Representaciones Vectoriales Orientadas a Conceptos y Estructura para Recuperación de Información

por

Maya Carrillo Ruiz

Tesis sometida como requisito parcial
para obtener el grado de

DOCTOR EN CIENCIAS EN EL ÁREA DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Noviembre 2010
Tonantzintla, Puebla

Supervisada por:

Dr. Aurelio López López, INAOE

Agradecimientos

Al doctor Aurelio López López, mi asesor de tesis, por su invaluable guía e ideas a lo largo de estos años, pero principalmente por soportar conmigo los momentos de crisis y enseñarme que en el trabajo de investigación la paciencia es indispensable. Muchas gracias.

Chris Eliasmith for his ideas and encouragement, as well as showing me a wonderful way to create science; where work is base on thought, discussion, a great deal of enthusiasm and friendship, to make knowledge a universal right.

A los doctores Manuel Montes y Gómez y Luis Villaseñor Pineda por el tiempo dedicado a la revisión de este trabajo, por el interés mostrado en el mismo, aportándome un sinfín de ideas interesantes a lo largo del tiempo que duró esta investigación. Muchas gracias.

A los doctores Jesús Ariel Carrasco Ochoa y Saúl E. Pomares Hernández por sus acertados comentarios para mejorar este trabajo.

Al Instituto Nacional de Astrofísica, Óptica y Electrónica por las facilidades y los recursos brindados durante estos años para la conclusión de mi investigación.

Al Consejo Nacional de Ciencia y Tecnología por el apoyo proporcionado mediante la beca número 208265.

A mis compañeros del Laboratorio de Tecnologías del Lenguaje y del Departamento de Ciencias Computacionales por compartir conmigo sus conocimientos y experiencias para enriquecer mí investigación.

Ningún trabajo es posible sin el apoyo de familiares y amigos. Agradezco a Hortensia y Manuel, soy verdaderamente afortunada por tenerlos como padres, apoyándome en todas mis decisiones y soportando mis arrebatos sin el menor reclamo, para luego recibirme con el corazón y los brazos abiertos en un ambiente lleno de amor, confianza y felicidad. Muchas gracias.

A mis hermanas Miriam, Rosalba y Hortensia, por sus innumerables acciones de ayuda y comprensión y por hacer de los momentos compartidos algo muy divertido.

A Hernando, mi tío, por sus valiosas correcciones a la redacción de este trabajo y por estar siempre ahí como un gran roble al que uno sabe que puede asirse en los momentos difíciles.

A mis queridos sobrinos Mau, Sebas, Rayhan, Rorris y Nao, porque de sus risas

y juegos cada fin de semana obtenía el aliciente necesario para continuar.

A Gareth por su amistad y por permitirme interrumpir la tranquilidad de su hogar para resolver mis dudas y ayudarme en la redacción de éste y los trabajos derivados.

A Blanca y Marco Aurelio por llenar de risas y entusiasmo los momentos compartidos. Gracias queridos amigos.

A la maestra Carmen por su amistad y palabras de aliento, pero sobre todo, porque en cada una de sus clases obtenía la dosis de alegría necesaria para hacer menos difícil el camino.

A Juan Carlos por ser tan divertido y enseñarme a mirar la vida desde una perspectiva llena de gratitud. Muchas gracias.

Al maestro Isaí, no sólo por ayudarme a romper mis limitantes físicas, sino por estar siempre ahí, escuchando las palabras o en el silencio, para darme el justo remedio a cualquiera de mis males. Muchas gracias.

Resumen

El lenguaje es una de las habilidades más impresionantes de los seres humanos. Las áreas especializadas del cerebro, como la de Broca y la de Wernicke, sugieren que genéticamente tenemos elementos neurológicos para el desarrollo del lenguaje. Así, a lo largo de la historia de la humanidad, el conocimiento se ha comunicado, guardado y manejado en forma de lenguaje natural (griego, latín, inglés, español, etc.). En la época actual el conocimiento sigue atesorándose en documentos, libros, revistas, aunque ahora se guarda también en forma digital. Este factor ha convertido a la computadora en una herramienta para acceder de manera eficiente a la información. Como seres humanos podemos interpretar el conocimiento almacenado en dichos documentos y hacer inferencias lógicas sobre su contenido. Sin embargo, para las computadoras dicha información es sólo una secuencia de caracteres y nada más. La recuperación de información es una disciplina cuyo objetivo es desarrollar métodos para suministrar automáticamente información relevante a solicitud de los usuarios. Las técnicas clásicas de recuperación de información representan los documentos como listas de palabras sin ningún orden ni relación. Esta representación ignora la estructura gramatical de los textos y entonces, elimina cualquier posibilidad de entender su contenido. La presente investigación propone una representación de documentos que considera aspectos léxicos, sintácticos y "semánticos", cada aspecto lingüístico se maneja en un espacio vectorial independiente. Los aspectos léxicos se capturan con la representación tradicional de bolsa de palabras; los sintácticos, con una representación tomada de la ciencia cognitiva propuesta por T.A. Plate, llamada representación holográfica reducida; y finalmente, los aspectos "semánticos" con la representación de bolsa de conceptos propuesta por Sahlgren y Cöster. Para crear estas dos últimas representaciones, se utiliza una metodología conocida como indexación aleatoria propuesta por Kanerva et al., la cual permite reducir el espacio vectorial producido por la aproximación de bolsa de palabras. Esta investigación, hasta donde se tiene conocimiento, es la propuesta inicial de un modelo de recuperación de información que integra las representaciones mencionadas, sin incrementar la dimensión del espacio vectorial. Los resultados experimentales prueban que la integración de estas tres representaciones mejora la media de la precisión promedio (MAP) de la recuperación de información, con respecto a la representación de bolsa de palabras.

Abstract

Language is one of the most impressive human abilities. Broca and Wernicke, specialized areas of the brain, suggest to us we genetically have neurological components for language development. Thus, throughout human history, most knowledge has been communicated, stored and managed in the form of natural language (Greek, Latin, English, Spanish, etc.). At the present time, knowledge continues to be treasured through documents, books and journals, although it is now also stored in digital form. This factor has turned the computer into an efficient tool for accessing information. As human beings, we can not only interpret the knowledge stored in those documents, but also perform logical inferences about content. However for computers, that information is just a sequence of characters and nothing else. Information retrieval is a discipline, which aims to develop methods for automatically providing relevant information to queries submitted by users. The classical information retrieval techniques represent documents as lists of words without any order or relation among them. This representation ignores grammatical structure of texts and then eliminates any possibility of understanding of their content. This research proposes a text representation, which considers lexical, syntactic and "semantic" aspects of documents; each linguistic aspect is handled in a separate vector space. Firstly, lexical representation is captured with traditional bag of words representation; secondly, syntactic, with a representation from Cognitive Science proposed by T.A. Plate, named holographic reduced representation, and finally, the "semantic" aspect with the bag of concepts representation proposed by Sahlgren and Cöster. The latter two representations need a methodology known as random indexing to be defined. Random indexing, proposed by Kanerva et al., reduces the vector space produced by the bag of words approach. This research, to the best of our knowledge, is the initial proposal for an information retrieval model, which integrates the mentioned representations without increasing the dimension of the vector space. The experimental results in several collections showed that the integration of these three representations can improve the information retrieval mean average precision (MAP), with respect to that produced by the bag of words representation.

Contenido

Índice de Figuras

Índice de Tablas

1

Introducción

1.1 Problema y motivación

La recuperación de información (IR [1]) es una disciplina relacionada con la organización, almacenamiento, representación y recuperación de elementos de información (Baeza-Yates y Ribeiro-Neto 1999). Los sistemas de IR están diseñados para proporcionar, en respuesta a una consulta emitida por un usuario, referencias a los documentos que probablemente contienen la información deseada por éste. Si un documento a juicio del usuario se relaciona con su consulta entonces se dice que el documento es relevante. Tanto los documentos como la consulta están expresados en lenguaje natural. Por lo tanto, la relación entre su contenido semántico o entre los diferentes elementos lingüísticos que los componen, no pueden especificarse de manera sencilla. Las principales razones son la flexibilidad y ambigüedad lingüística. La primera se refiere a que una misma idea puede ser expresada con diferentes palabras, mientras que la segunda a la posibilidad de que el texto se interprete de diferentes maneras. Ambas afectan el proceso de recuperación de información. La flexibilidad lingüística provoca que ciertos documentos relevantes no se recuperen, porque los términos en la consulta no son los mismos que se utilizaron en los documentos. La ambigüedad por otra parte produce ruido porque se recuperan documentos que tienen los términos presentes en la consulta pero en ellos el significado es diferente.

[1] Del inglés *Information Retrieval*. Todas las siglas en el documento provienen de los nombres en inglés.

En consecuencia, la comparación directa entre la consulta y los documentos expresados en lenguaje natural, es poco factible dadas las características propias del lenguaje. Por lo tanto, la tarea requiere reducir el vocabulario empleado y normalizar las palabras tanto en los documentos como en las consultas, para entonces realizar la comparación empleando representaciones de documentos y consultas. Las decisiones de:

(a) Qué atributos incluir: palabras, frases (motivadas lingüística o estadísticamente) o conceptos semánticos

(b) Qué relaciones considerar entre los mismos: relaciones sintagmáticas o paradigmáticas[2]

(c) De qué partes tomarlos: derivados de los documentos/consultas o asignados de una estructura de lenguaje controlada

son fundamentales para la IR. Las técnicas clásicas de IR utilizan palabras, derivadas de documentos e ignoran las relaciones existentes entre las mismas. Estas técnicas descansan en el siguiente supuesto: si un documento y una consulta (query) tienen una palabra en común, entonces el documento se refiere a la consulta, si el número de palabras (bolsa de palabras (BoW)) en común aumenta, entonces la relación es mayor. Bajo este acercamiento, la IR trata de determinar cuánto se parece la bolsa de palabras de la consulta a la bolsa de palabras de cada documento.

El lenguaje es más que una colección de palabras, se emplea para hablar acerca de entidades, conceptos y relaciones que deben ser expresadas en formas apropiadas. Por ejemplo, el orden de las palabras es importante, no es lo mismo *venetian blind* que *blind venetian*. Las palabras son combinadas en frases y estructuras mayores que se mantienen unidas mediante relaciones tales como: dependencias estructurales, correferencias, roles semánticos, dependencia del discurso, intenciones y demás.

Por lo tanto, representar los documentos sólo como listas de palabras ha mostrado ser insuficiente para representar el contenido textual. Por ejemplo, dos documentos pueden utilizar el mismo conjunto de palabras, pero uno discute un tópico de manera positiva y el otro se refiere a dicho tópico en sentido negativo. En consecuencia, los documentos se caracterizarían de manera más adecuada si se incluyera información sintáctica y semántica en su representación.

Existen investigaciones que han buscando incluir aspectos semánticos en la representación textual, como la indexación semántica latente (LSI) (Deerwester et al. 1990),

[2]Las relaciones sintagmáticas ocurren entre unidades léxicas consecutivas y las paradigmáticas entre unidades léxicas como parte de la estructura léxica establecida por el lenguaje.

que agrupa los términos que tienen significado similar utilizando el método de descomposición en valores singulares (SVD). Sin embargo, SVD es costoso tanto en tiempo de procesamiento como en la memoria de almacenamiento requerida.

Por otra parte hay más esfuerzos para representar conceptos más precisos que únicamente palabras. Por ejemplo Mitra et al (Mitra et al. 1997), Evans y Zhai (Evans y Zhai 1996), entre otros, han investigado la utilización de frases como parte de la representación textual desde los inicios de la IR. Sin embrago, las mejoras que generalmente han obtenido son marginales. Recientemente Vilares, et al. (Vilares et al. 2004) han extraído dependencias binarias (i.e. sustantivo-modificador, sujeto-verbo, verbo-complemento) consiguiendo cierta mejora.

En la presente investigación, como alternativa para representar conceptos, se considera la utilización de la indexación aleatoria (RI), una metodología de espacio vectorial propuesta por Kanerva et al.(Kanerva et al. 2000), para producir vectores de contexto, los cuales capturan la "semántica" implícita de los documentos y consultas sin emplear técnicas de reducción costosas como SVD. Los vectores de contexto generados con RI se utilizan para representar documentos como bolsa de conceptos (BoC). Sahlgren y Cöster en (Sahlgren y Cöster 2004) proponen la utilización de BoC, esta representación se basa en la intuición de que el significado de un documento puede representarse como la unión del significado de los términos que lo componen.

Por otra parte, además de la utilización de BoC, empleamos la representación holográfica reducida (HRR), propuesta por Plate (Plate 2003), para codificar relaciones sintácticas entre palabras. Es importante señalar que el propósito de esta investigación es explorar la utilización de estas representaciones novedosas en IR y no necesariamente obtener los mejores resultados posibles en una colección particular de datos.

En IR debe empezarse a pensar más allá de la aproximación de bolsa de palabras, ya que con precisiones típicamente por debajo del 50 %, hay mucho margen para mejorar (Lease 2007).

1.2 Objetivos

Como lo enuncia Smeaton en (Smeaton 1995), las investigaciones que emplean procesamiento de lenguaje natural (NLP) en IR, con la intención de mejorar su efectividad, han reemplazado la bolsa de palabras por una bolsa de palabras mayor, en la que incluyen términos compuestos como nuevos términos. Ante esta observación, en el presente trabajo se busca integrar los términos compuestos a las representaciones vectoriales sin incrementar el número de términos. A continuación los objetivos del presente trabajo:

1.2.1 Objetivo general

Establecer una representación de contenido de textos para la cual se localicen, extraigan e integren relaciones entre términos, para mejorar la expresividad en tareas de tratamiento de información, en particular en IR, con respecto a la representación vectorial de textos tradicional.

1.2.2 Objetivos particulares

(a) Determinar una representación adecuada para documentos que permita capturar asociaciones entre términos e integrarlas a la representación.

(b) Establecer cómo asociar los términos para que constituyan unidades de comparación.

(c) Validar la representación en recuperación de información.

(d) Determinar que la incorporación de aspectos semánticos latentes contribuye a mejorar la efectividad de la IR en colecciones de tamaño controlado.

Con respecto al último objetivo Stein en (Stein 2007) clasifica a las aproximaciones de recuperación de información, que reducen el espacio vectorial original a un espacio vectorial semántico, como situaciones de recuperación cerradas o semicerradas, y por lo tanto no escalables a colecciones grandes, se comprobará si esto aplica a la indexación aleatoria.

1.3 Organización del Documento

El resto del documento está organizado de la siguiente manera:

- *Capítulo 2*: presenta los conceptos básicos utilizados en el resto de la tesis, incluye principalmente una explicación del modelo vectorial, la metodología de indexación aleatoria, la representación de bolsa de conceptos y la representación holográfica reducida.

- *Capítulo 3*: muestra el trabajo relacionado con la investigación propuesta. Aquí se describen algunos métodos empleados para incorporar relaciones textuales en el proceso de recuperación de información. También se describen trabajos relacionados con la utilización en procesamiento de texto de: la indexación

aleatoria, la representación de bolsa de conceptos y la representación holográfica reducida.

- *Capítulo 4*: explica el método propuesto detallando la integración de la indexación aleatoria, la representación de bolsa de conceptos y la representación holográfica reducida, para representar documentos y recuperar información.
- *Capítulo 5*: presenta los experimentos realizados para evaluar las representaciones propuestas y los resultados obtenidos.
- *Capítulo 6*: resume las aportaciones de la presente investigación, las cuales se ubican en el área de recuperación de información. Al final del capítulo se describe el trabajo futuro y las líneas de investigación abiertas.

2

Conceptos generales

Este capítulo presenta los métodos y conceptos generales que se utilizarán a lo largo de esta tesis. Los temas que se abordan son: i) Modelo vectorial, ii) Representación de Bolsa de Conceptos y iii) Representaciones Distribuidas

2.1 Introducción

En la actualidad vivimos rodeados de un sinfín de información textual en forma de periódicos, revistas, libros, páginas en Internet, etc. Así que la mayor cantidad de información manipulada por las computadoras está en lenguaje escrito. Quizás, entonces, el mayor desafío del procesamiento de lenguaje natural (NLP) es la utilización de éste sin ninguna restricción. Venciendo dicho reto se habilitaría de manera transparente el intercambio de información entre nosotros los humanos y las computadoras.

