# Concept Based Representations as Complement of Bag of Words in Information Retrieval*

Maya Carrillo[1,2], Aurelio López-López[1]

[1]Coordinación de Ciencias Computacionales, INAOE
Luis Enrique Erro 1, Santa Maria Tonantzintla, Puebla, México, C.P.72840
{cmaya,allopez}@inaoep.mx
[2]Facultad de Ciencias de la Computación, BUAP
Av. San Claudio y 14 Sur Ciudad Universitaria, 72570 Puebla, México

**Abstract.** Information Retrieval models, which do not represent texts merely as collections of the words they contain, but rather as collections of the concepts they contain through synonym sets or latent dimensions, are known as Bag-of-Concepts (BoC) representations. In this paper we use random indexing, which uses co-occurrence information among words to generate semantic context vectors and then represent the documents and queries as BoC. In addition, we use a novel representation, Holographic Reduced Representation, previously proposed in cognitive models, which can encode relations between words. We show that these representations can be successfully used in information retrieval, can associate terms, and when they are combined with the traditional vector space model, they improve effectiveness, in terms of mean average precision.

**Key words:** Information Retrieval, Concept Based Representation, Vector Model, Random Indexing, Holographic Reduced Representation.

## 1 Introduction

Information Retrieval (IR) is a discipline involved with the representation, storage, organization, and access to information items [1]. IR systems are designed to provide, in response to a user query, references to documents which could contain the information desired by the user. To compare documents and queries, these have to be represented in an appropriate way to be processed. Sometimes, features are extracted from documents without performing any advanced processing; this produces what is known as Bag of Words representation (BoW), where the document attributes are words or word stems.

Merely considering the words of a document has shown not to be enough for representing content. For instance, consider two documents using the same set of words, but one discussing the topics in a positive sense, while the other refers

* The first author was supported by Conacyt scholarships 208265, while the second author was partially supported by SNI, Mexico. We would also like to thank the CNRG at the University of Waterloo, Canada for the resources provided.

to the same topic but in a negative sense. Then, a document would be more appropriately represented if syntactic and semantic information was included. There is research that has sought to include such semantic aspects as Latent Semantic Analysis [9], which has included implicit context information in the indexing process. The information is obtained by grouping terms that have similar meaning using Singular Value Decomposition (SVD). However, this method is quite computationally expensive.

On the other hand, there have been other efforts to represent more precise concepts than only words. For instance, Mitra et al., Evans and Zhai [5],[6], among others have investigated the use of phrases as part of text representation since the early days of information retrieval. Their overall performance improvement has been only marginal, however. Recently Vilares, et al. in [22] have extracted binary dependencies (i.e. noun-modifier, subject-verb and verb-complement), their experiments have shown some improvement.

This paper, as an alternative to representing concepts, considers the use of Random Indexing (RI) to produce context vectors, which capture the implicit "semantics" of documents and queries without expensive reduction techniques as SVD. Thereafter, the context vectors are used to represent documents as Bag of Concepts (BoC) [7]. Besides this, we present the use of Holographic Reduced Representation (HRR) [2] to include syntactic relations between words. These techniques, to the best of our knowledge, have not been used in IR.

Nowadays, the traditional IR engines are able to retrieve the majority of relevant documents for most collections, but generally the ranking of the retrieved results leads to poor performance, in terms of precision. Therefore, we propose to use BoC and HRR to re-rank the results generated by the traditional vector space model (VSM) [4]. Our assumption was that the BoW could be enriched with information from a concept-based representation to improve its precision. Our results achieved with the English CLEF2005 collection for Adhoc track, have confirmed our hypothesis showing an improvement of over 16% in mean average precision (MAP).

The remainder of this paper is organized as follows: Section 2 provides a brief description of related work, particularly on including phrases in information retrieval. Section 3 presents Random Indexing, how it is used to create BoC representations, and related work. Section 4 introduces the concept of Holographic Reduced Representations (HRRs) and presents how to use them to represent documents. Section 5 explains the experimental setup and the results obtained. Finally, section 6 concludes the paper and gives some directions for further work.

## 2 Previous Work

There are several previous works, suggesting the use of phrases to index and retrieve documents. For instance, Evans & Zhai [5] present an approach to index noun phrases for IR. They describe a hybrid method to extract meaningful sub-compounds from complex noun phrases. Mitra et al. [6] present a study that compares the usefulness of phrase recognition by using linguistic and sta-

tistical methods. Croft et al. describe an approach where phrases identified in natural language queries are used to build structured queries for a probabilistic retrieval model [17]. Despite their many implementations, retrieval experiments with phrases have shown inconsistent results. Recently, a number of retrieval approaches have investigated the effectiveness of language modeling approach in modeling statistical phrases such as n-grams or proximity-based phrases, showing promising results [18], [19].

