Representing Context Information for Document Retrieval*

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

[1] Coordinación de Ciencias Computacionales, INAOE,
Luis Enrique Erro 1, Sta.Ma. Tonantzintla, 72840, Puebla, Mexico
[2] Facultad de Ciencias de la Computación, BUAP,
Av. San Claudio y 14 Sur Ciudad Universitaria, 72570 Puebla, Mexico
{cmaya,villatoroe,allopez,mmontesg,villasen}@inaoep.mx
[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. The bag of words representation (BoW), which is widely used in information retrieval (IR), represents documents and queries as word lists that do not express anything about context information. When we look for information, we find that not everything is explicitly stated in a document, so context information is needed to understand its content. This paper proposes the use of bag of concepts (BoC) and Holographic reduced representation (HRR) in IR. These representations go beyond BoW by incorporating context information to document representations. Both HRR and BoC are produced using a vector space methodology known as Random Indexing, and allow expressing additional knowledge from different sources. Our experiments have shown the feasibility of the representations and improved the mean average precision by up to 7% when they are compared with the traditional vector space model.

Keywords: Information Retrieval, Vector Model, Context Information, Random Indexing, Holographic Reduced Representation.

1 Introduction

Information retrieval (IR) is the branch of computer science which studies the retrieval of information from a collection of written documents. An IR process initiates when a user introduces a query into an IR system. Queries are statements of information needs, for example: "Floods in European cities". A query does not match a single document in a collection. Instead, several documents may be related to the query with different degrees of relevancy. Moreover, traditional

* The first author was supported by scholarship 217251/ 208265, second author by scholarship 165545 granted by CONACYT, while the third, fifth and sixth author were partially supported by SNI, Mexico.

T. Andreasen et al. (Eds.): FQAS 2009, LNAI 5822, pp. 239–250, 2009.

IR techniques will not be able to produce an effective response to our example, since the user's information need is not explicit. Consequently, context information would be needed to match the word "cities", with actual cities names. Therefore, both the explicit and context information contained in documents and queries have to be interpreted to provide appropriate responses

Over the past years, several methods have been proposed for IR (i.e. Boolean, probabilistic, language models); however the Vector Space Model (VSM) [8] is widely used because of its simplicity and acceptable results. This model uses word lists (Bag of Words (BoW)) to represent document contents. A word list is a description that does not express anything about context information.

There have been efforts to enrich the BoW. For instance, Mitra et al, Evans and Zhai [6,7], among others have investigated the use of phrases as part of text representation since the early days of information retrieval. Certainly, there has been the feeling that phrases should improve the specificity of indexing language and, consequently, the quality of text representation. The experimental results obtained however, do not support this intuition. These results have gone from small improvements in some collections to a decrease in effectiveness in others. We have observed that commonly, these methods include phrases as new VSM terms. Our idea, however, is to include them using a representation that reflects, rather than adds, syntactic structure and distributes it across the document representation.

Therefore we propose to extract compound terms (i.e. *information retrieval*), and binary linguistic relations (i.e. *(eagle, eat), (eat, fish), (in, Europe)*) from texts. Thereafter, rather than representing these relations as syntactic trees, semantic frames, or conceptual graphs, they can be represented as holographic reduced representations (HRRs) as proposed by Plate [3]. Being simple vectors, they can be indexed and retrieved rapidly as a result.

Other efforts to enrich the BoW have used the inherent semantic structure that exists in the association of terms with documents in the indexing process. Deerwester et al., Hofman, and Wong et al. [14, 15, 16] report results, where their models show improvement over the VSM using short collections. However these methods are quite intense computationally, so our proposal attempts to lessen this factor.

Accordingly, our idea is to capture the inherent semantic structure using the Bag of Concepts (BoC) as proposed by Sahlgren and Cöster [2], where the meaning of a term is considered as the sum of contexts in which it occurs. Both BoC and HRR are vector representations built with the aid of Random Indexing (RI), a vector space methodology also proposed by the same authors, which is an efficient, scalable and incremental method of building context vectors. In addition, Fishbein and Eliasmith [9] have used HRR together with BoC for text categorization. To the best of our knowledge, they have never been employed in IR, which is our main contribution.