Ahora estamos interesados en que las computadoras sean capaces no sólo de realizar tareas mecánicas, sino también de auxiliarnos en tareas intelectuales como revisar textos, ejecutar instrucciones en lenguaje natural, comprender los textos para darnos respuestas razonables basadas en el entendimiento de los mismos y de esta manera dejarnos únicamente la tarea de decidir. Sin embargo, los principales avances en el área de NLP se han apoyado en el procesamiento de los niveles bajos del lenguaje, i.e. análisis morfológico, léxico y sintáctico, con poco entendimiento del nivel semántico. En este capítulo, a partir de uno de los modelos más sencillos y robustos como es el modelo vectorial, basado únicamente en análisis morfológico y léxico, se pre-

senta una serie de métodos con el objetivo de lograr una representación textual que capture "semántica"implícita y estructura compositiva y jerárquica de los conceptos expresados en los documentos.

2.2 Modelo Vectorial

La IR incluye dos actividades principales: indexar y buscar. La primera se refiere a representar el contenido de los documentos y la solicitud de información que expresa el usuario. La segunda, a la forma de examinar la representación de los documentos con respecto a la petición de información, para proporcionar como respuesta aquellos que se consideren de mayor relevancia.

Las características inherentes al lenguaje, previamente mencionadas, ocasionan que las operaciones de indexar y de buscar nunca recuperen información de manera perfecta, como puede ser el caso de los manejadores de bases de datos. Por lo tanto, ha sido necesario establecer métodos cuantitativos (funciones de relevancia) para evaluar de forma aproximada la similitud entre la consulta y los documentos, y entonces determinar cuáles recuperar.

En función de lo anterior, un modelo de IR se define con la representación que se da a los documentos y a las consultas (objetivo de esta investigación), y la función de relevancia que se emplea para compararlos.

Cuando se recupera un grupo de documentos se verá que algunos de ellos son irrelevantes y que algunos relevantes no se han recuperado. El éxito de la recuperación se establece mediante dos métricas: la precisión, que es la razón de los documentos relevantes recuperados entre el total de documentos hallados y el recuerdo, la razón del número de documentos relevantes recuperados entre el total de documentos relevantes existentes en la colección.

La recuperación de información de documentos escritos en lenguaje natural, requiere de un proceso para trasladar los textos a una representación adecuada para su manipulación. Hay diferentes representaciones: las basadas en probabilidad, álgebra lineal o teoría de conjuntos, por mencionar algunas. En este trabajo utilizamos representaciones basadas en álgebra lineal, de las cuales, quizás la más conocida por su simplicidad y resultados aceptables es el modelo vectorial, propuesto por Salton et al. (Salton et al. 1975).

La idea central del modelo vectorial es que el contenido de los documentos puede esbozarse con los términos que aparecen en ellos. Los documentos y consultas son representados como vectores, cuyas entradas están determinadas por la frecuencia de términos presentes en ellos, Figura 2.1. De esta manera un documento d se representará como un vector de n entradas, una para cada término distinto en la colección,

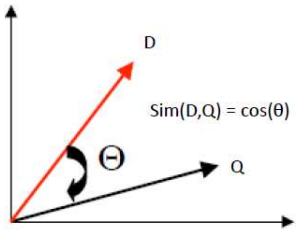

Figura 2.1: El documento D y la consulta Q son representados como vectores

los t_i representan la frecuencia dentro del documento de cada término:

$$\mathbf{d} = (t_1, t_2, \ldots, t_n) \tag{2.1}$$

De forma análoga el vector de la consulta q se representará como:

$$\mathbf{q} = (q_1, q_2, \ldots, q_n) \tag{2.2}$$

donde los q_i representan la frecuencia de los términos dentro de la consulta q.

Dado un conjunto de términos, puede notarse que no todos ellos tienen la misma utilidad para describir el contenido de un documento. Dicha utilidad se intenta capturar con diferentes esquemas de ponderación. En un esquema de ponderación se distinguen tres componentes: la ponderación local, es decir las veces que un término aparece en el documento; la global, las veces que un término aparece en la colección (conjunto de documentos), y el factor de normalización que compensa la discrepancia en la longitud de los documentos, esto puede expresarse como:

$$L_{ij} G_i N_j \tag{2.3}$$

donde L_{ij} es el peso del término i en el documento j, G_i es el peso global del término i y N_j es el factor de normalización para el documento j. Hay diversos esquemas de ponderación: binario, basado en frecuencia, logarítmico, frecuencia inversa, entropía, entre otros (Chisholm y Kolda 1999). Sin embargo, quizás el más conocido es el motivado por los primeros trabajos de Salton et al. (Salton et al. 1975) en los que considera el problema de recuperación como un problema de agrupamiento, donde dada una colección C, la consulta q, es una descripción vaga de los documentos que pertenecen a un grupo R (documentos relevantes para q). En un problema de agrupamiento deben resolverse dos situaciones: primero, determinar los atributos de los documentos que pertenecen a R (similitud *intra-clustering*) y segundo, los atributos que diferencian a los documentos de R de los restantes de la colección (diferencia *inter-clustering*). En el modelo vectorial la similitud *intra-clustering* se determina estableciendo la frecuencia del término k_i en el documento d_j, es decir establece una

medida de cuan bien el término describe el contenido del documento. Mientras que la diferencia *inter-clustering* se mide por la frecuencia inversa definida como relación inversa del número de documentos en los que aparece el término k_i entre el total de documentos de la colección. Este factor restará importancia a los términos que aparecen en muchos documentos ya que éstos carecen de utilidad para discriminar entre documentos relevantes y no relevantes (Baeza-Yates y Ribeiro-Neto 1999). Por lo tanto, un esquema de pesado efectivo tratará de equilibrar estos dos factores. El esquema de ponderación planteado por Salton et al. utiliza ponderaciones dadas por la fórmula:

$$w_{i,j} = tf_{i,j} \times idf_j \tag{2.4}$$

donde $tf_{i,j}$ se define como:

$$tf_{i,j} = \sharp(d_i, t_j) \tag{2.5}$$

en donde la función $\sharp(d_i, t_j)$ regresa el número de veces que el término t_j aparece en el documento d_i.

Por otra parte idf_j, se define como:

$$idf_j = log\frac{N}{df_j} \tag{2.6}$$

donde N es el número total de documentos de la colección y df_j es el número de documentos en los que aparece el término j.

Representados los documentos y consultas, estos deben compararse para establecer la similitud entre ellos. Hay diferentes medidas como son el coeficiente de Dice, de Jaccard y la más común el coseno del ángulo formado entre la consulta y los documentos, es decir:

$$sim(q, d) = \frac{\sum_{j=1}^{n} w_{qj} w_{dj}}{\sqrt{\sum_{j=1}^{n} (w_{dj})^2 \sum_{j=1}^{n} (w_{qj})^2}} \tag{2.7}$$

Dada una colección de m documentos y n términos diferentes, su representación vectorial estará dada por una matriz V de dimensiones n x m, conocida como matriz de términos contra documentos. En esta matriz los vectores que representan los términos se consideran ortogonales. Por lo tanto al comparar el documento d con la consulta q empleando el coseno, por ejemplo, su similitud estará dada únicamente por los términos semejantes que aparecen en ellos, sin importar los términos semejantes semánticamente. La representación de los documentos empleada en este modelo, por considerar únicamente la extracción de las palabras presentes en ellos, sin realizar ningún análisis sintáctico o procesamiento de lenguaje natural adicional, se conoce como bolsa de palabras (BoW).

El modelo vectorial, a pesar de su simplicidad, es una estrategia de recuperación robusta que funciona en general para cualquier colección, quizás la única desventaja que puede adjudicársele, es la suposición de que los términos en los documentos son independientes, ya que la expresión 2.7 no captura ninguna dependencia entre los mismos. El orden de los documentos generado por este modelo (*ranking*) a veces es difícil de mejorar, aun con técnicas como expansión de consulta o retroalimentación de relevancia (Baeza-Yates y Ribeiro-Neto 1999).

En la siguiente sección se presenta la representación de bolsa de conceptos (BoC) cuyo propósito es disminuir la limitante de considerar a los términos como completamente independientes. BoC captura ciertas relaciones entre términos, que están implícitas en el contexto de los documentos.

2.3 Representación de Bolsa de Conceptos

Existen modelos vectoriales que intentan capturar el "significado"de las palabras. Este tipo de espacios vectoriales se conocen como espacios de palabras (*word-spaces*) y los vectores que generan son llamados vectores de contexto (Gallant 2000). El término espacio de palabras fue introducido por Hinrich Schütze (Sahlgren 2006), para identificar a los modelos que representan información semántica de las palabras derivada de información de coocurrencia. Estas aproximaciones se apoyan en la hipótesis de distribución formulada por el lingüista Zellig Harris, que establece que términos con patrones de distribución similar tienden a tener el mismo significado (Sahlgren 2006). Por lo tanto, puede decirse que palabras en el mismo contexto serán semánticamente similares.

Lavelli et al. (Lavelli et al. 2004) plantean dos formas de construir vectores de contexto, la primera conocida como DOR (Document Ocurrence Representation) y la segunda TCOR (Term Co-Ocurrence Representation). Ambos métodos están motivados por la *hipótesis de distribución*, la cual establece que palabras con propiedades de distribución similares tienen significados similares. Esta hipótesis fue formulada por el lingüista Zelling Harris.

En DOR, dado un vocabulario τ, el término $t_j \in \tau$ se representa como el vector $\mathbf{t}_j = (w_{1j}, w_{2j}, \cdots, w_{|M|_j})$ donde $|M|$ es la cardinalidad de los documentos (contextos) de la colección y w_{kj} representa la frecuencia ponderada del término t_j en el contexto $k \in M$ es decir la contribución del contexto k a la especificación de la semántica del término t_j.

En TCOR el significado del término $t_j \in \tau$ se representa como el vector $\mathbf{t}_j = (t_{1j}, t_{2j}, \cdots, t_{|\tau|_j})$ donde t_{kj} representa la frecuencia ponderada del término t_j que ocurre simultáneamente con el término $t_k \in \tau$ en algún contexto. En esta repre-

sentación los contextos son conjuntos de términos definidos en ventanas simétricas centradas en t_k.

La elección de cualquiera de estos métodos tendrá por lo tanto repercusión en el tamaño de la matriz que representa la colección. Si se elige DOR la matriz será del orden $\mid \tau \mid \times \mid M \mid$ y si se opta por TCOR $\mid \tau \mid \times \mid \tau \mid$.

Utilizando los vectores de contexto, Sahlgren and Cöster (Sahlgren y Cöster 2004) introducen una representación que llaman Bolsa de Conceptos (BoC) basada en la intuición de que el significado de los documentos puede expresarse como la unión del significado de sus términos. En esta aproximación, utilizando información de co-ocurrencia, se generan vectores de contexto para cada término de un documento. Entonces, el vector de dicho documento será la suma ponderada de los vectores de contexto de sus términos. De esta manera, BoC tiene mayor probabilidad de mantener relaciones de sinonimia que las representaciones que identifican términos de manera superficial, como el modelo vectorial (Fishbein y Eliasmith 2008a).

2.3.1 Consideraciones Computacionales

En la representación de BoW las matrices que se generan usualmente son muy dispersas, pues los documentos sólo contienen una fracción pequeña de los términos del vocabulario total de la colección. En BoW, la construcción de los vectores es eficiente de manera computacional, lineal con respecto al número de términos presentes en los documentos. Por otra parte, la construcción de las representaciones de BoC es costosa. Tal es el caso de TCOR, especialmente cuando se consideran ventanas muy grandes. Para esta representación, construir los vectores de contexto tiene complejidad cuadrática con respecto a la ocurrencia de términos por documentos. La complejidad de sumar los vectores de contexto variará de acuerdo a la cantidad de términos. Sin embargo los vectores de documentos serán muy densos (vectores con pocos elementos igual a cero).

En DOR los vectores de contexto son el dual de los vectores de documento en BoW por lo tanto construir y sumar estos vectores tendrá un complejidad lineal en función del número de términos por documento. Sin embargo, los vectores de documento resultantes serán, al igual que en TCOR, muy densos.

2.3.2 Reducción de Dimensión

Considerando lo expuesto en la sección anterior, cualquier algoritmo que implementa un modelo de espacio vectorial tiene que manejar la dimensión potencialmente

alta de los vectores de contexto, para evitar afectar su eficiencia y escalabilidad. El balance, entre la cantidad de datos de coocurrencia utilizados y el tamaño de la matriz de coocurrencia que se utiliza para generar los vectores de contexto, debe mantenerse. Como se mencionó previamente, la mayoría de las celdas en la matriz de coocurrencia es cero, dado que la mayoría de las palabras ocurren en contextos limitados, por lo tanto los algoritmos de espacio de palabras buscarán reducirla a fin de hacer eficiente su utilización. Hay diferentes métodos para reducir la dimensión del espacio vectorial (Deerwester et al. 1990) (Wong et al. 1985) (Hofmann 1999). Uno de los más conocidos en recuperación de información es la indexación semántica latente (LSI) (Deerwester et al. 1990) que proyecta las consultas y los documentos a un espacio de dimensiones "semánticas latentes", donde términos coocurrentes son proyectados a la misma dimensión y los no coocurrentes a dimensiones diferentes. En LSI, una consulta y un documento pueden tener un índice de similitud alto, aunque no tengan términos en común, siempre y cuando sus términos sean semánticamente similares de acuerdo al análisis de coocurrencias. LSI es la aplicación de una técnica matemática llamada descomposición en valores singulares (SVD). Las técnicas como SVD son costosas en términos de memoria y de tiempo de procesamiento, ya que primero necesitan construir la matriz de términos contra documentos y después reducirla. Este proceso se repite cada vez que se agregan nuevos datos y debe terminarse antes de que pueda iniciarse cualquier otro procesamiento de datos, factor que hace impráctica su utilización. Como alternativa existe un método llamado Indexación Aleatoria (RI) (Sahlgren y Cöster 2004) (Sahlgren 2005), que permite de manera eficiente, escalable y con incrementos graduales construir vectores de contexto. La siguiente sección describe dicho método.

2.3.3 Indexación Aleatoria

La indexación aleatoria (RI del inglés *Random Indexing*) maneja el problema de eficiencia computacional acumulando, de manera gradual, vectores *índice* de dimensión k en una matriz de contexto R del orden $n \times k$, donde $k \ll m$, pero generalmente del orden de miles. Esto se hace en dos pasos:

(a) Una representación aleatoria única conocida como vector *índice* se asigna a cada contexto (documento o palabra). Los vectores *índice* son vectores con un pequeño número de elementos distintos de cero (ϵ), dichos elementos son igual a 1 o -1 en igual proporción. Los vectores *índice* sirven como etiquetas de las palabras o documentos (contextos).

(b) Los vectores *índice* son empleados para producir vectores de contexto recorrien-

do el texto y cada vez que una palabra dada (t) ocurre en un contexto (c), el vector *índice* del contexto (ic) se agrega al vector de contexto de la palabra (tc). Es decir, el vector de contexto de t es actualizado como $tc = tc + ic$.

De esta manera, R es una matriz de vectores de contexto de dimensión k que son la suma de los contextos de los términos. Debe notarse que estos pasos producen la matriz estándar de términos contra documentos V de orden $n \times m$ si se emplean vectores índice, con un sólo uno, de la misma dimensión que el número de contextos. Tales vectores de dimensión m serían ortogonales, mientras que los vectores *índice* de dimensión k generados aleatoriamente son sólo cercanamente ortogonales. Sin embargo, Hecht-Nielsen (Sahlgren 2005) establece que hay muchas más direcciones cercanamente ortogonales en un espacio de dimensión alta que direcciones realmente ortogonales, lo que significa que la matriz de contexto $R_{n\times k}$ será una aproximación de la matriz de términos contra documentos $V_{n\times m}$.