Our study differs from previous phrase-based approaches, in one aspect. We express phrases using a representation that reflects syntactic structure. This structure is then distributed across the document representation, rather than taking the phrases as new terms extending the space dimension.

## 3 Random Indexing

Random Indexing (RI) is a vector space methodology that accumulates context vectors for words based on co-occurrence data. The technique can be described via the following two steps:

1. A unique random representation known as index vector is assigned to each context (i.e. document). Index vectors are binary vectors 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 512- dimensional vector space, they will have ten +1s, ten -1s and 492 0s. Index vectors serve as indices or labels for contexts.
2. Index vectors are used to produce context vectors by scanning through the text and every time a given word occurs in a context; the index vector of the context is added to the context vector of the word [8].

The above steps can be exemplified as follows: Let's suppose we have an eight-dimensional space, two non-zero elements and two documents *D1: Regular Right Part Grammars and their Parsers*, whose index vector is [0,1,0,0,-1,0,0,0] and *D2: Boolean Matrix Methods for the Detection of Simple Precedence Grammars* with [0,1,0,0,0,-1,0,0] index vector. Then, the context vector for *Grammars* is the addition of both index vectors, since such word appears in them, producing the vector [0,2,0,0,-1,-1,0,0]. Word context vectors generated through this process are used to build document vectors as Bag of Concepts (BoC). Thus, a document vector is the sum of the context vectors of its words.

Random Indexing, in the same manner as Latent Semantic Analysis (LSA), attempts to capture implicit "semantic" relations, but RI has additional advantages as: a) it does not have to use reduction techniques like Singular Value Decomposition (SVD) to reduce the space dimensionality; b) It is an incremental method, which means that we do not have to process all the data before we can start using the context vectors [8].

There are several works that have validated the use of RI in text processing tasks: for example, Kanerva et al. in [20] 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. Sahlgren & Karlgren [21] demonstrated that Random Indexing can be applied to parallel texts for automatic bilingual lexicon acquisition. Sahlgren & Cöster [7] used Random Indexing to carry out text categorization.

## 4 A Representation from Cognitive Science

Distributed representation allows to integrate connectionism and cognitivism, where mental representations (symbols) are specified by distributed patterns of neural activities, while is possible to introduce formal algebraic operations on these distributed patterns, to mathematically model cognitive operations [10]. This approach, whose principles in cognitive science were formulated at the end of the 1960's [11], [12], [13], [14], [15], was culminated by the Holographic Reduced Representation (HRR) formulated by T. Plate [2]. The HRR, is a method for representing compositional structure in analogical reasoning. 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. 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 give $\mathbf{z} = (z_0, z_1, \ldots, z_{n-1})$ where $\mathbf{z} = \mathbf{x} \otimes \mathbf{y}$ is defined as:

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

Illustrating the use of HRRs to represent simple predicates, we suppose an instantiated frame, which is the superposition of the frame name and the role-filler bindings (roles convolved with their respective fillers). For example, in a very simplified frame for *eating*, the vector for the frame name is *eating* and the vectors for the roles are $eat_{agent}$ and $eat_{object}$. This frame can be instantiated with the fillers *Peter* and *the_fish* to produce "Peter eats the fish" as follows:

$$s_1 = eat + eat_{agent} \otimes peter + eat_{object} \otimes the\_fish \tag{2}$$

Accordingly, we adopt HRRs to build a text representation scheme in which two-word terms could be represented. 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) Syntactically tag the documents using a natural language processing tool; c) Bind the *tf.idf*-weighted index vector of each word to its role. This "side" role is an HRR which serves to distinguish the right side from the left side of the two-word term. d) Add the resulting HRRs (with the two-word nouns encoded) to obtain a single HRR vector; e) Multiply the resulting HRR by an attenuating factor $\alpha$; f) Normalize the HRR produced to obtain the vector which represents the document. In [23] and [24] there are some examples of this representation.

Fishbein, and Eliasmith have used the HRRs together with Random Indexing for text classification, having BoC as their baseline [3]. We have reported other result using this representation in [23] with smaller collections and adding the

HRRs to a special representation named Index Vector Representation. In [24] we reported results using GEOCLEF collection from 2005 to 2008 and using HRRs to code prepositional location phrases. It is important to mention that up to now, we are not aware of other research that uses RI together with HRRs.