The rest of this paper is organized as follows. In Section 2, we briefly review Random Indexing. Section 3 introduces Bag of Concepts, Holographic Reduced Representations, and how to use them to add context information to document

representations. Section 4 describes the experiments performed and the results obtained. Finally, Section 5 concludes the paper and gives directions for future work.

2 Random Indexing

Random Indexing (RI) is a vector space methodology, proposed by Sahlgren and Cöster [1,2], that accumulates context vectors for words based on co-occurrence data. This methodology assumes that semantically similar terms will occur in similar contexts. The technique is described as follows:

a) First, a unique random representation known as 'index vector' is assigned to each context (document or word). Index vectors are vectors with a small number of non-zero elements, which are either +1 or -1, with equal amounts of both. The dimensionality of the random index vectors is smaller than the number of contexts in a collection.
b) Second, index vectors are used to produce context vectors by scanning through the text. Every time a given word occurs in a context, the context's index vector is added to the word's context vector. Therefore, a word is represented by a context vector that contains traces of every context, i.e., word or document that the word has co-occurred with or in.

RI methodology is similar to latent semantic indexing (LSI) [14]. However, to reduce the co-occurrence matrix no dimension reduction technique such as singular value decomposition is needed, since the dimensionality t of the random index vectors is pre-established as a parameter (implicit dimension reduction). Consequently t does not change once it has been set; as a result, the dimensionality of context vectors will never change with the addition of new data: only the values of their vector entries change. This reduction makes RI more efficient, scalable and flexible (it can be used with different kinds of contexts) than LSI-like methods.

3 Text Representations

In this paper, RI is applied to represent documents using: 1.Bag of Concepts representation (BoC) for representing the meaning of a document as the addition of the meanings of its terms; 2. HRR, to encode syntactic structure which can directly capture relations between words (e.g., compound terms, subject-verb, verb-object, and spatial relations), as explained in the following sections.

3.1 Bag of Concepts

Bag of Concepts (BoC) is a recent representation scheme proposed by Sahlgren and Cöster in [2], which is based on the perception 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 from each term within the

document, and generating a document vector as the weighted sum of the term context vectors contained within that document. Therefore, we use RI to represent the meaning of a word as the sum of contexts (entire documents) in which it occurs. Illustrating this technique, suppose you have two documents: *D1*: Towards an Automata Theory of Brain, and *D2*: From Automata Theory to Brain Theory. Let us suppose that they have index vectors *ID1* and *ID2*, respectively: the context vector for "brain" will be the *ID1 + ID2*, because this word appears in both documents.

Once the context vectors have been built by RI, they are used to represent the document as BoC. For instance, after removing stop words and supposing *CV1*, *CV2* and *CV3* are the context vectors of *automata*, *theory* and *brain* respectively, then document *D2* will be represented as the weighted sum of these three context vectors.

3.2 Holographic Reduced Representation

HRRs are representations proposed by Plate [3], which permit the representation of structure using a circular convolution operator to bind terms, without increasing vector dimensionality. Circular convolution operator ($\otimes$) binds two vectors $\overrightarrow{x} = (x_0, x_1, ..., x_{n-1})$ and $\overrightarrow{y} = (y_0, y_1, ..., y_{n-1})$ to produce $\overrightarrow{z} = (z_0, z_1, ..., z_{n-1})$ where $\overrightarrow{z} = \overrightarrow{x} \otimes \overrightarrow{y}$ is defined as:

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

It can be thought of as a multiplication operator for vectors which has properties in common with scalar and matrix multiplication. As a result, it is not complicated to manipulate expressions containing additions, convolutions, and scalar multiplications [3].

HRR Document Representation. We adopt the HRR to build a text representation scheme in which certain syntactic structure can be captured and used to improve retrieval effectiveness. To define the HRR of a document, the following steps are done:

a) We first determine the index vectors for each word in a vocabulary by adopting the random indexing method, as described earlier.
b) For each syntactic relation in a document, the index vectors of the words involved in the relation are bound to their role identifier vectors (which are HRRs).
c) The tf×idf-weighted sum of the resulting HRRs is taken to obtain a single HRR vector representing the syntactic relation.
d) HRRs of the syntactic relations, multiplied by an attenuating factor α, are added in order to obtain a single HRR vector representing the document, which is then normalized.