El aproximar V de esta manera está fundamentado en el Lema de Johnson-Lindenstrauss (Sahlgren 2005), el cual establece que si proyectamos puntos de un espacio vectorial a un subespacio seleccionado aleatoriamente de dimensión suficientemente grande, la distancia entre los puntos se preserva de manera aproximada. Este lema se utiliza no sólo en la indexación aleatoria, sino también en otras técnicas de reducción de dimensión tales como la proyección aleatoria de Papadimitriou et al. (Papadimitriou et al. 1998) y mapeo aleatorio de Kaski (Kaski 1998). Por lo tanto, la dimensión de una matriz V dada puede reducirse proyectándola mediante una matriz P.

$$R_{n\times k} = V_{n\times m} P_{m\times k} \tag{2.8}$$

Evidentemente, las técnicas de reducción de dimensión que descansan en el lema de Johnson-Lindenstrauss tienen que seleccionar la matriz P de manera apropiada. Como se mencionó anteriormente, si los k vectores aleatorios de la matriz P son ortogonales, tal que $P^T P = I$, entonces $R = V$; si los vectores aleatorios son cercanamente ortogonales, entonces $R \approx V$ en términos de la similitud de sus renglones. Achlioptas (Achlioptas 2001) ha mostrado que distribuciones simples -prácticamente todas con media de distribución 0 y varianza 1- dan un mapeo que satisface este lema. La indexación aleatoria es equivalente a la propuesta de Achlioptas con parámetros:

$$r_{ij} = \frac{1}{\epsilon}\sqrt{k} \times \begin{cases} +1 & \text{con probabilidad} \quad (\epsilon/2)/k \\ +0 & \text{con probabilidad} \quad (k-\epsilon)/k \\ -1 & \text{con probabilidad} \quad (\epsilon/2)/k \end{cases}$$

La indexación aleatoria tiene varias ventajas:

(a) Se incrementa gradualmente, lo que implica que los vectores de contexto pueden ser utilizados para calcular similitudes, aun cuando solo unos cuantos documentos se hayan procesado. En contraste, otros espacios de palabras requieren que la totalidad de los datos se represente en una matriz V de coocurrencias, antes de que el cálculo de similitudes pueda realizarse.

(b) Utiliza dimensionalidad fija, lo que significa que al agregarse nuevos datos la dimensión de los vectores no se incrementa.

(c) Se tiene reducción de dimensión implícita, ya que la dimensión de los vectores es mucho menor que el número de contextos en la colección ($k \ll m$). Esto conduce a un ahorro significativo de tiempo de procesamiento y consumo de memoria. Por ejemplo, la complejidad de calcular la SVD es del orden O(nzm), donde n es el tamaño del vocabulario, m es el número de documentos, y z es el número de elementos diferentes de cero por columna. En contraste, la complejidad de producir los vectores de contexto con RI es sólo del orden O(nr) donde n es como se describió previamente y r es la dimensión de los vectores. Nótese que el método no depende de la construcción de la matriz de términos contra documentos (V).

(d) Es comparablemente robusto con respecto a la elección de los parámetros. Otros espacios de palabras, como LSI, son muy sensitivos a la elección de la dimensión del espacio reducido. Para RI, la elección de la dimensión vectorial es un intercambio entre eficiencia y rendimiento. Se ha mostrado que el rendimiento de RI alcanza un estado estable cuando la dimensión de los vectores es suficientemente grande, pero aun así con $k \ll m$.

El presente trabajo utiliza RI para generar vectores, que se emplean en la representación de documentos como BoC. Sin embargo, BoC ignora la gran cantidad de información sintáctica existente en los documentos, que no es capturada con información de coocurrencia entre palabras, entonces se plantea capturar dicha información sintáctica utilizando la representación holográfica reducida (HRR), discutida a continuación. Previamente se establece la relación de la HRR con otras representaciones utilizadas para definir modelos de procesos cognitivos.

2.4 Representaciones Distribuidas

Un modelo conexionista (red neuronal) se define por un conjunto de unidades de procesamiento interconectadas en una red, como se ilustra en la Figura 2.2

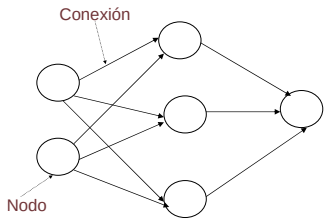

Figura 2.2: Modelo Conexionista

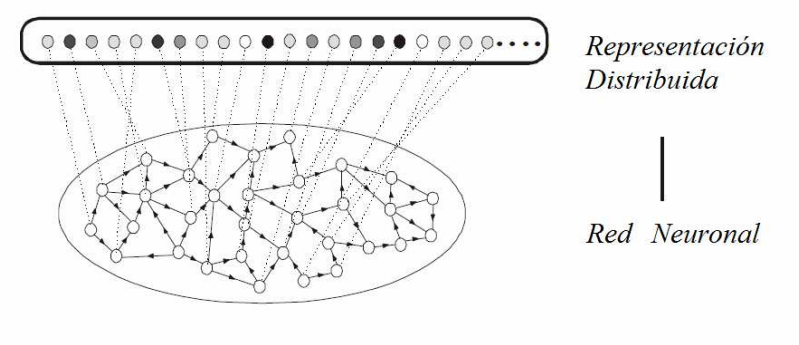

Figura 2.3: Transición de una red neuronal a una representación distribuida.

Existen dos representaciones principales para dichos modelos: las *localistas* y las *distribuidas*. En las *localistas*, cada concepto se representa por una unidad, mientras que en las *distribuidas* se emplea un número fijo de unidades para representar diferentes conceptos.

Las representaciones *localistas* están más ligadas al Cognitivismo, para el cual, la cognición, se define como la manipulación de símbolos mediante determinadas reglas y por lo tanto como un proceso secuencial y localizado. Sin embargo, este esquema no concuerda con las investigaciones recientes de la neurociencia, más ligadas al conexionismo que plantea modelos del cerebro donde las operaciones son distribuidas y se generan como producto de interconexiones masivas que cambian de acuerdo a la experiencia.

Frente a los planteamientos anteriores, las representaciones *distribuidas*, surgen como una alternativa para integrar al conexionismo con el cognitivismo (Kvasnicka 2004). Éstas permiten representar símbolos como patrones distribuidos (representados como vectores multidimensionales con entradas aleatorias) de actividades neuronales, como se muestra en la Figura 2.3. Sobre dichos patrones, es posible introducir operaciones algebraicas modelando matemáticamente las operaciones cognitivas.

Al trabajar con representaciones distribuidas, varios conceptos pueden representarse al mismo tiempo en el mismo conjunto de neuronas, superponiendo sus patrones.

Matemáticamente, superposición significa sumar componentes. Entre más patrones se superponen se complica el establecer si un patrón particular está en la superposición o no. Cuando algunos patrones aparecen en la superposición, sin que realmente existan, se produce lo conocido como interferencia o imágenes fantasma.

El número de patrones que pueden superponerse, antes de que las imágenes fantasma sean un problema, depende de diferentes aspectos de la representación: el número de neuronas, el número de patrones diferentes, el grado de tolerancia al ruido y la densidad de los patrones. La densidad de los patrones puede ajustarse para optimizar las propiedades de la representación, por ejemplo, minimizar la interferencia y maximizar la capacidad. Como ventajas de estas representaciones pueden mencionarse:

(a) Uso eficiente de los recursos: una representación localista con n neuronas puede representar n diferentes entidades, mientras que una representación distribuida puede representar 2^n diferentes entidades (utilizando todos los posibles patrones de unos y ceros).

(b) Continuidad: las representaciones están en un espacio vectorial continuo, lo que permite su utilización en técnicas de aprendizaje tales como retropropagación (*backpropagation*).

(c) Límites de capacidad y degradación suave: las representaciones tienen un límite, que se alcanza suavemente, con respecto al número de conceptos que pueden ser representados simultáneamente antes de que la interferencia afecte seriamente los resultados. También el rendimiento de una red neuronal que utiliza representaciones distribuidas, se degrada suavemente ante daños de la red o ruido agregado durante las activaciones.

Las áreas de investigación principales en representaciones distribuidas son: técnicas para representar datos más complejos que elementos simples; propiedades de los esquemas de representación, por ejemplo, capacidad, escalamiento, reconstrucciones exactas; y técnicas de aprendizaje con representaciones distribuidas.

Hablando específicamente de las representación de datos complejos, debe tenerse presente que muchas tareas cognitivas involucran el entendimiento y procesamiento de información cuya estructura es muy compleja, por ejemplo, los conceptos comunicados mediante el lenguaje humano siempre tienen una naturaleza compuesta y jerárquica. En la oración *"María escucha a Naomi tocar el piano"*, existen diferentes conceptos: *"María", "escuchar", "Naomi", "tocar", "piano"*. Representar la idea completa expresada por esta oración implica representar sus componentes y las relaciones entre ellos. Si ignoramos las relaciones no podrá diferenciarse *"María escucha a Naomi tocar el piano"* de *"Naomi escucha a María tocar el piano"*. El problema de identificar qué atributo pertenece a qué objeto se conoce como problema de enlace (*binding*

problem). En el ejemplo anterior, necesitamos enlazar los conceptos *Naomi* y *piano* a la relación *tocar*. Después de construir el concepto *"Naomi tocar piano"*, éste debe enlazarse junto con *María* a la relación *"escuchar"*. Esta naturaleza compuesta y jerárquica no es exclusiva de tareas lingüísticas, también está en el entendimiento de escenas visuales y recuperación de analogías, por mencionar dos ejemplos adicionales (Plate 2002).

Los principios de estas representaciones fueron formulados a finales de los noventa (Kvasnicka 2004) y culminaron en la Representación Holográfica Reducida (HRR) establecidas por Plate (Plate 2003), cuyo objetivo principal es la representación de conceptos con estructura compuesta y jerárquica, como el ejemplo presentado anteriormente. En la siguiente sección se presenta la HRR.

2.4.1 Representación Holográfica Reducida

Antes de describir qué es la Representación Holográfica Reducida (HRR) definamos lo que son las representaciones reducidas. Hinton (Plate 2003) considera a las representaciones reducidas como un método para representar estructuras complejas en representaciones distribuidas. Plantea que en lugar de almacenar una estructura jerárquica en una única representación, se pueden colapsar o reducir ramas de la jerarquía en un solo elemento, siempre y cuando éstas no sean el foco actual de utilización. Cuando el foco se traslada a las ramas colapsadas, entonces éstas se expanden reconstruyendo la jerarquía. En la Figura 2.4, el concepto D puede ser el foco del sistema y entonces E, F y G llenarán sus roles; o puede ser sólo el valor correspondiente al rol 3 del concepto A. Por lo tanto, las representaciones reducidas son versiones comprimidas de los datos originales, en las que se descarta cierta información, sin embargo, al descomprimirlas, la información importante se mantiene. Así, este método permite almacenar datos jerárquicos en representaciones distribuidas de tamaño fijo.

Ahora bien, la Representación Holográfica Reducida (HRR), que es una implementación concreta de la noción de representaciones reducidas de Hinton (Plate 2003), fue introducida por Plate (Plate 2003), como un método para representar estructura compositiva en representaciones distribuidas. La HRR define vectores cuyos elementos siguen una distribución normal N(0,1/n) que mantienen las propiedades favorables de las representaciones distribuidas y además:

(a) Utilizan la representación de llenado de roles (role-filler), sugerida por Hinton et al. (Hinton et al. 1986). Hinton propone que la codificación de representaciones reducidas se realice como una operación lineal:

$$W_R = R_1 f_1 + R_2 f_2 + R_3 f_3 \tag{2.9}$$

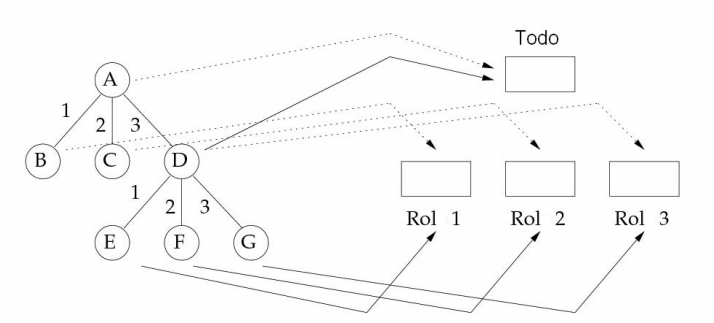

Figura 2.4: Cuando D es el foco de atención, los objetos que participan en sus roles se asocian a las unidades correspondientes (flechas sólidas). Cuando A es el foco de atención, la descripción reducida de D es asociada a la unidad correspondiente (flechas punteadas).

Donde W_R (un vector) es una representación reducida de un concepto total; f_1, f_2, f_3 llenan los roles del concepto y R_1, R_2 y R_3 son matrices aleatorias cuadradas que representan los roles 1, 2 y 3

(b) Emplean la convolución circular, en lugar del producto punto como sugiere Hinton, para vincular variables de manera recursiva de acuerdo a su rol y representar estructuras compositivas. La convolución circular permite mantener fija la dimensión vectorial.

Esta representación fue propuesta por Tony A. Plate para recuperar analogías en (Plate 1998). Él plantea que el empleo de representaciones vectoriales para procesar analogías ha sido limitado por la suposición de que es difícil o imposible representar estructura empleando vectores. Afirma que esta suposición es falsa y cita los trabajos de Smolensky (con productos tensoriales en (Smolensky 1990)), Pollack (con RAAMs en (Pollack 1990)) y Kanerva (con *binary spattercodes* en (Kanerva 1996)). Subsecuentemente, para ejemplificar el empleo de los HRRs en el procesamiento de analogías, presenta varios episodios que involucran tres personas (Jane, John y Fred), tres perros (Fido, Spot y Rover), un gato (Felix) y un ratón (Mort). Los miembros de la misma especie son considerados similares entre sí pero no similares a los de otras especies. El episodio que sirve para compararse con los otros es: (P) *"Spot bit Jane, causing Jane to flee from Spot"*. Define cinco episodios diferentes para realizar la comparación: LS (Similitud Literal) *"Fido bit John, causing John to flee from Fido"*; SF(Atributos Superficiales) *"John fled from Fido, causing Fido to bite John"*; CM (Mapeo Cruzado) *"Fred bit Rover, causing Rover to flee from Fred"*; AN (Analogía)

"Mort bit Felix, causing Felix to flee from Mort" y FOR (Sólo relaciones de primer orden) *"Mort fled from Felix, causing Felix to bit Mort"*. Cada uno de estos episodios es representado empleando HRRs. Por ejemplo, la parte de $P =$ *"Sport bit Jane"* se representará como:

$$K_{P-bite} = bite + bite_{agt} \otimes spot + bite_{obj} \otimes jane \tag{2.10}$$

Plate muestra que la recuperación de analogías empleando HRR sigue el mismo patrón observado en la gente: $LS > CM > SF > AN > FOR$. Este trabajo, aunque es la propuesta inicial del empleo de HRRs para representar estructura, no permite evaluar su utilidad para tareas de procesamiento de texto, las seis oraciones presentadas como ejemplo, constituyen poca evidencia de su utilidad.

2.4.2 Convolución Circular

La convolución circular ($\otimes$) es un operador asociativo que puede utilizarse recursivamente para representar estructuras jerárquicas en representaciones distribuidas. Su principal característica es que no incrementa la dimensión vectorial, esto es, la convolución circular de dos vectores de n elementos tendrá n elementos. En contraste, la convolución ordinaria de un par de vectores de n elementos dará como resultado un vector de $2n-1$ y el producto de tensores producirá un vector de n^2 (Plate 2003). Estos hechos explican la eficiencia y utilidad de los HRRs en una aplicación donde la cantidad de datos es del orden de miles.

