pair is only determined by the terms they 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 Analysis (LSA) [17], a method of word co-occurrence analysis to compute semantic vectors (context vectors) for words. LSA applies singular-value decomposition (SVD) to V (the term-document matrix) in order to construct context vectors. As a result, the dimension of the produced vector space will be significantly smaller by grouping together words that mean similar things;

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 (RI) [3], which represents an efficient, scalable, and incremental method for building context vectors, which express the distributional profile of linguistic terms.

RI overcomes the efficiency problems by incrementally accumulating k - dimensional index vectors into a context matrix R of order n x k, where $k \ll m$, but usually on the order of thousands. This is done in a two steps: 1) 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 (ϵ) of non-zero elements, which are either +1 or -1, with equal amounts of both. For example, if 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; 2) Index vectors are used to produce context vectors by scanning through the text. Every time a target word (t) occurs in a context (c), the index vector of the context (ic) is added to the context vector of t (tc). Thus, the context vector of t is updated as: $tc += ic$.

In this way, R is a matrix of k-dimensional context vectors that are the sum of the terms' contexts. Notice that these steps will produce a standard term-document matrix V of order n x m if we use unary index vectors of the same dimensionality as the number of contexts. Such m-dimensional unary vectors would be orthogonal, whereas the k-dimensional random index vectors are only nearly orthogonal. However, Hecht-Nielsen [21] stated that there are many more nearly orthogonal directions in a high dimensional space than truly orthogonal directions, which means that context matrix R n x k will be an approximation of the term-document matrix F n x m.

The approximation is based on the Johnson-Lindenstrauss lemma [21], which states that if we project points in a vector space into a randomly selected subspace of sufficiently high dimensionality, the distances between the points are approximately preserved. Then, the dimensionality of a given matrix V can be reduced by projecting it through a matrix P.

$$R_{nxk} = V_{nxm} P_{mxk} \tag{1}$$

Random Indexing has several advantages: 1. It is incremental, which means that the context vectors can be used for similarity computations even after just a few documents have been processed; 2. It uses fixed dimensionality, which means that new data do not increase the dimensionality of the vectors; 3. It uses implicit dimensionality reduction, since dimensionality is much lower than the number of contexts in the data ($k \ll m$).

There are works that have validated the use of RI in text processing tasks: for example, Sahlgren & Karlgren [12] demonstrated that Random Indexing can be applied to parallel texts for automatic bilingual lexicon acquisition. Sahlgren & Cöster [3] used Random Indexing to carry out text categorization. This technique, as far as we know has not been used in IR, but similar techniques as SVD are well known and used in the area.

4 BoC Document Representation

BoC is a recent representation scheme introduced by Sahlgren & Cöster [3], which 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. Thus, the m documents in a collection D are represented as:

$$\mathbf{d_i} = \sum_{j=1}^{s} h_j \mathbf{g_j} \qquad i = 1, ..., m \tag{2}$$

where s is the number of terms in document d_i, $\mathbf{g_j}$ is the context vector of term j, and h_j is the weight assigned to term j according to the weighting scheme considered.

The context vectors used in BoC are generated using RI and 'Document Occurrence Representation' (DOR). DOR is based on the work of Lavelli et al. [13] and considers the meaning of a term as the bag of documents in which it occurs. When RI is used together with DOR, the term t is represented as a context vector:

$$\mathbf{t} = \sum_{k=1}^{u} \mathbf{b_k} \tag{3}$$

where u is the number of documents containing t, and $\mathbf{b_k}$ is the index vector of document k, then the contribution of document k to the specification of the semantics of term t. For instance, the context vector for a term t, which appears in the documents d_1 = [1, 0, -1, 0] and d_2 = [1, 0, 0, -1] would be [2, 0, -1, -1]. If the term t is encountered again in document d_1, the existing index vector of d_1 would be added one more time to the existing context vector to produce a new context vector for t of [3, 0, -2,-1]. Context vectors generated through this process are used to build document vectors as BoC. Thus, a document vector is the sum of the context vectors of its terms.

5 HRR Document Representation

In addition to BoC, we explore the use of syntactic structures (prepositional phrases such as *'in Asia')* to represent spatial relations and re-rank the retrieved documents. The traditional IR methods that include compound terms, extract and include them as new VSM terms [4], [5]. 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. Fishbein, and Eliasmith have used the HRRs together with Random Indexing for text classification, where they have shown improvement under certain circumstances,

having BoC as the baseline [1]. It is important to mention that up to now, we are not aware of other work that uses RI together with HRRs.