For instance, suppose we want to represent the compound term $R = hot\ dog$. R will be represented using the index vectors $\overrightarrow{r_1}$ for *hot* and $\overrightarrow{r_2}$ for *dog*. Each term plays a different role in the structure (e.g. right word/left word, or subject/verb, or verb/object, etc.). Thus, to encode these roles two special vectors (HRRs) are needed: $\overrightarrow{role_1}$ (left word), $\overrightarrow{role_2}$ (right word). The relation vector is therefore:

$$\overrightarrow{R} = (\overrightarrow{role}_1 \otimes \overrightarrow{r}_1 + \overrightarrow{role}_2 \otimes \overrightarrow{r}_2) \tag{2}$$

Considering another example, suppose the spatial relation $R_2 = in\ Europe$. Therefore, R_2 will be represented using the index vectors $\overrightarrow{r_3}$ for *Europe*, where $\overrightarrow{r_3}$ will be joined to its location role, an HRR $\overrightarrow{role_3}$, which represents the relation *in*. Thus the *in Europe* vector will be:

$$\overrightarrow{R_2} = (\overrightarrow{role}_3 \otimes \overrightarrow{r}_3) \tag{3}$$

Given a document D, with compound terms L_1, L_2 among terms t_{x1} and t_{y1}; t_{x2} and t_{y2} respectively, its vector will be built as:

$$\overrightarrow{D} = \langle \alpha((\overrightarrow{role}_1 \otimes \overrightarrow{t}_{x1} + \overrightarrow{role}_2 \otimes \overrightarrow{t}_{y1}) + (\overrightarrow{role}_1 \otimes \overrightarrow{t}_{x2} + \overrightarrow{role}_2 \otimes \overrightarrow{t}_{y2})) \rangle \tag{4}$$

where $\overrightarrow{t}_{x1}$, $\overrightarrow{t}_{y1}$, $\overrightarrow{t}_{x2}$, $\overrightarrow{t}_{y2}$ are index vectors,$\langle\ \rangle$ 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.

4 Experiments

We conducted two experiments: the first was focused on verifying if HRRs would be appropriated for representing linguistic relations. Therefore we use CACM collection, which is no longer used for IR experiments because of its small size. However, we did need a testbed that could be easily handled and allow us to verify the soundness of HRRs. The second was done to prove whether some latent semantic context could be captured by BoC and that the HRRs worked in a larger collection. We decided to test the BoC representation in the Geographic Information Retrieval (GIR) environment using the CLEF collection for evaluating the Geo-CLEF task [10, 11]. Our decision was taken because GIR is a specialized IR branch, where the search of documents is based not only on conceptual keywords, but also on spatial information [12,13]. Going back to the query used as an example in the introduction "Floods in European cities", it is evident that GIR needs to go beyond lexical analysis and then capture or use some context information to satisfy the user's information needs, for example by matching "cities" with actual city names. Finally, our baseline was the results produced by the VSM using in all cases the cosine as a similarity function to compare documents and queries.

4.1 Evaluation Measure

The evaluation of the results was carried out with a metric that has demonstrated its 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:

$$AvgP = \frac{\sum_{r=1}^{m} P(r) \times rel(r)}{n} \tag{5}$$

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 and m is the number of relevant documents retrieved for q.

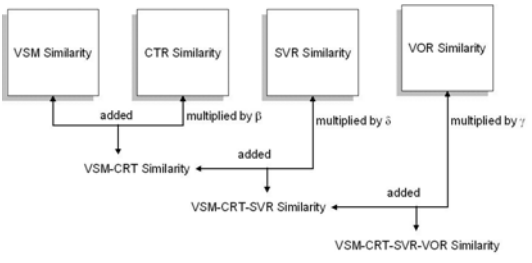

Fig. 1. Proccess to calculate the similarity values among documents and queries after having compared their HRRs representation

4.2 First Experiment

We selected a set of 10 queries from CACM, a collection with 3204 documents and 64 queries. The selected queries met the condition of having at least six relations to be represented and they had to include compound terms, subject-verb and verb-object relations.