La convolución circular puede considerarse la compresión del producto tensorial de dos vectores, como se ilustra en la Figura 2.5. Este operador enlaza dos vectores $\mathbf{x} = (x_0, x_1, \cdots, x_{n-1})$ y $\mathbf{y} = (y_0, y_1, \cdots, y_{n-1})$ para tener un $\mathbf{z} = (z_0, z_1, \cdots, z_{n-1})$ donde $\mathbf{z} = \mathbf{x} \otimes \mathbf{y}$ se define como:

$$z_i = \sum_{k=0}^{n-1} x_k y_{i-k} \qquad i = 0 \text{ to } n-1 \text{ (subíndices son módulo } n) \tag{2.11}$$

La convolución circular tiene propiedades en común con la multiplicación de matrices y escalares, como que es conmutativa: $x \otimes y = y \otimes x$; asociativa: $x \otimes (y \otimes z) = (x \otimes y) \otimes z$; y bilinear: $x \otimes (\alpha y + \beta z) = \alpha x \otimes y + \beta x \otimes z$. Tiene también un vector identidad: $I \otimes x = x$, $I = (1, 0, 0, \cdots)$; y un vector cero: $0 \otimes x = 0$, $0 = (0, 0, 0, \cdots)$. Para todos los vectores existe una inversa $\mathrm{x}^{-1} \otimes x = I$. En expresiones algebraicas se le da a la convolución circular la misma precedencia que a la multiplicación (i.e. $x \otimes y + z = (x \otimes y) + z$) y precedencia mayor que al producto

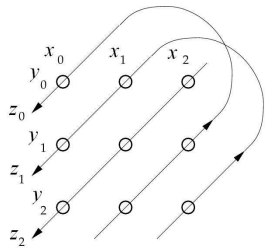

Figura 2.5: Convolución circular ($\otimes$) representada como el producto tensorial comprimido para $n = 3$. Los círculos representan el producto tensorial de x por y. Los elementos de la convolución son la suma a lo largo de las flechas de los elementos del producto punto.

punto (i.e. $x \otimes yw \otimes z = (x \otimes y)(w \otimes z)$). Como resultado, no es complicado manipular expresiones que contienen sumas, convoluciones y multiplicación por un escalar (Plate 1998). La convolución circular tiene también una inversa aproximada que es la correlación circular ($\oplus$) también considerada como una compresión del producto tensorial. Dados dos vectores, este operador enlaza dos vectores $\mathbf{x} = (x_0, x_1, \cdots, x_{n-1})$ y $\mathbf{z} = (z_0, z_1, \cdots, z_{n-1})$ para tener un $\mathbf{t} = (t_0, t_1, \cdots, t_{n-1})$ donde $\mathbf{t} = \mathbf{x} \oplus \mathbf{z}$ se define como:

$$t_i = \sum_{k=0}^{n-1} x_k t_{k+j} \qquad j = 0 \text{ to } n-1 \text{ (subíndices son módulo } n) \tag{2.12}$$

Supongamos que tenemos la convolución circular de dos vectores $\mathbf{z} = \mathbf{x} \otimes \mathbf{y}$. La correlación circular de x con z reconstruirá una versión distorsionada de y : $c = x \oplus z \approx y$. Una condición para que la correlación aproxime a la convolución es que los elementos de cada vector (de dimensión n) sean independientes e idénticamente distribuidos con media 0 y varianza $1/n$, un ejemplo de una distribución apropiada para los elementos es la normal con valores equiprobables $1\sqrt{n}$.

Ya que la convolución circular tiene sólo n elementos parece extraño que varios pares de vectores puedan almacenarse en ella, ya que estos vectores tienen también n elementos. La razón es que los vectores se almacenan con poca fidelidad, únicamente la necesaria para reconocerlos pero no para reconstruirlos íntegramente. Para distinguir los vectores, sólo necesitamos almacenar información suficiente para diferenciar un elemento de otro. Suponga que se tienen m elementos equiprobables, cada uno representado en vectores n dimensionales. Solo $2klog_2M$ bits de información se necesitan para representar k pares de estos objetos y reconocerlos posteriormente. Por ejemplo si tenemos 1024 objetos, cada uno representado por un vector, entonces el

número de bits necesarios para identificar cuatro pares de ellos es ligeramente menos que $2 \times 4 \times log_2 1024 = 80$ bits.

Una manera simple de almacenar un conjunto de vectores es superponerlos (sumarlos). Esta manera de almacenarlos no permite reconstruirlos, pero permite reconocerlos, es decir, determinar si un objeto particular se ha almacenado o no. Este tipo de memoria de superposición es útil, ya que si vectores con elementos distribuidos aleatoriamente e independientes se suman, el vector resultante será más similar a los constituyentes que a ningún otro. Cuando la superposición se aplica a memorias basadas en convolución, como los HRRs, la propiedad se mantiene, permitiendo que el resultado de varias convoluciones se almacene en una sola representación. Plate (Plate 1998) determinó una expresión para la probabilidad de reconocer todos los vectores almacenado en una superposición, dado un conjunto de vectores almacenados. Él llama a esta probabilidad $Pr(\text{All Correct})$ y muestra que puede ser calculada para vectores de longitud n como:

$$Pr(\text{All Correct}) = max_t Pr(S_a > t)^k Pr(S_r < t)^{m-k} \tag{2.13}$$

donde k es el número de vectores en la superposición, m es el total de vectores en el conjunto de vectores preestablecido, S_a es la señal de aceptación definida como $S_a = N(1, (k+1)/n)$ y S_r es la señal de rechazo definida como $S_r = N(0, k/n)$. En (Plate 2002) Plate proporciona valores para t que calcula mediante optimización numérica de la ecuación 2.13.

2.5 Resumen

En este capítulo se presentó el modelo vectorial, basado en análisis morfológico y lexicográfico de los textos para representar documentos y consultas como una combinación lineal de vectores. Los vectores utilizados representan términos que están en los documentos. Este modelo supone que dichos vectores son linealmente independientes, eliminando la posibilidad de capturar cualquier relación semántica entre los mismos. Como alternativa se presentan los modelos de espacio de palabras que construyen vectores de contexto, para capturar el "significado"de los términos con base en la información de coocurrencia de los mismos dentro del conjunto de documentos. Dado que los vectores de contexto son densos y generalmente de alta dimensión, se presentó un método conocido como Indexación Aleatoria, con el objetivo de reducir la dimensión de los vectores de contexto y hacer factible su escalamiento y empleo en aplicaciones del mundo real. Finalmente, para incorporar a la representación de documentos conceptos compuestos en los que se observa cierta jerarquía, se presentó la

representación holográfica reducida que mediante el operador de convolución circular permite vincular las partes que forman un concepto. En el siguiente capítulo se resumen trabajos que utilizan los métodos descritos.

3

Estado del Arte

En este capítulo se describen trabajos relacionados con la presente investigación, principalmente aquellos enfocados a incluir frases en el proceso de IR. Por otro lado, se discuten investigaciones que validan la utilización de la indexación aleatoria y BoC, en tareas de procesamiento de texto. Finalmente se exponen ejemplos de la utilización de los HRRs para representar textos.

3.1 Incorporación de Relaciones en IR

Como se ha mencionado, el propósito de la IR es desarrollar métodos que permitan automáticamente proporcionar información relevante a solicitudes de información. Si dichas solicitudes están expresadas en lenguaje natural, entonces el éxito de la tarea estará completamente ligado a cuán bien se modele y entienda el lenguaje. En la práctica es difícil entender el lenguaje de manera completa con las técnicas actuales, mientras que las aproximaciones basadas en bolsa de palabras (BoW) continúan prevaleciendo (Lease 2007).

Dada la consideración de independencia de los términos de los modelos basados en BoW, parece obvio que modelar alguna noción de cómo los términos se relacionan, podría ser un paso hacia contar con descripciones más ricas de los documentos y consultas.

El uso de frases como parte de la representación de documentos se ha investigado quizás desde los inicios de la IR. Dos tipos de frases son consideradas en los trabajos relacionados: las generadas estadísticamente y las obtenidas sintácticamente.

Una frase estadística es cualquier par (terna, cuádrupla, etc.) de palabras no necesariamente funcionales que frecuentemente ocurren contiguas en un corpus. Por otra parte, las frases sintácticas están formadas por cualquier conjunto de palabras que satisfacen ciertas relaciones sintácticas o constituyan cierta estructura sintáctica especificada (Mitra et al. 1997).

Empezando con los trabajos relacionados, presentándolos de manera cronológica, tenemos que Cleverdon, por ejemplo, incluye indexación basada en frases en los estudios de Cranfield en 1966 (Croft et al. 1991), después Salton (Salton 1986) muestra resultados recuerdo-precisión estándar para cuatro colecciones CACM , CISI , MED y CRAN con 3204, 1460, 1033 y 1398 documentos respectivamente. Salton utiliza fases extraídas estadísticamente, complementa la indexación realizada con términos simples y muestra porcentajes de cambios promedio en precisión para las colecciones mencionadas de 6.8 %, -8.6 %, 1.6 % y 4.2 %. Salton concluye que al menos en el previsible futuro, los sistemas de procesamiento de texto que utilicen identificadores de información complejos deben limitarse a áreas con temas restringidos.

Posteriormente, Fagan (Fagan 1987) define un método para indexado estadístico de frases, que es una extensión del presentado por Salton (Salton 1986). El método considera para la extracción de frases parámetros como *longitud*, máximo número de elementos de la frase; *dominio*, a nivel oración o documento; *proximidad* de los elementos de la frase; un umbral (DFh) para el número de documentos en los que debe aparecer al menos uno de los componentes de la frase y un umbral (DFp) para el número de documentos en los que debe aparecer la frase completa. Fagan presenta resultados en 5 colecciones cuyo tamaño varía entre 1033 y 12684 documentos; y entre 30 y 225 consultas. Con valores para los parámetros mencionados de: $longitud = 2$, $dominio = documento$, $proximidad = 1$, $DFh = 1$, $DFp = 1$, el autor alcanza cambios promedios en precisión que van de -2.2 % a 7.6 %. Estos porcentajes de cambio son con respecto a la indexación empleando términos simples. Debe mencionarse que Fagan obtiene resultados superiores a los proporcionados previamente, cuando optimiza el valor de los parámetros para cada colección en particular. El autor concluye que el método presentado, después de aplicarlo a cinco colecciones, no conduce de manera consistente a mejoras substanciales en la efectividad de la recuperación.

Otro trabajo que evalúa la utilidad de frases en IR es el de Croft et al. (Croft et al. 1991) donde modelan los documentos y consultas utilizando redes de inferencia. Los autores buscan comprobar si la utilización de consultas estructuradas que incluya frases mejora la efectividad de la recuperación. Las frases son seleccionadas utilizando un analizador sintáctico, un etiquetador de partes de la oración y dicho etiquetador combinado con un diccionario. Los investigadores extraen las frases de las consultas de manera manual y comparan los resultados de este método con los obtenidos al

seleccionar las frases de las consultas, mediante las herramientas mencionadas previamente. Y concluyen que hay poca variabilidad entre la efectividad de seleccionar las frases de las consultas, de manera manual o automática. La colección que utilizan para sus experimentos es CACM. Algunos de sus experimentos muestran que indexar únicamente con frases, ignorando los términos simples, degrada la precisión. Prueban tres métodos para incorporar las frases a su modelo: a) Considerar las frases como la conjunción de los términos simples. Y entonces, la certeza de una frase se calcula como el producto de la certeza de los términos que la componen, aunque no todos estén presentes; b) Tratar las frases como una relación de proximidad, donde sólo a los documentos que contienen todos los términos de la frase, cumpliendo cierto criterio de proximidad, se les asigna un nivel de certeza mayor al default; c) Utilizar un método híbrido, que es una combinación de los dos anteriores. El criterio de proximidad se valida primero y si no se cumple entonces se considera la conjunción de los términos presentes. Los autores reportan los mejores resultados con el modelo híbrido, cuando las frases son extraídas utilizando un etiquetador de partes de la oración y validadas contra un diccionario de frases. Con el uso del diccionario reportan un cambio promedio en precisión de 3.1 %. La utilización del etiquetador permite alcanzar 2.4 % de cambio. Los resultados más bajos los obtienen con el analizador sintáctico con un -10.1 %. Estos porcentajes de cambio son con respecto a los valores recuerdo-precisión estándar, reportados por Salton, utilizando únicamente términos simples.

Posteriormente Evans y Zhai (Evans y Zhai 1996) proponen la identificación de cuatro tipos de frases como elementos de indexación: a) átomos léxicos (por ej. "hot dog"), b) pares modificadores de encabezado (head-modifier pairs), c) subcomponentes y d) pares de modificación de preposiciones cruzadas. Los casos mencionados, pueden ilustrarse considerando el siguiente texto: "the quality of surface of treated stainless steel strip". Ejemplo de cada categoría serían: b) "treated strip", c) "stainless steel strip", y d) "surface quality". Los autores generan las frases candidatas y aceptan como válidas sólo aquellas para las cuales encuentran evidencia de su existencia independiente en el corpus (ocurrencias en el corpus). La extracción de átomos léxicos, la efectúan en múltiples fases. En cada fase, sólo dos unidades adyacentes son consideradas, así que al inicio sólo se detectan átomos léxico de dos palabras. Estos átomos, en la siguiente fase, se consideran como términos simples y entonces la longitud de los nuevos átomos puede incrementarse. Los autores definen cuatro heurísticas, en función de la frecuencia de los candidatos, para la postulación de átomos léxicos. Posteriormente, Evans y Zhai asignan a cada par generado un puntaje de acuerdo a cuatro fórmulas definidas con parámetros, como la frecuencia de los términos, de los bigramas adyacentes y discontinuos, entre otros. La idea detrás de este proceso,

que requiere de múltiples pasadas por el corpus, es que la agrupación de palabras, efectuada en diferentes pasos, eventualmente producirá la estructura más restrictiva e informativa de las frases. Realizan experimentos con la colección Associated Press Newswire de 1989 del TREC, que consiste de 84,678 documentos y 50 consultas, su base de comparación son los resultados obtenidos por el proceso estándar de CLARIT, sistema participante en el TREC 5 en la tarea de NLP. Sus resultados muestran 1 % de mejora en recuerdo y un 4.38 % en precisión-recuerdo estándar. Los autores concluyen diciendo que el análisis de las subestructuras existentes en las frases nominales, puede mejorar la efectividad en IR y otras tareas como son agrupación de conceptos y generación de resúmenes. Además, ellos comentan que la estructura extraída podría reflejar directamente las relaciones lingüísticas entre términos.

Strzalkowski y Vauthey (Strzalkowski y Vauthey 1992) extraen frases utilizando un analizador sintáctico y representándolas como pares cabeza-modificador, donde la cabeza es el elemento principal (verbo principal, nombre principal, etc.) y los modificadores son alguno de los argumentos adyacentes de la cabeza. Realizan experimentos con CACM obteniendo resultados similares los de Croft et al. (Croft et al. 1991).

Mitra et al. (Mitra et al. 1997) conducen un estudio que tiene por objetivos comprobar si extraer las frases de manera estadística o sintáctica repercute en la efectividad de la recuperación de información; y observar el efecto de las frases con diferentes esquemas de pesado. Los autores definen frase estadística como un par de palabras contiguas que aparecen en al menos 25 documentos del disco 1 del TREC, utilizan truncamiento y ordenan los componentes alfabéticamente, es decir "United States"se convierte en "stat unit". Para extraer las frases sintácticas utilizan un etiquetador de partes de la oración y extraen los componentes identificados como frases nominales (NP). Un subsistema de procesamiento de lenguaje natural (NLP), retorna una lista de las NP máximas encontradas en los documentos. Una NP es considerada máxima si no forma parte de otra NP mayor. Las palabras vacías son eliminadas y se truncan las restantes para obtener las frases nominales que se emplean para indexar los documentos. Las frases formadas por tres palabras o más, además de emplearse en su totalidad, se emplean para obtener frases de dos palabras con todas las combinaciones posibles de los términos. El esquema de pesado empleado es *tf.idf*. Como el análisis sintáctico de los documentos es costoso desde el punto de vista computacional, no indexan toda la colección y experimentan aproximando la *idf* de las frases con 6 fórmulas diferentes: a) tomar la *idf* máxima de los componentes, b) la mínima, c) la media aritmética de los componentes, d) la media geométrica, e) calcular la *idf* como el número de documentos en los que aparecen todos los componentes de las frases sin importar el orden y f) aproximar la *idf* de manera probabilística, teniendo en cuenta los documentos que contienen cada término de manera independiente y los

que contienen la frase completa. Los experimentos son realizados con las secciones de Wall Street Journal, AP Newswire, and Ziff-Davis del disco 2 del TREC con un total de 211,395 documentos y 50 consultas. Para este conjunto de consultas se tienen 4273 documentos relevantes. La colección es indexada con términos simples y estos resultados se establecen como referencia (*baseline*). Para cada consulta se eligen sólo los 100 documentos con rango mayor, se procesan para obtener las frases sintácticas y se reordena el subconjunto de documentos empleando frases. La similitud final se calcula combinando la similitud obtenida con los términos simples más la obtenida con las frases. Los resultados reflejan una mejora de 0.1 % con frases estadísticas y de 1.1 % con frases sintácticas. Debe mencionarse que las frases estadísticas las limitan a dos palabras. Inicialmente los experimentos se realizan con frases sintácticas de mayor tamaño pero se comprueba que son de más utilidad las frases de dos palabras. Los autores prueban diferentes esquemas de pesado con frases de dos palabras y observan que la diferencia en el resultado obtenido es insignificante. Otro experimento fue reordenar los 100 documentos para cada consulta sin tomar los resultados de los términos simples y utilizar únicamente frases sintácticas y estadísticas. Las frases sintácticas resultan ser mejores que las estadísticas con una diferencia en precisión a 20 documentos del 10 %. Finalmente se realizan experimentos reordenando 500 documentos y los resultados mejoran. Con frases sintácticas a 20 documentos se obtiene 1.6 % y a 500 documentos 2.2 % de mejora en precisión promedio; con frases estadísticas 1.4 % y 2.5 % respectivamente. Los investigadores concluyen que las frases nominales ayudan a mejorar el rango de los documentos a niveles de precisión bajos y no a niveles altos.