## 5 Experimentation

The proposed document representation was generated for the English document collection used in CLEF 2005 for the Adhoc track. This is composed of 56,472 news articles taken from the Glasgow Herald (British) 1995 and 113,005 from LA Times (American) 1994 for a total of 169,477. The queries used were 50 from number 251 to number 300. They contain title, a brief description, and a narrative. The experiments, described below, were done taking only the title and description fields.

The experiments were divided into two stages. The aim of the first was to obtain as many relevant documents as possible; this was carried out by Lemur[1], an open source system designed to facilitate research in information retrieval. The VSM using *tf.idf* weighting scheme and a cosine measure to determine vector similarity were configured in Lemur and the results generated were used as a baseline.

In the second stage, the list of documents generated by Lemur, with elements of three attributes *query-document-similarity* was re-ranked. In the re-ranking process the Lemur similarity list was combined (adding the similarity values) with two additional similarity lists. These lists were multiplied by 1/4 before the merging process to give more importance to the similarities obtained by Lemur. Of these additional lists, the former one was generated when documents and queries were represented as BoC and compared; the latter list was produced when the same documents and queries were represented as HRRs and compared. These lists were built taking only the first 1000 documents for each query to diminish the processing time. However, it should be mentioned that, on average, the time to build the BoC representation for a query and its associated documents was 2.9 min., and 4.6 min. to build the HRRs. In contrast, the time for comparing the query with the one thousand documents was only 0.124 seconds in both representations.

The HRR document representations were built as specified in section 4 where $\alpha$ was equal to 1/6, taking only two-word terms. MontyLingua [16] was used for parsing the documents and queries and extracting the two-word terms. Examples of these composed terms are: *civil war, serial killer, Russian force, military action*. Both BoCs and HRRs representations were weighted and compared using the same schemes used for the VSM. We carried out several previous experiments intended to assess the effects of dimensionality, limited vocabulary, vector density and context definition. In our experiments the vector dimensionality was 4096 and the density 20 non-zero elements. We removed stop words and used

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

stemming in the same way as the traditional VSM. These parameters, which were determined after our preliminary experiments, produced suitable data for our proposal. However, defining their right value is an open research topic.

### 5.1 Evaluation

The results after re-ranking the documents were evaluated with two metrics: Mean Average Precision (MAP), which is defined as the average of all the *AvgP* obtained for each query. *AvgP* is defined as:

$$AvgP = \sum_{k=1}^{m} P(r) \text{ x } rel(r)/n \tag{3}$$

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 $q$.

The second metric is R-Precision (*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.

### 5.2 Results

Table 1 compares Lemur results, with those produced after adding to it, BoC and HRR similarity lists. This process thereby produces Lemur + BoC and Lemur + BoC + HRR lists. Notice how BoC improves MAP and R-Prec by always being above 15%. Although BoC gave a high improvement in MAP, HRR additionally improved it by 1.3%. In contrast, the R-Prec decreased 2% when the HRRs were added. Consequently the HRRs, as was expected, help to emphasize the representation of specific concepts as observed for the MAP increment, but they cause a loss of generality reflected in the R-Prec decrement. The results obtained by both Lemur + BoC and Lemur + BoC + HRR, performing a paired t-student test, were found to be statistically significant in a 99% confidence interval in terms of MAP.

Table 2 shows the same comparison, but now in terms of precision at the indicated number of documents. The proposed representations increased the precision in all cases, but it should be noted that the difference is higher at low recall levels. It is difficult to compare our results with those mentioned in related work, because the authors worked with different collections, metrics and environments. However, as mentioned in [25] "Sparck Jones has suggested that a difference in the scores between two runs that is greater than 5% is noticeable, and a difference that is greater than 10% is material".

## 6 Conclusions and Future Works

In this article, we have presented a proposal for representing documents and queries that according to the experiments, has shown itself to be feasible and

**Table 1.** MAP and R-Prec results for ADHOC CLEF 2005 English document collection

| Metric | Lemur | Lemur+BoC | %Diff | Lemur+BoC+HRR | %Diff |
|---|---|---|---|---|---|
| MAP | 0.2903 | 0.3392 | **16.84** | 0.3430 | **18.15** |
| R-Prec | 0.3103 | 0.3579 | **15.34** | 0.3514 | **13.24** |