Representation. To produce the HRRs for documents and queries, the syntactic relations were extracted from the selected queries and documents of CACM using Link Grammar [4] and MontyLingua 2.1 [5]. The stop words were eliminated and stemming was done for all the relations. If one of the elements of the compound terms or subject-verb relations had more than one word, only the last word was taken. The same criterion was applied for the verb in the verb-object relations; the object was built only with the first set of words extracted (direct object), and the last word taken, but only if the first word of the set was neither a preposition nor a connective. After extracting the syntactic relations their HRRs were built using vectors of 4096 dimensionality. This vector dimension was empirically determined after considering 1024, 2048, 4096 and 8192 dimensionalities. The precision increases as expected with the vector dimension; however after 4096, there was only a slight improvement compared with the doubled memory space needed to store the vectors. The generated HRRs were used

to represent the documents and queries as the sum of their compound terms (compound term representation (CTR)), afterwards as the sum of their subject-verb relations (subject- verb representation (SVR)) and finally as the sum of their verb-object relations (verb-object representation (VOR)). Thereafter, documents and queries represented as HRRs were compared using the cosine. As a result three similarity files were produced, one for each type of syntactic relation (CTR similarity, SVR similarity and VOR similarity files). Thus, to calculate the final similarity values, as is shown in Figure 1, the VSM was used to generate the initial similarity file. This initial similarity file was then added to the CTR similarity file, multiplied by a constant of less than one, and the documents were sorted again according to their new value. Afterwards, the SVR and VOR similarity files also multiplied by a constant of less than one were added and the documents once again sorted. Consequently, the last similarity between a document d given a query q is calculated with (6), where:β, δ, and γ are coefficients of less than 1.

$$similarity(q,d) = VSMsimilarity(q,d) + \beta\ CTRsimilarity(q,d) + \delta\ SVRsimilarity(q,d) + \gamma\ VORsimilarity(q,d) \quad (6)$$

Results. Table 1 shows the $AvgP$ for the selected queries. The value given to β in (6) was 1/16 and to δ and γ 1/32, determined by experiments where their values were varied from 1/2 to 1/64. The table also shows the MAP reached by the VSM and by adding to it the similarity files of the other relations. The average percentage of change after adding compound terms (CTR), subject-verb (SVR), and verb-object (OVR) similarity files to the VSM similarity file was 4.68%. Only one query had an unfavorable percentage of change in AvgP, and seven a favorable one. Two queries of the last category illustrate a change of 39% and 26%. We observed that when the relation extracted were real concepts, the representation works appropriately. For example, the compound terms extracted

Table 1. Queries with all the specified relations and at least 6 to be represented

Query ID	*VSM AvgP*	*VSM-CTR AvgP*	*VSM-CTR-SVR AvgP*	*VSM-CTR-SVR-VOR AvgP*	*% Diff*
1	0.1432	0.1551	0.1590	0.1814	26.68
4	0.0681	0.0777	0.0774	0.0777	14.10
7	0.2383	0.2388	0.2402	0.2414	1.30
8	0.1386	0.1379	0.1375	0.1932	39.39
9	0.1793	0.1893	0.1892	0.1978	10.32
31	1.0000	1.0000	1.0000	1.0000	0.00
32	0.4667	0.4667	0.4667	0.4667	0.00
37	0.1881	0.1866	0.1860	0.1862	-1.01
39	0.3006	0.3077	0.3091	0.3094	2.93
45	0.2327	0.2343	0.2356	0.2402	3.22
MAP	0.2956	0.2994	0.3001	0.3094	4.68

from query 1 were: *IBM computer, operating system, share system*, while the terms extracted from query 37 were: *data types, synchronization attempt, bit stream, passing system*. Analyzing the results for query 8, we observed that it has three relevant documents, six syntactic relations (2 of each type of relation); however, this query only shares a verb-object relation with the relevant document 2625, which in the VSM similarity file has the third position. When the syntactic relations were added, it was moved to the second position. Despite the small size of this experiment, we had favorable evidence to think that HRRs could be used to represent syntactic relations.

4.3 Second Experiment

This experiment had the aim to test the HRRs in a bigger collection and determine the effect of including context information using BoC in the GIR environment. We consider two phases in this experiment. The objective of the first was to retrieve as many relevant documents as possible for a given query (i.e. acceptable recall) and reduce the number of documents that have to be parsed to extract relations (since this process is costly). The function of the second, however, was to improve the final ranking (i.e. retrieval effectiveness) of the retrieved documents by applying BoC and HRR representations. We used Lemur[1] in the first phase, using the results produced by the VSM, configured in Lemur, as our baseline.