Turpin y Moffat (Turpin y Moffat 1999) revisan y extienden el trabajo de Mitra et al. (Mitra et al. 1997) utilizando la misma colección (disco 2 del TREC). Los autores consideran dos palabras continuas como frase si aparece en al menos 25 documentos del disco 2 del TREC y no del disco 1 como lo hacen Mitra et al. También para ellos es suficiente, en otro caso de estudio, que el par de palabras aparezca en al menos un documento. Otras variaciones introducidas son aceptar frases de cualquier longitud y no ordenarlas alfabéticamente. Emplean la medida de similitud BD-ACI-BCA recomendada por Zobel and Moffat para propósitos de recuperación general (Zobel y Moffat 1998). Con estas variaciones, Turpin y Moffat obtienen mejoras entre el 4.4 % y 6 % pero con respecto a líneas de referencia (*baselines*) menores que las utilizadas por Mitra. Los autores concluyen diciendo que a pesar de todas las variaciones realizadas para incluir las frases, no son capaces de identificar ningún uso de las frases que resulte en una mejora substancial en precisión. No obstante, queda abierta la posibilidad de considerar las frases en un espacio vectorial separado al de los términos, como lo sugiere Smeaton y Kelledy (Smeaton y Kelledy 1998).

Doucet y Ahonen-Myka (Doucet y Ahonen-Myka 2004) generan las frases emplean-

do secuencias frecuentes maximales (MFS) calculadas con un algoritmo que combina métodos ascendentes y exhaustivos. Ellos calculan dos valores que llaman estado de la recuperación (RSV): uno para la recuperación con el VSM clásico y otro con MFS. Los autores dividen la colección en subcolecciones, agrupando los documentos empleando k-means. Las MFS son calculadas en cada subcolección y después unidas para tener la extracción total. El cálculo de las MFS no tiene restricción de longitud. Sin embargo, la similitud entre documentos y consultas es calculada tomando únicamente grupos de dos palabras. Estos grupos se generan dividiendo las MFS en pares, permitiendo separaciones entre sus elementos de máximo una palabra. La importancia de los pares generados se determina con *idf* y si las palabras están separadas, este valor decrece en un factor α. La similitud final se calcula sumando la obtenida con el VSM, multiplicada por un factor λ; y la obtenida con las frases, multiplicadas por $(1 - \lambda)$. Los resultados de cada modelo se normalizan. La validez de esta propuesta se verificó con la colección INEX (*Initiative for the Evaluation of XML retrieval*) con 12,107 documentos de textos científicos escritos en inglés de revistas de la IEEE. Los resultados que sirvieron como base fueron los generados por VSM. La métrica utilizada para comparar los resultados fue precisión promedio a 10, 50 y 100 documentos. Las mejoras obtenidas son de -1.2 %, 0 %, 26.6 %, respectivamente. Doucet y Ahonen-Myka concluyen diciendo que las frases, como lo estableció Mitra, son más útiles a niveles de recuerdo alto y que las MFS son más útiles en casos que requieren búsqueda de información exhaustiva. Finalmente los investigadores mencionan que falta comprobar si el método proporciona los mismos resultados con colecciones cuyos textos no manejen una terminología común, como es el caso de INEX y comprobar su utilidad con otras colecciones, por ejemplo de noticias de periódicos.

En los últimos años, Vilares et al., (Alonso et al. 2002), (Vilares et al. 2004) y (Vilares et al. 2005), presentan trabajos en los que la extracción de relaciones ha mejorado la precisión de la IR. Los autores utilizan datos etiquetados para construir árboles correspondientes a frases nominales y a sus variaciones sintácticas y morfológicas. Los árboles construidos son analizados en busca de todas las dependencias binarias (nombre-modificador, sujeto-verbo y verbo-complemento) posibles; y combinados, para obtener el árbol que representa el patrón sintáctico de la frase. Dicho patrón lo transforman en una expresión regular, que conserva las dependencias binarias y que permite extraer términos multipalabra para indexar documentos. Los investigadores presentan experimentos realizados con la colección del CLEF 2001, sin superar a la aproximación de bolsa de palabras. Los autores prosiguen con sus experimentos y plantean la utilización de un analizador sintáctico superficial (shallow parser) de cinco capas para identificar dependencias sintácticas. Las dependencias que obtienen, las utilizan como términos compuestos para indexar los documentos; y realizan prue-

bas con el corpus del CLEF 2003, sin éxito. Posteriormente, los autores utilizan la colección del CLEF 2001/02 formada por 215,738 documentos. En los experimentos consideran indexar tanto los documentos como la consulta por términos simples (palabras) y términos compuestos (pares de dependencias sintácticas). Primero aplican lematización a la consulta, realizan la consulta y extraen los t mejores términos de los n documentos evaluados como más aproximados a la consulta. Los t términos son empleados para expandir la consulta y realizan una nueva consulta para obtener el conjunto final de documentos recuperados (retroalimentación por relevancia). Esta vez sus resultados muestran mejora. En sus trabajos se presentan resultados de experimentos que permiten observar que la mejora se mantiene, aun utilizando sólo frases nominales. La mejora que obtienen con precisión a 1000 documentos es de 1.35 %, y a 30 documentos 13.08 %, con respecto al modelo vectorial tradicional.

La tabla 3.1 muestra un resumen de los principales trabajos revisados donde puede observarse el tipo de representación empleado, el tamaño de las colecciones, el número de consultas, la línea base de referencia, la métrica utilizada y el porcentaje de mejora obtenido.

Tabla 3.1: Trabajos principales que han utilizado frases en IR

Autor	Representación	Método de extracción	Colección	Número Documentos	Número Consultas	Línea base	Métrica	% Cambio
Salton 1986	BoW	Estadístico	CACM, CISI, MED, CRAN	Entre 1033 y 3204	Entre 30 y 225	VSM	Precisión	Entre -8.6 y 6.8
Fagan 1987	BoW	Estadístico	CACM,CISI,MED CRAN,INSPEC	Entre 1033 y 12684	Entre 30 y 225	VSM	Precisión	Entre -2.2 y 7.6
Croft 1991	Redes de Inferencia	Sintáctico	CACM	3204	50	VSM	Precisión	3.1
Evans 1996	BoW	Estadístico	Associated Press Newswire	84678	50	CLARIT	Precisión	4.38
Mitra 1997	BoW	Estadístico y Sintáctico	WSJ,AP NewsW and Ziff-Davis	211395	50	VSM	Precisión 100 docs.	0.1 est. 1.1 sin.
Turpin 1999	BoW	Sintáctico	WSJ,AP NewsW, and Ziff-Davis	211395	50	VSM	Precisión 500 docs.	Entre 4.4 y 6
Doucet 2004	BoW	MFS	INEX	12107	30	VSM	Precisión 50 y 100	0 y 26.6
Vilares 2005	BoW	Árboles Sintácticos	CLEF 2001 y 2002	215738	46	VSM	Precisión 1000 docs.	1.35

Ciertamente, como lo muestran los trabajos comentados, desde los inicio de la IR ha existido la intuición de que las frases, si se utilizan adecuadamente, pueden mejorar la especificidad del lenguaje, y en consecuencia, la calidad de la representación de los documentos y por ende la IR. Sin embargo, los resultados no han sido homogéneos y tampoco concluyentes. Incluso se ha cuestionado la utilidad del procesamiento de lenguaje natural (NLP) en IR (Sparck 1999), (Lewis y Sparck 1996), (Brants 2004), (Arampatzis et al. 1999). Tal vez uno de los mayores retos ha sido escalar los resultados alentadores que se obtuvieron al inicio con colecciones pequeñas. A pesar de este panorama, en el presente trabajo, se retoma la intuición de que modelar la sintaxis de los textos puede mejorar la efectividad de la IR. Cabe preguntarse: ¿Por qué retomar un tema cuyo número de publicaciones en las últimas conferencias del área de 2001 a 2003 (SIGIR, ECIR, ACL) ha representado apenas el 2 % del total de los trabajos (Brants 2004)?

Si analizamos los trabajos en el área, observamos que ahora quizás más que nunca, lograr incrementos en la precisión con los paradigmas existentes, se ha tornado difícil de alcanzar. Esto hace evidente que se necesita empezar a pensar más allá del marco de trabajo existente, sí se desea mejorar la efectividad de la IR. Como Lease en (Lease 2007), parafraseando a Rijsbergen (Rijsbergen 1979), plantea: "El tiempo parece madurar para intentar una vez más con el procesamiento del lenguaje natural en recuperación de información, así, considerar de nuevo esta línea de investigación y ver si hay alguna luz que se haya derramado sobre ella en los años intermedios". Por estas razones en el presente trabajo se ha tomado una representación propuesta en la ciencia cognitiva y se ha experimentado con ella, para complementar la BoW.

En las descripciones de los trabajos relacionados encontramos que se varió el esquema de pesado, los métodos de extracción de frases, las métricas utilizadas, pero sólo Croft et al. variaron la representación.

Por otra parte, la sintaxis nos proporciona información de la composición del lenguaje, de cómo las palabras se combinan en frases y las frases forman oraciones completas. Si logramos representar cuáles palabras modifican a otras, estaremos dando un paso a tener una interpretación más precisa de los documentos, que la simple BoW. Recuperar e interpretar de manera apropiada la sintaxis, son tareas necesariamente precursoras de la construcción de sistemas capaces de entender el lenguaje natural, objetivo final de las investigaciones en NLP.

Finalmente, pareciera que el uso del NLP se enfatizó en torno al TREC-5. Sin embargo, aunque las técnicas reportadas mejoraron la recuperación, lo hicieron de manera irrelevante, a un alto costo de procesamiento. Posteriormente el interés en este tipo de técnicas decreció, como se refleja en el escaso número de trabajos relacionados. En la presente investigación, pensamos que la existencia de herramientas

de procesamiento de lenguaje natural con porcentajes de error menor al de las que existían durante el TREC-5, además de la existencia de colecciones más grandes y consultas más variadas, podrían ser factores que condujeran a resultados satisfactorios.

3.2 Indexación Aleatoria y BoC

Hay trabajos que validan el uso de la indexación aleatoria (RI) en tareas de procesamiento de texto: por ejemplo, Kanerva et al. (Kanerva et al. 2000) utilizan RI para resolver parte del TOEFL, en el cual, dada una palabra, la persona debe seleccionar su sinónimo de una lista de alternativas. Los resultados obtenidos son entre 48-51 % de respuestas correctas utilizando DOR. Karlgren y Sahlgren (Karlgren y Sahlgren 2001), (Sahlgren 2001) usan TCOR para mejorar los resultados en la misma tarea, obteniendo entre 64.5 % - 67 % respuestas correctas, porcentajes comparables a los resultados reportados por los solicitantes extranjeros a universidades de Estados Unidos, que obtienen aproximadamente el 64.5 % de respuestas correctas. Sahlgren & Karlgren (Sahlgren y Karlgren 2005) demuestran que RI puede aplicarse a textos paralelos para construir vocabularios bilingües de manera automática. En sus experimentos extraen vocabulario bilingüe de manera automática trabajando con datos paralelos de textos en sueco -español e inglés- alemán. Ellos calculan la superposición entre el vocabulario extraído automáticamente y su punto de referencia que son los diccionarios *Lexin's online Swedish-Spanish lexicón* y *TU Chemnitzónline English-German*. Los autores obtienen 60 % de traducciones correctas, cuando incluyen únicamente términos con frecuencia por arriba de 100 ocurrencias en el lenguaje fuente. Sahlgren & Cöster (Sahlgren y Cöster 2004) utilizan RI en la tarea de clasificación. Ellos utilizan la colección Reuters-21578 y BoC para representar los documentos. Dichas representaciones son la entrada a un clasificador de máquinas de vectores de soporte (SVM). Los resultados que obtienen son comparables a la representación estándar del modelo vectorial con alrededor del 82 % de precisión.

3.3 Representaciones Holográficas Reducidas

La utilización de las representaciones holográficas reducidas para representar texto es una idea novedosa, de la cual sólo se tiene conocimiento de los trabajos reportados por Fishbein y Eliasmith (Fishbein y Eliasmith 2008a) (Fishbein y Eliasmith 2008b) para clasificar documentos. En (Fishbein y Eliasmith 2008a) los autores utilizan HRRs y RI para clasificar documentos experimentando con la colección 20 Newsgroups de

20,000 documentos. Los HRRs se forman con los vectores de contexto obtenidos con RI y las etiquetas de parte de la oración asignadas al texto. Los experimentos reportados con un clasificador basado en SVM, método de aprendizaje uno contra todos y validación cruzada de diez pliegues, produce un valor de F_1 (promedio-macro) de 58.19 % para la representación HRR, en contraste con la obtenida por BoC, de 56.55 %. La dimensión vectorial que emplean es de 512. Estos resultados son importantes para demostrar la utilidad de HRR para representar documentos en tareas de tratamiento de información, sin embargo, no puede establecerse si dicha representación contribuye a mejorar la efectividad de la tarea dado que sólo se comparan con BoC, que generalmente está por debajo del VSM.

En (Fishbein y Eliasmith 2008b), los autores exploran diferentes métodos para representar semántica y sintaxis de los documentos, empleando RI para generar vectores de contexto. Para tal propósito etiquetan la colección 20 Newsgroups utilizando un etiquetado de partes de la oración (PoS). Después, todas las etiquetas sintácticas generadas, las reducen a alguna de ocho categorías (sustantivo, verbo, pronombre, preposición, adjetivo, adverbio, conjunción e interjección). Los investigadores exploran tres métodos para codificar la estructura sintáctica (etiqueta sintáctica) con la semántica de las palabras (vector de contexto). En dos de dichos métodos, cada etiqueta sintáctica se representa como un HRR. Dicho HRR, según el método, se multiplica por el vector de contexto de la palabra a la que está asociada la etiqueta; o se une al vector de contexto mediante la convolución circular. En el tercer método cada palabra se concatena con su etiqueta sintáctica. Con este mecanismo se generan nuevos términos, que son la entrada para indexar la colección con RI. Realizan experimentos de clasificación utilizando SVM, vectores de dimensión de 512, método de aprendizaje uno contra todos y validación cruzada de diez pliegues. Evalúan los resultados con la medida F_1 (promedio-macro) y obtienen mejores resultados cuando emplean la convolución circular para codificar sintaxis y semántica en un sólo vector.

3.4 Resumen

En este capítulo se comentaron trabajos que utilizan frases con el objetivo de mejorar la precisión en IR. Algunos aspectos a resaltar son que la mayoría utilizan BoW para representar documentos y consultas. La extracción de frases la realizan estadísticamente o sintácticamente sin que se identifique alguna diferencia notoria entre dichos mecanismos. Las frases son extraídas empleando diferentes herramientas (etiquetadores de partes de la oración, analizadores sintácticos, diccionarios) y aunque la longitud de las mismas en algunos casos no se limita, los mejores resultados se obtienen con frases de dos palabras. También, hay experimentos con diferentes

esquemas de pesado, sin que se establezca una diferencia considerable entre ellos. La posibilidad de considerar espacios separados para los términos simples y las frases, se deja abierta como investigación futura.

Los trabajos relacionados con RI, BoC y HRR, ponen de manifiesto cuán novedosos son estos métodos en el área de procesamiento de lenguaje natural. En el siguiente capítulo se explica cómo se incorporan dichos mecanismos a la IR.

4

Método Propuesto

En este capítulo se describe la utilización de los métodos presentados en el capítulo 2, con el objetivo de construir representaciones de documentos que capturen información de contexto de los términos presentes en dichos documentos y además tengan la posibilidad de representar las relaciones sintácticas existentes entre términos.