The Holographic Reduced Representation, HRR, was introduced by Plate [2] as a method for representing compositional structure in distributed representations. HRRs are vectors whose entries follow a normal distribution N(0,1/n). They allow to express structure using a circular convolution operator to bind terms. This circular convolution operator ($\otimes$) binds two vectors $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})$ and $\mathbf{y} = (y_0, y_1, \ldots, y_{n-1})$ to produce $\mathbf{z} = (z_0, z_1, \ldots, z_{n-1})$ where is defined as:

$$z_i = \sum_{k=0}^{n-1} x_k \; y_{i-k} \qquad i = 0 \; to \; n-1 (\text{subscripts are modulo} - n) \tag{4}$$

Circular convolution is an operator which does not increase vector dimensionality, making it excellent for representing hierarchical structures. We adopt HRRs to build a text representation scheme in which spatial relations (SR) could be captured. Therefore, to define an HRR document representation, the following steps are done: a) Determine the index vectors for the vocabulary by adopting the random indexing method, as described earlier; b) Tag text of documents using a Name Entity Recognition System; c) Bind the $tf.idf$-weighted index vector of each location entity to its location role. This location role is an HRR which represents a preposition (i.e. *in, near, around, across*, etc.) extracted from the text considering the preposition preceding the location entity; d) Add the resulting HRRs (where the spatial relations are encoded) to obtain a single HRR vector; e) Multiply the resulting HRR by an attenuating factor α; f) Normalize the HRR obtained so far, to get the vector which represents the document. For example, when given a spatial relation: $R = in \; Asia$, 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:

$$\mathbf{R} = (\mathbf{role_1} \otimes \mathbf{r_1}) \tag{5}$$

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

$$\mathbf{D} = \langle \alpha((\mathbf{role_1} \otimes \mathbf{t_{x1}}) + (\mathbf{role_1} \otimes \mathbf{t_{y1}}) \rangle \tag{6}$$

where α is a factor less than one intended to lower the impact of the coded relations. Queries are processed and represented in a similar way.

6 Experimental Setup

We used in our experiments Lemur[5]. The results produced by the VSM configured in Lemur were taken as our baseline.

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

Our experiments were conducted using the English document collection for the GeoCLEF track. This collection is composed of news articles taking 56, 472 from the Glasgow Herald (British) 1995 and 113, 005 from the LA Times (American) 1994 to total 169,477 news articles.

We worked with the queries of GeoCLEF 2007 and GeoCLEF 2008, a set of 50 queries (from number 51 to 100). These queries are described in three parts: a) the main query or title; b) a brief description; and c) a narrative. We took the title and description for all our experiments, except for the query representations with HRR, where we also considered the narrative statement in order to have improved relations for representation. It is worth mentioning that Lemur results worsen when the narrative is included.

To investigate whether combining word-based and concept-based representations can be used to improve the GIR, we considered two phases. The aim of the first was to retrieve as many relevant documents as possible for a given query, whereas the purpose of the second was to improve the final ranking of the retrieved documents by applying BoC and HRR representations.

Lemur was used to process the 169,477 documents, first with the queries for 2007 and then with the queries for 2008. Thereafter, only the top 1000 documents ranked by the VSM were selected for each query. These sub-collections were processed to generate the BoC representations of its documents and queries. BoC representations were generated by first stemming all words in the sub-collections, using the Porter stemmer. We then used Random Indexing to produce context vectors for the given sub-collection. The dimensionality of the context vectors was fixed at 4096. The index vectors were generated with 10 +1s and 10 -1s, distributed over the 4096 dimensions. This vector dimension and density were empirically determined. These context vectors were then $tf.idf$-weighted and added up for each document and query, as described earlier to produce BoC representations.

On the other hand, HRRs were generated by firstly tagging all sub-collections with the Named Entity Recognition System of Stanford University [6]. Afterwards, the single word locations preceded 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 locations by a Fast Fourier Transform implementation of circular convolution, $tf.idf$-weighted, added, and multiplied by $\alpha = 1/6$ to represent each document, as described earlier to generate spatial relations representations.

Finally, the evaluation of the results after re-ranking the documents was carried out with the Mean Average Precision (MAP).

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

7 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 use of BoC and HRR representations) b) The objective of the second was to compare our strategies against a traditional re-ranking mechanism known as Pseudo Relevance Feedback (PRF).

First Experiment. Table 1 compares Lemur results, with the results produced by adding the Lemur similarity values with its corresponding values from BoC to produce Lemur-BoC, which is a new list re-ranked according to the new values. Then the same process as described above was followed, but now adding Lemur-BoC values to HRR values to produce Lemur-BoC-HRR. We only considered the set of supported queries, that is, the queries that have at least one relevant document: 22 queries in 2007 and 24 in 2008. Notice how MAP is incremented in a constant way, always at above 7%.