Data. We used the English CLEF collection for GIR, which is composed of 56,472 news articles taken from the British Glasgow Herald (1995) and 113,005 from the American LA Times (1994) to have a total of 169,477 articles.

```
<top>
 <num>GC008</num>
 <EN-title>Milk Consumption in europe</EN-title>
 <EN-des>Provide statistics or information concerning milk
 consumption in European countries </EN-desc>
 <EN-narr>Relevant documents must provide statistics or other information
 about milk consumption in Europe, or in single European nations. Reports on
 milk derivatives are not relevant</EN-narr>
</top>
```

Fig. 2. Topic GC008: Milk consumption in Europe

Queries. We worked with the queries from GeoCLEF 2005 to GeoCLEF 2008. A total of 25 queries were emitted per year to give in 2008 a set of 100 queries. Figure 2 shows the structure of each topic. The main query or title is between labels $\langle \text{EN} - \text{title} \rangle$ and $\langle /\text{EN} - \text{title} \rangle$. A brief description ($\langle \text{EN} - \text{desc} \rangle, \langle /\text{EN} - \text{desc} \rangle$) and a narrative ($\langle \text{EN} - \text{narr} \rangle, \langle /\text{EN} - \text{narr} \rangle$) are given too. These last two fields usually increase the requirement specificity of the original query. Participants at GeoCLEF were free to employ any or all of the three fields in their experiments. We took the title and description for our BoC

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

experiments, and for HRR we added the narrative statement in order to have more relations as representation. It should be mentioned that Lemur results are lower when the narrative is included than when only title and description are considered.

Representation. Lemur was used to process the 169,477 documents; first with the queries for 2005 and after with the queries for 2006-2008. 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.

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 corpus. We then used random indexing to produce context vectors for each sub-collection. The dimension of the context vectors was fixed at 4096, determined by experimentation as described for CACM. These context vectors were then tf×idf-weighted and added up for each document and query, to produce BoC representations.

On the other hand, the HRRs for spatial relations were generated by firstly tagging all sub-collections with the Named Entity Recognition of Stanford University[2]. 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 vectors of the identified location words, tf×idf-weighted and added to each document, as described in section 3.2 to generate spatial relation representations (SRR).

We present the results of three processes. The first is named VSM-BoC, which is created by combining the VSM similarity value with its corresponding value from BoC. The second, which was given the name of VSM-SRR, follows the same process as described above, but now with the similarity lists generated by VSM and SRR multiplied by a constant $\lambda = 1/16$ (as described in section 4.2). Finally, the three similarity lists (VSM, BoC, and SRR multiplied by λ) are combined to form VSM-BoC-SRR. The similarity values were calculated by the cosine function in all cases.

Results. Table 2 compares VSM results, with those produced after adding to it, BoC, SRR and BoC-SRR similarity lists. Notice how VSM-BoC increments MAP in a constant form, always above 5%. On the other hand, the increment with VSM-SRR is very slight, at lower than 1%. But when added together with BoC, the difference is raised by a further 2% to a total of 7% in 2008 and by a lesser degree in 2005 and 2007. Table 2 illustrates these favorable percentages to our proposals (numbers in bold).

Because the results for 2006 in terms of MAP were unfavorable, we tried to find the reason: Table 3 displays statistics for each sub-collection where vocabulary size; number of different documents per sub-collection; number of words,

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

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

	2005	*2006*	*2007*	*2008*
VSM	0.3191	0.2618	0.1612	0.2347
VSM-BoC	0.3380	0.2495	0.1695	0.2475
%Diff	**5.92**	-4.7	**5.15**	**5.45**
VSM-SRR	0.3193	0.2619	0.1623	0.2357
%Diff	**0.06**	**0.04**	**0.68**	**0.43**
VSM-BoC-SRR	0.3381	0.2495	0.1699	0.2512
%Diff	**5.95**	-4.7	**5.40**	**7.03**

Table 3. 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
in relations/query	0.72	2	2	7
relevant docs./query	41	15	29	31
total of relevant docs.	1028	378	650	747

Table 4. Queres with the highest number of relevant and relevant retrieved documents

Year	*Id.Qry*	*Rel*	*Rel. Ret.*	*VSM*	*VSM-BoC-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

number of *in* spatial relations, relevant document per query, and the total number of relevant documents per year are all shown. 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 *in* 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, which we believe represents the problem for data of this year, since 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 queries that have many relevant documents but may harm queries 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 4 shows queries that meet these conditions for each year. From this, it is clear that even for 2006 (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.