4.1 Motivación

Se ha argumentado que los procesos cognitivos de alto nivel, en particular el procesamiento lingüístico, puede comprenderse mejor integrando estructura y significado (Eliasmith y Thagard 2001). Los HRRs tienen varias propiedades que los hacen apropiados para modelar la estructura del lenguaje. Primero, la similitud de un HRR con sus componentes es nula. Entonces al combinar los términos simples, estamos creando nuevos términos sin incrementar la dimensión vectorial, que es una de las propiedades importantes de los HRRs. Por otra parte, ya que la superposición de HRRs produce un vector que es similar a sus constituyentes, se pueden representar documentos utilizando la superposición. De esta manera, los vectores de documentos serán similares a los vectores de sus componentes. Debe notarse que los HRRs ya se han utilizado para crear modelos de lenguaje: Eliasmith y Thagard (Eliasmith y Thagard 2001) han mostrado que los HRRs pueden modelar similitud semántica. Eliasmith (Eliasmith 2005) muestra la utilización de los HRRs para construir modelos cognitivos del lenguaje. Sin embargo, estas investigaciones se ubican dentro de la ciencia cognitiva. Hasta donde se tiene noticia, este trabajo es el primer intento de

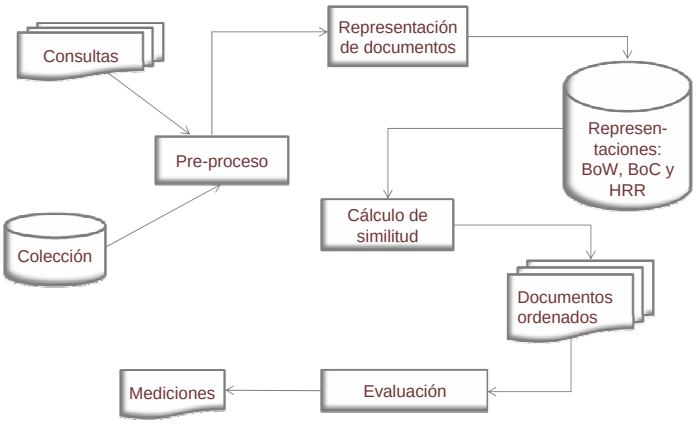

Figura 4.1: Procesos principales para generar las representaciones de documentos y establecer su impacto en la precisión.

emplear los HRRs para modelar estructura sintáctica en recuperación de información.

Por otra parte, hay métodos que no representan los textos como colecciones de palabras aisladas sino que intentan capturar el significado de las mismas, ya sea mediante conjuntos de sinónimos o dimensiones latentes (Deerwester et al. 1990). Aquí utilizamos BoC precisamente con la idea de capturar el "significado"implícito en el contexto de las palabras y de esta manera evaluar el efecto en la precisión.

El método propuesto, entonces, incorpora a la representación de BoW aspectos "semánticos" capturados con BoC y además enriquece esta representación especificando estructura sintáctica, con los HRRs.

En la Figura 4.1 se observan los principales procesos, que intervienen en la generación de las representaciones propuestas, para recuperar información. En las siguientes secciones, se detallará el funcionamiento y algoritmos empleados en los mismos.

4.2 Preprocesamiento del texto

El propósito de este componente es identificar los términos que se utilizarán en las representaciones propuestas. Los primeros pasos para crear las representaciones BoC y HRR, son: eliminar las palabras vacías, identificar los términos simples, y reducirlos a un elemento común (por ejemplo: playing, played y player serán reducidas a play). Para ello, las palabras de los documentos se separan en función de espacios y signos de puntuación. Posteriormente, se toman las palabras y con base en un grupo de

reglas, se eliminan sus terminaciones morfológicas y flexivas [1], es decir se realiza un truncamiento. La salida de este proceso es un conjunto de términos simples por documento, que denotaremos como el conjunto S_d.

En cuanto a la representación con HRRs, además de necesitar los términos simples, el texto se divide en oraciones. Dichas oraciones se etiquetan para identificar las partes de la oración, utilizando el etiquetado basado en reglas, definido por Brill (Brill 1994) de dos etapas: en la primera la etiqueta más verosímil se asigna a cada palabra y en la segunda, se aplican reglas de transformación de la forma "remplaza la etiqueta X por la etiqueta Y en un ambiente desencadenante (triggering) Z". Los ambientes desencadenantes se expanden hasta tres componentes léxicos en cada dirección y consideran las palabras, etiquetas o propiedades de las palabras dentro de la región. El etiquetador Brill alcanzó menos de 3.5 % de error en el corpus de Wall Street Journal (WSJ) (Peshkin y Savova 2003). Una vez que el texto se ha etiquetado, se identifican bloques delimitados por NX que permiten extraer las frases nominales. En la Tabla 4.1, se observa un ejemplo del preprocesamiento del texto.

De acuerdo a los trabajos de Mitra et al. (Mitra et al. 1997) y Turpin y Moffat (Turpin y Moffat 1999), donde utilizar frases de dos palabras resultó más eficaz que utilizar frases de mayor longitud, se extrajeron únicamente las frases nominales de dos palabras, que en lo subsecuente llamaremos términos compuestos (CT). En el caso de frases de más de dos palabras, se tomaron las dos últimas palabras de las mismas. Las frases con frecuencia 1 se eliminaron, ya que no existía evidencia estadística de que fuesen correctas.

Al finalizar el preproceso tendremos los términos simples y compuestos, sin palabras vacías y truncados.

[1] El algoritmo utilizado para reducir los términos a un elemento común fue el de Porter Stemmer (Porter 1980)

Tabla 4.1: Procesamiento del texto para extraer las frases nominales

Texto inicial	After a disappointing year for a lot of unit holders, fund managers are generally taking an optimistic line on 1995. Whatever we may feel about things, the investment experts, looking at the global picture, seem to have captured a feel-good factor.
Oraciones (delimitadas por apóstrofos en una lista)	[‘After a disappointing year for a lot of unit holders, fund managers are generally taking an optimistic line on 1995.’, ‘Whatever we may feel about things, the investment experts, looking at the global picture, seem to have captured a feel-good factor.’]
Etiquetado de partes de la oración	[‘After/IN a/DT disappointing/JJ year/NN for/IN a/DT lot/NN of/IN unit/NN holders/NNS ,/, fund/NN managers/NNS are/VBP generally/RB taking/VBG an/DT optimistic/JJ line/NN on/IN 1995/CD ./.’, ‘Whatever/WDT we/PRP may/MD feel/VB about/IN things/NNS ,/, the/DT investment/NN experts/NNS ,/, looking/VBG at/IN the/DT global/JJ picture/NN ,/, seem/VB to/TO have/VB captured/VBN a/DT feel-good/JJ factor/NN ./.’]
Separación en bloques delimitados por NX	[‘After/IN (NX a/DT disappointing/JJ year/NN NX) for/IN (NX a/DT lot/NN NX) of/IN (NX unit/NN holders/NNS ,/, fund/NN managers/NNS NX) (VX are/VBP generally/RB taking/VBG VX) (NX an/DT optimistic/JJ line/NN NX) on/IN (NX 1995/CD NX) ./.’, ’(NX Whatever/WDT NX) (NX we/PRP NX) (VX may/MD feel/VB VX) about/IN (NX things/NNS NX) ,/, (NX the/DT investment/NN experts/NNS NX) ,/, (VX looking/VBG VX) at/IN (NX the/DT global/JJ picture/NN NX) ,/, (VX seem/VB to/TO have/VB VX)(VX captured/VBN VX) (NX a/DT feel-good/JJ factor/NN NX) ./.’
Frases Nominales de los bloques delimitados por NX	[‘disappointing year’, ‘lot’, ‘unit holders , fund managers’, ‘optimistic line’, ‘1995’] [‘Whatever’, ‘we’, ‘things’, ‘investment experts’, ‘global picture’, ‘feel-good factor’]

4.3 Representación de documentos

En el capítulo 3 se comenta el trabajo de Croft et al. (Croft et al. 1991) en el que experimentalmente comprobó que indexar únicamente empleando frases degrada la precisión. Los trabajos reportados posteriormente (Evans y Zhai 1996), (Mitra et al. 1997), (Turpin y Moffat 1999), (Doucet y Ahonen-Myka 2004), (Zobel y Moffat 1998), (Strzalkowski y Vauthey 1992) siempre utilizan las frases como complemento de la indexación con términos simples. De igual manera, en este trabajo se propone indexar los documentos empleando BoC y HRR para complementar la representación de BoW. La representación de BoC, como se ha mencionado, aportará componentes "semánticos", mientras que la HRR elementos sintácticos.

Experimentalmente se comprobó que la utilización de BoC + HRR de manera independiente decrece el MAP (Apéndice A). Esto puede suceder porque la representación de BoC construye vectores de contexto (CV) adecuados si existen suficientes contextos (documentos), pero si los temas de los documentos son muy variados la calidad de los CV decrece, dado que no pueden definirse adecuadamente los contextos. Por otra parte, los HRR necesitan de relaciones textuales que puedan representarse, condición que no satisfacen ni todos los documentos, ni todas las consultas. Por lo tanto, es necesario contar con la recuperación inicial del VSM, y las representaciones propuestas actuarán como elementos de refinamiento.

4.3.1 Representación BoC

Para producir la representación de BoC de los documentos, éstos se indexan empleando RI, siguiendo los pasos descritos en la sección 2.3.3. Inicialmente a cada contexto se le asigna una representación aleatoria única conocida como vector índice. En la Figura 4.2 se observa este paso donde a cada uno de los n documentos de una colección D, se les asigna un vector índice conformado por 1 y -1, en igual cantidad, y el resto de elementos igual a 0. Para utilizar RI debe definirse previamente la dimensión (k) de los vectores índice que se generarán (reducción implícita de dimensión), la cantidad de elementos diferentes de cero que contendrán los vectores índice, llamada densidad (u) y el tipo de contexto a considerar (documento o términos).

RI toma todos los documentos de D a la vez y construye un modelo de la utilización de los diferentes términos en la colección. Sea S el conjunto de las términos diferentes en D es decir $S = \bigcup_{d=1}^{n} S_d$ donde n es el número de documentos de D. RI toma el conjunto S y para cada término $s_i \in S$ crea un vector de contexto que denotaremos por c_i. Los vectores de contexto de los términos se construyen sumando los vectores índice

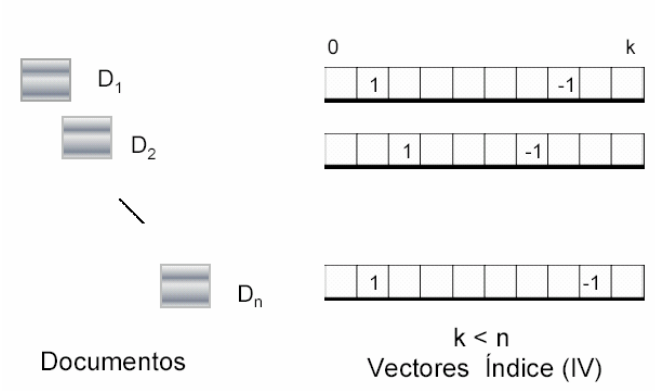

Figura 4.2: Reducción implícita de la dimensión vectorial

asignados a los contextos. Formalmente, tomando como contextos los documentos, tenemos:

$$c_i = \sum_{k=1}^{l} b_k \tag{4.1}$$

donde l es el número de documentos que contienen s_i y b_k es el vector índice del documento k. Por lo tanto, el resultado del indexado aleatorio puede definirse como una función $R : S \rightarrow C$ donde C es el conjunto de todos los c_i generados en el proceso de indexado. Entonces los resultados del indexado aleatorio, nos permitirán asociar un término s_i con su vector de contexto c_i. Ejemplificando la descripción anterior, en la Figura 4.3 suponga que se establece como parámetros de entrada a RI: $k = 10$, $u = 2$ y DOR. Entonces para los documentos D_1 y D_2 se generarán los vectores índice $IV\_D_1 = [0, 1, 0, 0, 0, -1, 0, 0, 0, 0]$ e $IV\_D_2 = [1, 0, 0, 0, 0, 0, 0, -1, 0, 0]$ que se utilizarán para construir los vectores de contexto de los términos. En la figura, el vector de contexto para *brain* se define como la suma de $IV\_D_1 + IV\_D_2 = [1, 1, 0, 0, 0, -1, 0, -1, 0, 0]$, dado que aparece en ambos documentos.

Los vectores de contexto generados de esta manera tendrán almacenada la contribución de cada uno de los contextos a la "semántica"de los términos. Finalmente, para cada s_i se define un factor de peso $w_i = tf_i.idf_i$ donde tf_i denota la frecuencia de s_i como se define en la fórmula 2.5, e idf_i la frecuencia inversa de s_i como se define en la fórmula 2.6. Con estos factores de peso y los vectores de contexto representamos cada documento d_i como:

$$d_i = \sum_{j=1}^{r} c_j w_j \tag{4.2}$$

donde r es el número de términos en d_i y c_j corresponde al vector de contexto de cada

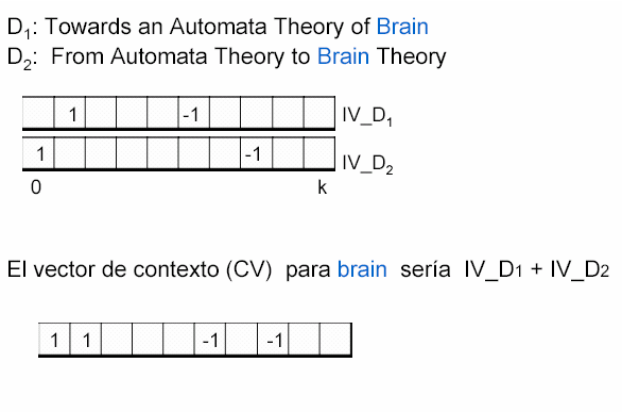

Figura 4.3: Definición de vectores de contexto para los términos simples

Tabla 4.2: Algoritmo Indexación Aleatoria DOR

Para todo documento d en D hacer

// Crea un vector índice de dimensión k y densidad u

$\mathbf{e_d} \leftarrow$ CreaVectorIndiceDoc(k, u)

Para todas las ocurrencias del término $t \in d$ hacer

$\mathbf{R_t} \leftarrow \mathbf{R_t} + \mathbf{e_d}$ // Agrega el vector índice aleatorio al renglón t de la matriz R

fin para

fin para

---

término en el documento. La generación de vectores de contexto se hace utilizando la representación de ocurrencia de documento (DOR), como se describe en la sección 2.3. El algoritmo para este proceso se muestra en la Tabla 4.2

4.3.2 Representación HRR

Los métodos tradicionales de IR que incluyen términos compuestos (CT), los extraen, y entonces los incluyen como nuevos términos dentro del modelo vectorial tradicional (Croft et al. 1991), (Evans y Zhai 1996), (Mitra et al. 1997), (Turpin y Moffat 1999), (Doucet y Ahonen-Myka 2004), (Zobel y Moffat 1998), (Strzalkowski y Vauthey 1992), incrementando la dimensión del espacio vectorial. En este trabajo, se explora una representación diferente para tales términos, llamada representación

holográfica reducida (HRR), que refleja cierta estructura sintáctica y distribuye dicha información a lo largo de los vectores que representan documentos. Con los HRRs, buscamos representar las relaciones identificadas entre términos para mejorar la expresividad y en consecuencia la precisión. La representación de dichas relaciones permite expresar cierta estructura superficial, haciendo más detalladas las descripciones de los textos.

La codificación de estructura requiere una forma de vincular elementos particulares. Para este propósito, se emplea el operador de convolución circular, descrito en la sección 2.4.2. La convolución circular permite codificar asociaciones entre términos y de esta manera representar estructura superficial. Este operador de vinculación mantiene el mismo tamaño de los vectores, puede ser decodificado y puede aplicarse recursivamente (Plate 2003), (Plate 1998).

En la codificación de relaciones, además de emplear los vectores índice de los términos que intervienen en ellas, también se utilizan los HRRs (sección 2.4.1) para identificar el papel (rol) de los términos (por ejemplo: parte derecha del término compuesto, parte izquierda del mismo). Los HRRs, junto con los vectores índices de los términos, se utilizan para codificar las relaciones empleando la convolución circular.

Dada una relación $R(r_1, r_2)$ donde r_1 y r_2 son los términos que intervienen en la relación, si éstos desempeñan un papel diferente, para codificar la relación se necesitarán dos HRRs: **izq** y **der** . El vector $\mathbf{R}$ que representa la relación será:

$$\mathbf{R} = (\mathbf{izq} \otimes \mathbf{r_1} + \mathbf{der} \otimes \mathbf{r_2}) \tag{4.3}$$

Por ejemplo, considere el término compuesto *fund manager*, éste será representado como: $\mathbf{izq} \otimes \mathbf{fund} + \mathbf{der} \otimes \mathbf{manager}$, donde **fund** y **manager** son los vectores índice de los términos.