Table 1. MAP results for Geo-CLEF collection (2007 - 2008)

	Lemur	Lemur-BoC	%Diff	Lemur-BoC-HRR	% Diff
2007	0.1832	0.2079	13.48	0.2085	13.81
2008	0.2445	0.2619	7.12	0.2628	7.48

From the queries considered in 2007, 1 query kept the same MAP produced by Lemur after adding BoC. The MAP of 5 queries decreased. Positively, there are 16 queries improved by BoC. The favorable percentages of improvement for 10 queries are observed in Table 2 above the 14%.

When HRRs were added to Lemur-BoC, only the query 64 that was not improved by BoC (and in consequence, not in Table 2) was affected. This query had a percentage of change equal to -4.35%, which was raised to 30.43% by the representation of its 5 spatial relations. From the queries shown in Table 2, the unaffected queries have none or one spatial relation, while the queries enhanced by adding the HRRs have on average 4.

We found that HRRs improve precision when there are distinctive and specific spatial relations, for example: *in Finland* instead of *in northern Europe*. Therefore when geographical information given is more precise, HRRs help to achieve improved effectiveness. However, when the number of retrieved relevant documents is low with few relations to compare, it is difficult to affect the ranking with the HRRs.

In 2008, 3 queries kept the same MAP produced by Lemur after adding BoC. The MAP of 9 queries decreased and 12 queries improved. Table 2 shows 10 queries improved by BoC where favorable percentages of improvement are depicted. From these 10 queries, those that were improved after adding the HRRs, have at least 2 spatial relations. Our conclusion is that the relative small

Table 2. MAP for query improvement by BoC in 2007 and 2008 and their spatial relations.

	Qry-ID	Lemur	Lemur - BoC	% Diff	SR	Lemur-BoC-HRR	%Diff. additional
Results 2007	52	0.0022	0.0038	72.73	0	0.0038	0.00
	57	0.204	0.2473	21.23	6	0.2577	4.21
	58	0.0197	0.0268	36.04	0	0.0268	0.00
	60	0.0022	0.0397	1704.55	1	0.0397	0.00
	61	0.0959	0.1321	37.75	1	0.1318	-0.23
	67	0.2569	0.2950	14.83	0	0.2950	0.00
	69	0.0701	0.0964	37.52	1	0.0963	-0.14
	70	0.043	0.0509	18.37	0	0.0509	0.00
	72	0.4859	0.6179	27.17	1	0.6179	0.00
	75	0.3522	0.4580	30.04	2	0.4612	0.70
Results 2008	76	0.44	0.4857	10.39	12	0.5000	2.94
	80	0.2518	0.2555	1.47	1	0.2555	0.00
	82	0.0005	0.0015	200.00	3	0.0018	20.00
	84	0.1385	0.2183	57.62	0	0.2183	0.00
	85	0.4554	0.4767	4.68	0	0.4767	0.00
	86	0.0592	0.1101	85.98	2	0.1130	2.63
	91	0.0625	0.1667	166.72	1	0.1667	0.00
	93	0.7375	0.8340	13.08	1	0.8340	0.00
	95	0.491	0.5320	8.41	6	0.5337	0.26
	96	0.2232	0.2418	8.33	11	0.2454	1.49

contribution to improve precision demonstrated by HRR is due to the limited amount of spatial relations presents in the set of queries used. We believe that the higher the number of spatial relations to be represented, the greater the contribution of this representation.

We perform a paired t-student test to measure the statistical significance of our MAP results. The MAP differences for GeoCLEF 2007 resulted significant in a confidence interval of 95% for both Lemur-BoC and Lemur-BoC-HRR; however the results are below the median of the year (0.2097) by 0.57%. In this year, the top system at CLEF reached a MAP of 0.2859 [9]. However, it used a very complex configuration and several external resources (four Geographical Gazetteers, a Feature Type Thesaurus to categorize geo-terms and a Shape Toolbox a database, which contains a "shape file" available for each country).

The MAP improvement for 2008 is not statistically significant. Even so, the MAP median of the participants in Geo-CLEF 2008 was of 0.2370 [15], which is 6.45% lower than that generated by our proposal. This year the team at the top obtained a MAP of 0.3040 [6]. They used two ontologies constructed manually, employing information from narratives. In addition they used Wikipedia in the retrieval process. In contrast we do not use any complex external resource.

Second Experiment. Finally, we compare the Lemur-BoC-HRR results with a traditional re-ranking method known as Pseudo Relevance Feedback (PRF). In order to apply this approach, we used the VSM, representing queries and documents as *tf-idf* vectors, and computing similarity with the cosine function. PRF treats the n top ranked documents as true relevant documents for a given query, then queries are expanded by adding the k words selected from the n top documents, and then a second IR process is done with the expanded query. Table 3 presents results (also for queries with relevant documents) when the top

2 and 5 documents are taken to extract 5, 10, and 15 words. Query texts are built from title and description fields. The values that improve Lemur MAP are depicted in bold and those obtained with our proposal in italics. The difference in MAP between PRF technique and our Lemur-BoC-HRR proposal is about 6.21% or higher in favor of our method in 2007 and 1.23% or higher in 2008.