We found that the spatial relation representation (SRR) contributes towards improving 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. In addition, we thought that SRR representation helped to improve precision at high recall levels, meaning our initial baseline is low. Table 2 shows that without considering 2006 data (the number of relevant documents for 2006 is low) the highest improvement is for 2007, then for 2008 and finally for 2005, which has the highest initial baseline (VSM MAP).

5 Conclusions and Future Work

In this paper we have presented two document and query representations: BoC and HRR, aimed at improving IR effectiveness, which was measured by MAP. Our first and second experiments demonstrated that HRRs are suitable representations for word relations and, if the number of relations to be represented is enough, they may contribute to improving effectiveness mainly at high recall levels. Specifically for Geo-CLEF collection, we only considered one type of spatial relations: the *in* relation, we think if more types of relations *across, near, far*, etc.) are added as long as they are present in the queries; it could lead to a higher improvement. Also, the second experiment proved that by capturing context information with BoC, the IR effectiveness improves. In particular, the improvements for Geo-CLEF collection were at above 5% for the years 2005, 2007 and 2008. Our results were compared with one of the customary IR models: the VSM. In addition, we observed that a lack of relevant documents in a collection produces low effectiveness (2006 data), and that by combining similarity files, the IR effectiveness improved. Our results show an improvement of over 5% with BoC and of 7% combining BoC and HRR in 2008 data.

Notice that the representations we are proposing allow expressing additional knowledge of diverse kinds and from different sources, for instance: inter-document (by BoC indexing), thematic (compound terms), syntactic (relations subject-verb and verb-object), location (named entities) and spatial (*in* relation) knowledge. When compared, they are treated accordingly.

We will continue working with other collections that provide us with more specific contexts to be represented. We are thinking of representing concepts extracted from texts as HRRs, not only syntactic relations. This allows us to thoroughly explore the usefulness of the proposed representations to improve IR effectiveness.

References

1. Sahlgren, M.: An Introduction to Random Indexing. In: Methods and Applications of Semantic Indexing Workshop at the 7th Int. Conf. on Terminology and Knowledge Engineering (2005)
2. Sahlgren, M., Cöster, R.: Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. In: Procs. of the 20th Int. Conf. on Computational Linguistics, pp. 487–493 (2004)

3. Plate, T.A.: Holographic Reduced Representation: Distributed representation for cognitive structures. CSLI Publications, Stanford (2003)
4. Grinberg, D., Lafferty, J., Sleator, D.: A Robust Parsing Algorithm for Link Grammars, Carnegie Mellon University, Computer Science, Technical Report CMU-CS-95-125 (1995)
5. Liu, H.: MontyLingua: An end-to-end natural language processor with common sense (2004), http://web.media.mit.edu/~hugo/montylingua
6. Mitra, M., Buckley, C., Singhal, A., Cardie, C.: An Analysis of Statistical and Syntactic Phrases. In: Procs. of RIAO 1997, 5th Int. Conf., pp. 200–214 (1997)
7. 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)
8. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
9. Fishbein, J.M., Eliasmith, C.: Integrating structure and meaning: A new method for encoding structure for text classification. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 514–521. Springer, Heidelberg (2008)
10. Cross-lingual evaluation forum (May 2009), http://www.clef-campaign.org/
11. 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 Workshop, Denmark (2008)
12. Henrich, A., Luedecke, V.: Characteristics of Geographic Information needs. In: Procs. of Workshop on Geographic Information Retrieval, Lisbon, Portugal. ACM Press, New York (2007)
13. Andrade, L., Silva, M.J.: Relevance ranking for geographic IR. In: Procs. of 3rd Workshop on Geographic Information Retrieval, SIGIR 2006, Seattle, USA. ACM Press, New York (2006)
14. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the ASIS 41, 391–407 (1990)
15. Hofmann, T.: Probabilistic latent semantic indexing. In: Procs. of the 22st Annual International ACM SIGIR Conf. on R&D in Information Retrieval (SIGIR 1999), Berkeley, CA, pp. 50–57. ACM, New York (1999)
16. Wong, S.K.M., Ziarko, W., Raghavan, V.V., Wong, P.C.N.: On modeling of information retrieval concepts in vector spaces. ACM Trans. on Database Systems 12, 299–321 (1987)