Dado un documento d_i con términos $t_1, t_2, \cdots, t_{x1}, t_{y1}, \cdots, t_{x2}, t_{y2}, \cdots, t_{xn}, t_{yn}, \cdots, t_n$ y términos compuestos $R_1, R_2, \cdots, R_n$ entre los términos $t_{x1}, t_{y1}; t_{x2}, t_{y2}$; y t_{xn}, t_{yn}, respectivamente, su vector HRR será construido como:

$$\begin{aligned} \mathbf{d}_i = \langle w_{x1,y1}(\mathbf{izq} \otimes \mathbf{t_{x1}} + \mathbf{der} \otimes \mathbf{t_{y1}}) + w_{x2,y2}(\mathbf{izq} \otimes \mathbf{t_{x2}} + \\ \mathbf{der} \otimes \mathbf{t_{y2}}) + \cdots + w_{xn,yn}(\mathbf{izq} \otimes \mathbf{t_{xn}} + \mathbf{der} \otimes \mathbf{t_{yn}})\rangle \end{aligned} \tag{4.4}$$

donde $\langle \ \rangle$ denota que el vector es normalizado y los w_{ij} son la ponderación *tf.idf* asignada a las relaciones. Las consultas se representan de igual manera. La Tabla 4.3 muestra el algoritmo empleado para representar los documentos como HRRs.

Ilustrando la utilización del operador de convolución circular, supongamos que el documento d_i tiene los términos t_1, t_2 y t_3, cuyos vectores índice respectivos, de

Tabla 4.3: Algoritmo para representar documentos empleando HRRs

```
para todo documento d en D hacer
  representacionDocumento ← CreaVectorVacio(k)
  relacionesMap ← ExtraccionRelaciones(d)
  RolesMap ← GenerarHRR(relacionesMap)
  para toda r en relacionesMap hacer
    representacionRelacion ← CreaVectorVacio(k)
    elementosRelacion ← ObtenElementos(r)
    para todo m en elementosRelacion hacer
      m_v ← ObtenVectorIndice(m)
      rol_v ← ObtenRol(RolesMap,m)
      representacionRelacion ← representacionRelacion +
                    CircularConvolucion(m_v, rol_v)
    fin para
  fin para
fin para
```

dimensión 10 y densidad 2, se representan a continuación: $[v_0, v_1, 0, 0, 0, 0, 0, 0, 0, 0]$, $[0, 0, v_2, v_3, 0, 0, 0, 0, 0, 0]$ y $[0, 0, 0, 0, v_4, v_5, 0, 0, 0, 0]$, donde los v_i son los elementos diferentes de 0. Además, existe un término compuesto entre t_2 y t_3 y los roles parte izquierda (**izq**) y parte derecha (**der**) representados por los vectores $[0, 0, 0, 0, 0, 0, s_6, s_7, 0, 0]$ y $[0, 0, 0, 0, 0, 0, 0, 0, s_8, s_9]$ (sólo con fines ilustrativos, recuérdese la definición de un HRR, sección 2.4.1), entonces la convolución circular para dos vectores de dimensión diez se define como:

$$z_i = \sum_{k=0}^{9} x_k y_{i-k} \qquad i = 0, \cdots, 9$$

y por lo tanto:

$izq \otimes t_2 = [s_7 v_3, 0, 0, 0, 0, 0, 0, 0, s_6 v_2, s_6 v_3 + s_7 v_2]$
$der \otimes t_3 = [0, 0, s_8 v_4, s_8 v_5 + s_9 v_4, s_9 v_5, 0, 0, 0, 0, 0]$

de donde la representación HRR de d_i será:

$$(izq \otimes t_2) + (der \otimes t_3) = w_{t2,t3}[s_7v_3, 0, s_8v_4, s_8v_5 + s_9v_4, s_9v_5, 0, 0, 0, s_6v_2, s_6v_3 + s_7v_2]$$

La convolución circular es calculada de manera eficiente en un tiempo *O(nlogn)*, utilizando la transformada rápida de Fourier para convertir los vectores al dominio de frecuencias y simplemente efectuar una multiplicación elemento a elemento, para obtener el vector resultante. En contraste con el tiempo $O(n^2)$ para calcular la convolución directamente de la fórmula 2.11. De hecho, Plate en (Plate 2002) muestra que los HRRs pueden construirse y operarse enteramente dentro del dominio de frecuencias de la transformada rápida de Fourier, eliminado la necesitad de tener que trasformar continuamente los vectores de un espacio a otro. La ecuación que relaciona la convolución con la transformada de Fourier es:

$$\mathbf{x} \otimes \mathbf{y} = \mathbf{f}^{-1}(\mathbf{f}(\mathbf{x}) \circ \mathbf{f}(\mathbf{y})) \tag{4.5}$$

donde $\mathbf{f}$ representa la transformada rápida de Fourier, $\mathbf{f^{-1}}$ su inversa y $\circ$ la multiplicación de vectores elemento a elemento. Al crear la representación de los documentos, se empleó este mecanismo con el fin de reducir el tiempo de procesamiento.

La Figura 4.4 esquematiza la relación entre RI y la generación de las representaciones de documentos como BoC y HRR. Puede observarse que RI genera tanto los vectores de contexto de los términos simples para producir la representación de BoC, como los vectores índice empleados en la generación de HRRs.

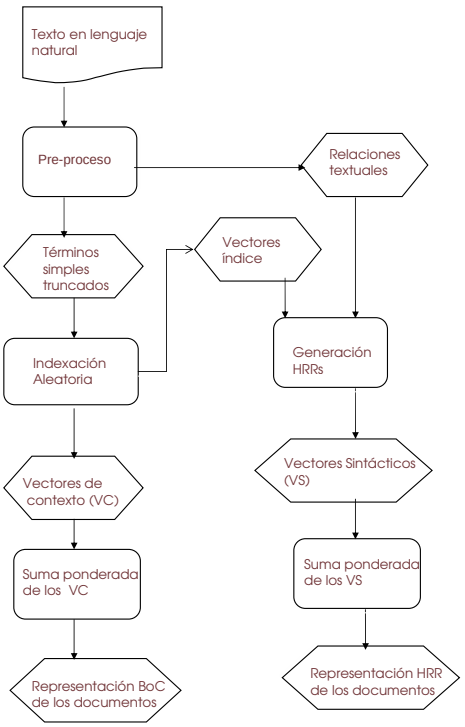

Figura 4.4: Proceso completo para generar las representaciones BoC y HRR

4.4 Cálculo de Similitud

La similitud empleando sólo términos, se calcula con el coseno del ángulo formado por los vectores de la consulta y el documento de acuerdo a la expresión 2.7. Si los documentos contienen relaciones, también empleamos el coseno:

$$Sim(d,q) = Sim(d,q)_{terminos} + (\alpha_c \cdot cos(f_{dc}, f_{qc})) + (\alpha_h \cdot \cos(f_{dh}, f_{qh})) \qquad (4.6)$$

donde f_{dc} y f_{qc} representan vectores construidos con BoC y f_{dh} y f_{qh} son los vectores construidos con las relaciones textuales (HRR). Por lo tanto, la similitud final entre un documento y una consulta, está dada por la similitud obtenida al compararlos empleando únicamente términos simples, más la multiplicación de un factor α_c por la similitud obtenida al comparar los documentos representados como BoC, más la multiplicación de otro factor α_h por la similitud entre el documento y la consulta representados como HRR. Los factores α_c y α_h son valores menores que uno, utilizados para atenuar el valor de la similitud obtenida con las codificaciones HRR y BoC.

Evidentemente, con esta nueva similitud obtendremos un nuevo ordenamiento de los documentos, que tendrá que evaluarse para comprobar si la inclusión de "semánti-

ca" y sintaxis ha contribuido a mejorar la precisión.

4.5 Reordenamiento

Existen técnicas para mejorar la efectividad de la recuperación dado un ordenamiento inicial. Una de estas técnicas es la retroalimentación de relevancia (PRF), la cual considera documentos relevantes, a los documentos colocados en las primeras n posiciones para cada consulta. Entonces las consultas son expandidas agregándoles k palabras seleccionadas de los n primeros documentos, y un segundo proceso de IR se realiza con las consultas expandidas.

Las representaciones propuestas pueden utilizarse para reordenar los resultados producidos por otro método de recuperación de información. En este caso, dadas m consultas, únicamente los primeros n documentos determinados como relevantes para ellas se indexan como BoC y HRR. De est manera contaremos con $2m$ indexaciones de documentos que se compararán con las m consultas representadas primero como BoC y luego como HRR. Finalmente, las similitudes obtenidas se combinan mediante la fórmula 4.6.

4.6 Resumen

En este capítulo se ha explicado cómo las técnicas descritas en el capítulo 2 se combinan para que, mediante la utilización de la indexación aleatoria, los documentos se representen como: a) BoC capturando algunos aspectos semánticos, y b) HRR representando cierta estructura sintáctica. Se explica la preparación previa que debe realizarse sobre los documentos y consultas de una colección. Se muestran los principales procesos y algoritmos que intervienen en la construcción de las representaciones y finalmente, se establece cómo comparar los documentos con las consultas, a fin de evaluar estos resultados y comprobar si se obtiene alguna mejora en la recuperación. En el siguiente capítulo se presenta la parte experimental de esta tesis.

5

Experimentación

En este capítulo, se discuten los resultados obtenidos al integrar las representaciones textuales, presentadas en los capítulos previos, al modelo vectorial para recuperar información. Las preguntas de investigación que se busca responder son:

- ¿Cuál será el efecto de incluir representaciones de BoC y HRR en la tarea de recuperación de información en términos de MAP?
- ¿La incorporación de HRRs afecta negativamente a los documentos que no contienen información adecuada para representarse?
- ¿La incorporación de HRRs con condiciones apropiadas de procesamiento contribuye a mejorar la precisión?

Los resultados generados por el modelo vectorial clásico (VSM), se tomaron como referencia para compararlos con los obtenidos al complementar la recuperación con la "semántica" capturada por BoC y la estructura sintáctica representada con la HRR de manera directa. El VSM se implementó usando el esquema de ponderación *tf.idf* y el coseno como medida de similitud entre documentos. Adicionalmente se realizaron experimentos de re-ordenamiento que se compararon con la implementación del VSM en Lemur y el método de retroalimentación de relevancia (PRF).

5.1 Evaluación

En esta sección se describen las métricas empleadas para evaluar los resultados obtenidos. Dos métricas que han demostrado su estabilidad para comparar sistemas de IR (Buckley y Voorhees 2000) son:

- La media de la precisión promedio (MAP), que se define como el promedio de todos los *AvgP* obtenidos para cada consulta. *AvgP* se define como:

$$AvgP = \sum_{r=1}^{m} P(r) \times rel(r)/n \tag{5.1}$$

  donde *P(r)* es la precisión a r documentos, *rel(r)* es una función binaria que indica si el documento r es relevante o no para una consulta dada q; n es el número de documentos relevantes para q; m es el número de documentos relevantes recuperados para q.

- La segunda métrica es R-Precision (R-Prec), que se define como la precisión alcanzada cuando se han recuperado R documentos, donde R es el número de documentos relevantes para la consulta q.

Por otra parte, una métricas utilizadas para establecer la consistencia de la recuperación, buscando contar con sistemas robustos, es la media geométrica de la precisión promedio obtenida para cada consulta (GMAP):

$$GMAP = exp\left(\frac{\sum_{i=1}^{n} log(x_i + \epsilon)}{n}\right) - \epsilon \tag{5.2}$$

donde ϵ es una constante positiva pequeña, n el número de consultas y x_i la precisión promedio para cada consulta.

5.2 Representaciones previas

Al iniciar la investigación, buscábamos una representación vectorial con el potencial de manejar relaciones entre términos. A diferencia de VSM, donde cada término se representa como una entrada de un vector que depende de la frecuencia del término, en la propuesta inicial se representó cada término como un patrón de cinco dígitos binarios contiguos y su correspondiente posición en el vector total. Dicha representación era semejante a los vectores índice de la indexación aleatoria, pero en este caso los patrones no se traslapaban, lo que incrementaba la dimensión del modelo vectorial por

un factor constante. Sin embargo, de esta manera se contaba con una representación de términos que podían asociarse mediante la convolución circular, para construir representaciones HRRs de las relaciones textuales. Los documentos se representaron como la suma de los patrones de sus términos simples, más la suma de sus términos compuestos, codificados como HRRs. La similitud entre documentos y consultas (codificadas de la misma manera) se obtenía con el coseno. En la Tabla 5.1 se muestran los resultados obtenidos para tres colecciones pequeñas con la propuesta descrita, y su comparación con los obtenidos por el VSM, en términos de precisión a niveles de recuerdo estándar.

Aunque esta propuesta permitía codificar las relaciones textuales como HRRs, al superponer los vectores, para representar documentos, aparecían patrones que realmente no existían (imágenes fantasma). Este fenómeno, que aumentaba con el incremento de HRRs en los documentos, inhabilitó la utilización de la propuesta inicial en colecciones de tamaño mayor.

La siguiente propuesta fue utilizar la indexación aleatoria, ahora utilizando los vectores índice (IV) como los patrones únicos de la representación anterior. Entonces los documentos se representaron como la suma de los vectores índice (términos simples) y la codificación de los términos compuestos como HRRs utilizando los IV. A esta representación se le llamó representación con vectores índice (IVR). Esto permitió reducir considerablemente la dimensión del modelo vectorial, dado que se utilizaron vectores de una dimensión fija a 4096. En esta etapa se inició la investigación del efecto que tenía complementar la IVR con la "semántica" capturada por DOR y la sintaxis representada con HRRs. También se construyeron espacios vectoriales separados para cada relación textual. La Tabla 5.2 muestra los resultados obtenidos con esta propuesta para dos colecciones pequeñas. Puede observarse que la precisión de IVR es menor que la del VSM por la reducción de dimensión. Cuando IVR es complementada con DOR, la precisión mejora. En la siguiente parte de la tabla, se observa que al agregar los términos compuestos (CT) tanto al VSM como al IVR, la precisión de esta última mejora con respecto a la primera para NPL. Los términos compuestos se agregaron al VSM como nuevos términos. En la última línea de la tabla, se observa que al integrar las tres representaciones se incrementa la precisión de manera importante para NPL.

En la colección CISI el complementar IVR con DOR no mejoró la presión. CISI tiene en total 1460 documentos con un vocabulario de 6977 palabras y 112 consultas. El número de contextos es reducido y el vocabulario casi tan amplio como el de NPL, que tiene 11429 documentos. En estas condiciones, DOR no cuenta con suficientes documentos para construir vectores de contexto, que caractericen de manera apropiada a los términos. El cambio en precisión, en consecuencia, fue de -13 % con respecto

al VSM. Por esta deficiencia, en los experimentos posteriores dejamos de considerar esta colección.

Además de representar términos compuestos con HRRs, se representaron relaciones sujeto-verbo y verbo-complemento. Un factor que actuó en contra de esta indagación, fue la escasez de este tipo de relaciones en las consultas. Como puede observarse en la Tabla 5.3, en las consultas de NPL prácticamente sólo se identificaron términos compuestos. En CACM se tomó un grupo de 21 consultas que incluían los tres tipos de relaciones bajo estudio. La función empleada para obtener la similitud después de agregar todas las relaciones fue:

$$\begin{aligned} Similitud(q,d) = Similitud\_IVR(q,d) + \beta Similitud\_CT(q,d) + \\ \delta Similitud\_SV(q,d) + \gamma Similitud\_VO(q,d) \end{aligned} \tag{5.3}$$

donde $\beta = 1/16$ y $\delta, \gamma = 1/32$, factores determinados por experimentación. La Tabla 5.4 muestra el cambio en la precisión después de agregar todas las relaciones. Contrastando la cantidad de procesamiento necesario para extraer las relaciones, contra el incremento en precisión después de estos experimentos, se decidió enfocar exclusivamente el estudio a los términos compuestos.

Los resultados mostrados en está sección, y advertir, que en el trabajo relacionado el VSM se complementaba con los resultados obtenidos al extraer los sustantivos compuestos, dirigieron la investigación hacia está dirección.

En esta nueva etapa, se utilizó el método descrito en el capítulo anterior para representar documentos y consultas. Primero realizamos experimentos para determinar el valor adecuado de los parámetros de la indexación aleatoria y de la función de ponderación (4.6). Posteriormente, se experimentó con diferentes colecciones de mayor tamaño cuyos resultados se exponen a continuación.