Table 3. Difference between PRF MAP and Lemur-BoC-HRR MAP

	Lemur-BoC-HRR	PRF with 2 documents			PRF with 5 documents		
		5 terms	10 terms	15 terms	5 terms	10 terms	15 terms
GeoCLEF 2007	*0.2085*	**0.1925**	0.1617	0.1533	**0.1963**	0.1703	0.1593
% Difference		8.31	28.94	36.01	6.21	22.43	30.89
GeoCLEF 2008	*0.2628*	**0.2539**	**0.2596**	**0.2505**	0.2306	0.2242	0.2101
% Difference		3.51	1.23	4.91	13.96	17.22	25.08

8 Conclusion and Future Work

In this paper, we have presented two document representations for re-ranking documents and improving precision for GIR. RI was used to build context vectors to create BoC representations, which capture context information. It also defines index vectors used in the HRR representations. When working with RI, the appropriate selection of the values for vector length and vector density is an open research topic. Our results have been compared with the VSM in its Lemur implementation. They have showed that: i) BoC can improve the initial ranker. ii) HRR representation improved the ranking of queries. However, its utility could not be totally verified because of the lack of spatial relations to be represented; ii) we foresee that when more relations are added to the HRRs, a better ranking is achieved. It should be noted that in the experiments conducted, only one type of spatial relation (*in*) was considered: we think if more types of relations (*near, around, across, far*, etc.) are added as long as they are present in the queries; it could lead to improved results; iii) comparing our method against PRF produces higher scores for this new method. Therefore, the overall results demonstrate that our approach is appropriate for re-ranking documents in GIR

We will continue working with other collections where queries have not only spatial relations but other syntactic relations (i.e. compose nouns, verb-subject) which could be represented and together with the context information, allow us to explore in-depth the usefulness of the proposed representations as a mechanism for re-ranking documents to improve precision.

References

1. Fishbein, J.M., Eliasmith, C: Integrating structure and meaning: A new method for encoding structure for text classification. In: Advances in IR: Procs. of the

30th ECIR Conf. 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. 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)
5. Evans D., Zhai C.: Noun-phrase Analysis in Unrestricted Text for Information Retrieval. In: Procs. of the 34th Annual Meeting on ACL, pp. 17-24 (1996)
6. Wang R., Neumann G.: Ontology-based query construction for Geoclef. In: Working notes for the CLEF Workshop, Aarhus, Denmark (2008).
7. 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 Workshop, Spain (2006)
8. Larson R. R.: Cheshire at Geoclef 2008: Text and fusion approaches for GIR. In: Working notes for the CLEF 2008 Workshop, Aarhus, Denmark (2008)
9. Ferrés D., and Rodríguez H.: TLAP at GeoCLEF 2007: Using Terries with Geographic Knowledge Filtering. In Working notes for the CLEF 2007 Workshop, Budapest, Hungary (2009).
10. 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)
11. 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).
12. Sahlgren. M., Karlgren J.: Automatic bilingual lexicon acquisition using Random Indexing of parallel corpora. Journal of Natural Language Engineering Special Issue on Parallel Texts, vol. 11 n. 3:327-341, (2005).
13. Lavelli A., Sebastiani 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, ACM Press. (2004).
14. Cross-lingual evaluation forum. http://www.clef-campaign.org/, (2009).
15. Mandl T., Carvalho P., Gey F., Larson R., Santos D., Womser-Hacker C., Di Nunzio G., Ferro N.: Geoclef 2008: the CLEF 2008 Track Overview. In: Working notes for the CLEF Workshop, Aarhus, Denmark (2008).
16. 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).
17. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis, Journal of the ASIS, 41, pp.391-407. (1990)
18. Henrich A., Lüdecke V.: Characteristics of Geographic Information needs. In Procs. of Workshop on Geographic Information Retrieval, Lisbon, Portugal, ACM Press. (2007)
19. Andrade. L, Silva. M. J.: Relevance ranking for geographic IR. In: Procs. of 3rd Workshop on Geographic Information Retrieval, SIGIR'06, ACM Press. (2006).
20. Kanerva P., Kristoferson J., and Anders Holst A. Random indexing of text samples for latent semantic analysis. In Procs. of the 22nd Annual Conf. of the Cognitive Sc. Society, USA. (2000).
21. Sahlgren M.: An introduction to random indexing. In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Denmark, (2005).