**Table 2.** Precision at 5, 10, 15, 20, 30 and 100 documents

| RI System | P@5 | P@10 | P@15 | P@20 | P@30 | P@100 |
|---|---|---|---|---|---|---|
| Lemur | 0.4600 | 0.4160 | 0.3960 | 0.3610 | 0.3120 | 0.1854 |
| Lemur + BoC | 0.5280 | 0.4640 | 0.4293 | 0.3980 | 0.3540 | 0.1996 |
| % Difference | **14.78** | **11.54** | **8.41** | **10.25** | **13.46** | **7.66** |
| Lemur + BoC + HRR | 0.5360 | 0.4740 | 0.4387 | 0.3990 | 0.3527 | 0.1998 |
| % Difference | **16.52** | **13.94** | **10.78** | **10.53** | **13.04** | **7.77** |

able to capture "semantic relations" and encode two-word terms. BoC improved the initial ranker. HRRs produced a slight gain in precision. However, they have the potential to encode other relations between concepts (e.g. syntactical relations such as subject-verb, or additional information as identifying named entities).We think if more types of relations are considered it could lead to a higher improvement. Based on our results, it seems reasonable to conjecture that these new representations when combined with the VSM to re-rank the documents, increase precision. Our results showed an improvement above 16% in MAP. Additional study and experimentation will be necessary to quantify the usefulness of the proposed representations. Then, we will continue working with other collections that provide us with more specific contexts to be represented. This allows us to thoroughly explore the usefulness of the proposed representations to improve IR effectiveness.

## References


1. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, (1999).
2. Plate, T.A.: Holographic Reduced Representation: Distributed representation for cognitive structures. CSLI Publications, (2003).
3. 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).
4. 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).
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. 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)

7. 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).
8. 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).
9. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis, Journal of the ASIS, 41, pp.391-407. (1990)
10. Kvasnicka, Vladimir: Holographic Reduced Representation in Artificial Intelligence and Cognitive Science In: Neural Network World, vol. 14; part 6, pp. 521-532, (2004)
11. Gabor, D.: Holographic model for temporal recall. In: Nature, 217, pp.1288-1289, (1968).
12. Metcalfe Eich, J.: Levels of processing, encoding specificity, elaboration, and charm, Psychological Review, 92, pp. 1-38, (1985).
13. Murdock, B. B.: A theory for the storage and retrieval of item and associative information. In: Psychological Review, pp. 316-338, (1982).
14. Slack, J. N.: The role of distributed memory in natural language processing. In: Advances in Artificial Intelligence: Procs. of the Sixth European Conference on Artificial Intelligence, ECAI-84, Elsevier Science Publishers, New York, 1984.
15. Willshaw, D. J., Buneman, O. P.,Longuet-Higgins, H. C.: Non-holographic associative memory. Nature, 222 (1969), pp. 960-962
16. Liu, Hugo: MontyLingua: An end-to-end natural language processor with common sense. web.media.mit.edu/ hugo/montylingua (2004).
17. Croft W.B., Turtle H.R., Lewis D.D.: The use of phrases and structured queries in information retrieval. In: Procs. of the 14th Annual International ACM/SIGIR Conference(1991) pp. 32-45.
18. Metzler D., Croft W. B.: A markov random field model for term dependencies. In: Procs. of SIGIR '05, (2005) pp. 472-479.
19. Gao J., Nie J., Wu G., and Cao G.: Dependence language model for information retrieval. In Procs. of SIGIR '04, (2004) pp. 170-177.
20. Kanerva P., Kristoferson J., Anders Holst A. Random indexing of text samples for latent semantic analysis. In Procs. of the 22nd Annual Conf. of the Cognitive Sc. Society, New Jersey: Erlbaum, pp. 103-106 (2000)
21. 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).
22. Vilares J., Gómez-Rodríguez C., Alonso M.A., Managing Syntactic Variation in Text Retrieval, In: Procs. of the ACM Symposium on Document Engineering. Bristol, United Kingdom, ACM Press, New York, USA, pp. 162-164. (2005).
23. Carrillo M., Eliasmith C., López-López A., Combining Text Vector Representations for Information Retrieval, In: Procs. Text, Speech and Dialogue, 12th International Conference, TSD 2009, Pilsen, Czech Republic, LNCS, Springer, pp. 24-31 (2009).
24. Carrillo M.,Villatoro-Tello E, López-López A., Eliasmith C., Montes-y-Gómez M., Villaseñor-Pineda L.: Representing Context Information for Document Retrieval. In: Procs. Flexible Query Answering Systems, 8th International Conference, Roskilde, Denmark, pp.239-250(2009).
25. Buckley, C. and Voorhees, E. M.: Evaluating Evaluation Measure Stability, ACM SIGIR 2000 Procs., pp. 33-40, (2000).