Tabla 5.1: Recuerdo-Precisión para colecciones pequeñas

Colección	Recuerdo	Precisión VSM / CT	Precisión Propuesta /CT	% de Cambio
CISI (76	0	0.5871	0.6423	9.40
consulta)	0.1	0.4787	0.4797	0.21
	0.2	0.3849	0.3909	1.56
	0.3	0.3077	0.3151	2.40
	0.4	0.2636	0.2698	2.35
	0.5	0.2271	0.2344	3.21
	0.6	0.181	0.1912	5.64
	0.7	0.1319	0.1375	4.25
	0.8	0.0961	0.0973	1.25
	0.9	0.063	0.0641	1.75
	1	0.0242	0.0246	1.65
	Promedio	0.2496	0.2588	3.06
CACM (50	0	0.6099	0.5842	-4.21
consultas)	0.1	0.5580	0.5723	2.56
	0.2	0.4456	0.4292	-3.68
	0.3	0.3828	0.3709	-3.11
	0.4	0.3160	0.3162	0.06
	0.5	0.2422	0.2505	3.43
	0.6	0.2159	0.2159	0.00
	0.7	0.1709	0.1693	-0.94
	0.8	0.1340	0.1310	-2.24
	0.9	0.0942	0.0932	-1.06
	1.0	0.0801	0.0798	-0.37
	Promedio	0.2954	0.2920	-0.87
NPL (92	0	0.4430	0.5137	15.96
consultas)	0.1	0.3851	0.4421	14.80
	0.2	0.3044	0.3519	15.60
	0.3	0.2397	0.2590	8.05
	0.4	0.2060	0.2200	6.80
	0.5	0.1599	0.1665	4.13
	0.6	0.1301	0.1283	-1.38
	0.7	0.1038	0.0998	-3.85
	0.8	0.0782	0.0753	-3.71
	0.9	0.0501	0.0485	-3.19
	1.0	0.0239	0.0242	1.26
	Promedio	0.1931	0.2118	4.95

Tabla 5.2: MAP comparando el VSM contra IVR e IVR-DOR

		Términos	simples		
	VSM	IVR	% Cambio	IVR+DOR	% Cambio
CACM	0.2622	0.2541	-3.08	0.2634	0.45
NPL	0.2015	0.1994	-1.04	0.2291	13.69
		Términos	compuetos		
	VSM+CT	IVR+CT	% Cambio	IVR+CT+DOR	% Cambio
CACM	0.2715	0.2538	-6.54	0.2631	-3.10
NPL	0.1857	0.1988	7.05	0.2291	23.40

Tabla 5.3: Número de consultas con las relaciones seleccionadas por colección

Colección	Término compuestos	Subject- Verb	Object-Verb
CACM	48	28	33
NPL	66	1	3

Tabla 5.4: MAP comparando el VSM con el IVR considerando todas las relaciones

VSM	IVR	% Cambio	IVR+ CT	% Cambio
0.2563	0.2570	0.28	0.2582	0.74
	IVR+CT+SV	% Cambio	IVR+CT+SV+VO	% Cambio
	0.2610	1.82	0.2693	5.07

5.3 Colección CACM

Al iniciar los experimentos no había fundamentos suficientes para seleccionar los valores de los parámetros de la indexación aleatoria que condujeran a resultados de recuperación adecuados. En los trabajos revisados, dichos parámetros se determinan de manera empírica. Así que iniciamos con la colección CACM, una colección con 3204 documentos (resúmenes) publicados en *Communications of the ACM* de 1958 a 1979 referentes a ciencias computacionales y 64 consultas. El número promedio de palabras [1] por consulta es de 13 y por documento de 29. Las consultas tienen en promedio 15 documentos relevantes. Esta colección es pequeña y por lo tanto útil para ajustar parámetros de manera empírica.

5.3.1 Determinando la Dimensión

Kaski (Kaski 1998) muestra que entre más alta sea la dimensión de los vectores, la matriz identidad será mejor aproximada por $P^T P$, y por lo tanto, la matriz generada por la indexación aleatoria será más próxima a la matriz de términos contra documentos (expresión 2.8). Por otra parte, sabemos que cuando la dimensión de los vectores es muy pequeña, RI empieza a ser deficiente por la imposibilidad de asignar vectores índice únicos a los diferentes contextos. Con estos antecedentes, para evaluar el efecto de la dimensión en IR, se exploró la efectividad de la recuperación incrementando la dimensión de manera logarítmica, y la cantidad de elementos distintos de cero varió de 2 a 40. Como se muestra en la Figura 5.1, el MAP más alto se alcanza con vectores de dimensión 4096 y 20 elementos distintos de 0. Puede observarse un incremento considerable del MAP de 512 a 1024, para después tener incrementos más suaves que tienden a estabilizarse. Con base en lo anterior y buscando reducir la dimensión empleada por el modelo vectorial clásico, se decidió trabajar con vectores de 4096.

5.3.2 Determinación de factores de ponderación

Una vez que los parámetros para la indexación aleatoria se ajustaron, se procedió a investigar el efecto de variar el factor α_c en 4.6, decreciendo su valor de 1 a 1/8. La

[1] Todas las estadísticas de número de palabras que se presentan se calcularon después de eliminar palabras vacías y efectuar truncamiento.

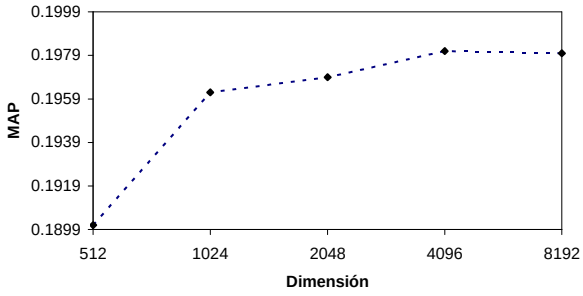

Figura 5.1: Efectividad de BoC en CACM con vectores de diferente dimensión

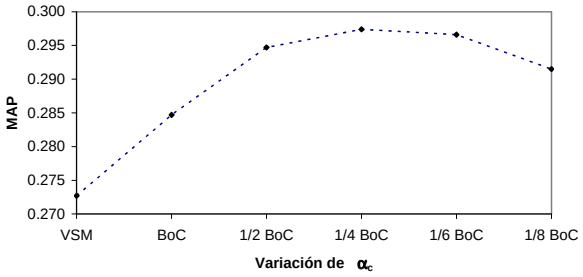

Figura 5.2: Efecto de complementar el VSM con BoC asignando ponderaciones diferentes a éste último.

Figura 5.2 muestra el efecto de complementar los resultados del VSM con los de BoC, variando el peso asignado a estos últimos. El MAP se incrementa hasta $\alpha_c = 1/4$, valor a partir del cual empieza a decrecer. Por lo tanto se resolvió hacer $\alpha_c = 1/4$ en 4.6.

Determinado el valor de α_c se procedió a ajustar el valor de α_h. En la Figura 5.3 se ve el efecto de este factor variando su valor de 1 a 1/8. Puede observarse una caída drástica del MAP cuando no se reducen los valores obtenidos por la representación HRR. El valor más alto se obtiene con el factor de peso $\alpha_h = 1/6$. Con 1/8 hay una ligera caída del -0.78 %. Finalmente para integrar las tres representaciones, a los valores del VSM complementados con 1/4 BoC, se agregaron los CT pesados por diferentes valores. En la Figura 5.4 puede observarse que el efecto es similar al obtenido al complementar únicamente los valores del VSM. Como derivación de estos resultados, $\alpha_h = 1/6$ se utilizó en los experimentos consecutivos. Todas las mediciones del MAP en estos experimentos y los siguientes se hicieron a 1000 documentos.

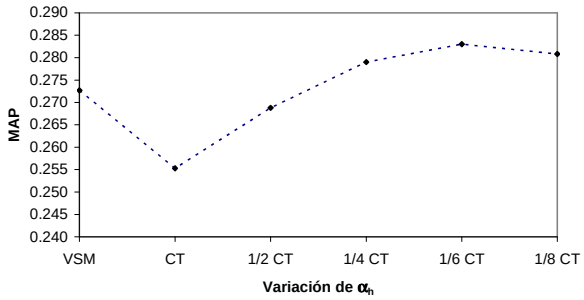

Figura 5.3: Efecto de complementar el VSM con términos compuestos representados como HRRs y variando el factor de ponderación

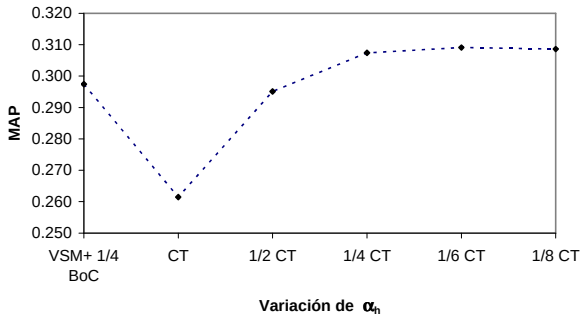

Figura 5.4: Integración de las tres representaciones. El VSM integrado con 1/4 BoC se complementa con los términos compuestos (representados como HRRs) variando el factor de ponderación.

5.3.3 Resultados

En la Tabla 5.5 se muestran características generales de la colección CACM. Únicamente 50 consultas de las 64 proporcionadas cuentan con documentos relevantes. El número de términos compuestos extraídos fue de 6415 de los cuales 5220 aparecen en un sólo documento. Los términos con frecuencia igual a uno no se tomaron en cuenta, pues no existe certeza estadística de que sean correctos.

La Tabla 5.6 muestra la comparación entre el VSM y los resultados al añadirle la "semántica" capturada por BoC. De las 50 consultas consideradas, 32 tuvieron un porcentaje de cambio en precisión promedio (AP) positivo, 16 negativo y 2 nulo. Es claro de manera inmediata, que la precisión mejora al incorporar BoC al VSM, tanto en MAP como en los diferentes cortes desde 5 a 30 documentos. Según el criterio expresado por Sparck Jones en (Buckley y Voorhees 2000), "Una diferencia entre dos

Tabla 5.5: Estadísticas de CACM y NPL después de eliminar palabras vacías, truncarlas y eliminar términos compuestos de frecuencia uno

Característica	CACM	NPL
No. de documentos	3204	11429
No. de consultas	50	93
No. de términos simples	4997	7714
No. de términos compuestos	1195	4098
No. de relevantes	792	2079
No. de relevantes recuperados	692	1888

Tabla 5.6: Comparación del VSM y los resultados al sumarle BoC (CACM)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.2727	0.2986	0.36	0.294	0.2587	0.23	0.1953
VSM+BoC	0.2974	0.3067	0.388	0.334	0.2933	0.261	0.2113
% Cambio	9.06	2.71	7.78	13.61	13.37	13.48	8.19

ejecuciones que es mayor del 5 % es notable y una diferencia más grande del 10 % es material". Encontramos diferencias materiales en P@10, P@15 y P@20 documentos. De las 32 consultas cuya AP se incrementó, 22 tuvieron un incremento mayor del 10 % y 7 del 5 %.

En la Tabla 5.7 se muestra el efecto de agregar los términos compuestos (CT) codificados como HRRs al VSM y posteriormente al VSM+BoC. El promedio de términos compuestos por documento es 3 y por consulta 1 [2].Puede observarse que al agregar los CT al VSM aunque la precisión se incrementa el cambio es pequeño. Sin embargo, al combinar las tres representaciones encontramos que los cambios en precisión, a excepción del R-Prec y precisión a 5 documentos, están arriba del 10 % y estos dos últimos arriba del 5 %. Indagando para conocer con exactitud el efecto de los HRRs en la recuperación, encontramos que de las 50 consultas de la colección sólo 32 tienen al menos un término compuesto. De esas 32, sólo 15 exhiben cambios en

[2] Estos números incluyen los términos compuestos con frecuencia igual a uno.

Tabla 5.7: Comparación del VSM y los resultados al sumarle BoC y CT (CACM)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.2727	0.2986	0.36	0.294	0.2587	0.23	0.1953
VSM+CT	0.283	0.3012	0.368	0.298	0.2653	0.234	0.198
% Cambio	3.78	0.87	2.22	1.36	2.55	1.74	1.38
VSM+BoC+CT	0.3091	0.3253	0.392	0.346	0.2987	0.262	0.2113
%Cambio	13.35	8.94	8.89	17.69	15.46	13.91	8.19

precisión al añadir los CT y los restantes mantienen el MAP obtenido con el VSM, porque no existe similitud entre sus codificaciones HRR y la de los documentos de la colección. La Tabla 5.8 muestra los resultados para las 15 consultas mencionadas, 8 de la cuales tienen cambio positivo y 7 negativo, sin embargo los cambios positivos son de mayor magnitud. De tal manera que al agregar los CTs al VSM, se obtiene un cambio en MAP del 13 % y al integrar las tres representaciones del 18 %.

5.3.4 Análisis cualitativo

En esta sección se analizan algunas consultas a detalle con el objetivo de entender cómo actúan las representaciones propuestas sobre el ordenamiento de los documentos producido por la recuperación de información.

Bolsa de Conceptos. En la Figura 5.5 se muestra el MAP de seis consultas después de añadir BoC, las tres con el mayor porcentaje de cambio positivo y las tres con el menor porcentaje negativo. En la consulta 48 con un 120 % de cambio y 12 documentos relevantes, se encontró que a 20 documentos BoC ubica 5 documentos relevantes en contraste a 2 ubicados en ese rango por el VSM. Para esta consulta, los documentos relevantes 1729, 1862 y 2223 se colocaron en las posiciones 11,12 y 13, a diferencia de sus posiciones en el VSM de 41, 42, y 43. Algo curioso fue que al revisar el contenido de los documentos relevantes mencionados, éstos tienen el mismo contenido:

Minit algorithm for linear programming (algorithm 333 [h]).

razón por la cual, los vectores de contexto de los términos presentes en ellos enfatizaron su importancia moviéndolos a posiciones superiores.

Tabla 5.8: Consultas cuyo MAP varió al agregar los CT (CACM)

Consulta	VSM	VSM+ BOC	% Cambio	VSM+ CT	% Cambio	VSM+ BOC+CT	% Cambio
1	0.2089	0.1972	-5.60	0.0625	-70.08	0.0712	-65.92
3	0.1023	0.0964	-5.77	0.1772	73.22	0.1768	72.83
10	0.6262	0.5232	-16.45	0.6174	-1.41	0.5458	-12.84
15	0.1148	0.1314	14.46	0.1039	-9.49	0.1178	2.61
18	0.2368	0.2365	-0.13	0.2953	24.70	0.2981	25.89
19	0.3617	0.3829	5.86	0.3701	2.32	0.385	6.44
20	0.0472	0.0528	11.86	0.0453	-4.03	0.0475	0.64
22	0.5832	0.6033	3.45	0.5624	-3.57	0.587	0.65
25	0.2261	0.3086	36.49	0.2314	2.34	0.2988	32.15
32	0.447	0.4034	-9.75	0.697	55.93	0.6749	50.98
40	0.2388	0.2221	-6.99	0.3244	35.85	0.304	27.30
42	0.0439	0.0649	47.84	0.0416	-5.24	0.0603	37.36
48	0.1027	0.2267	120.74	0.3342	225.41	0.4418	330.19
49	0.1109	0.0898	-19.03	0.0949	-14.43	0.0831	-25.07
63	0.3386	0.3721	9.89	0.3501	3.40	0.4028	18.96
Promedio	0.2526	0.2608	3.23	0.2872	13.69	0.2997	18.63

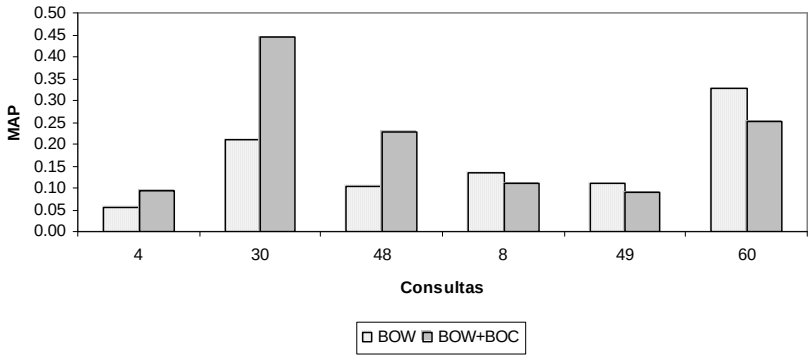

Figura 5.5: Consultas con mayor beneficio y deterioro al agregar BoC

La consulta 30 cuyo porcentaje de cambio es 112 % tiene sólo 4 documentos relevantes. BoC mueve dos documentos relevantes a las primeras posiciones 1, 3, mientras

que en el VSM aparecen en las posiciones 2 y 10. Finalmente en la consulta 4 con 12 documentos relevantes, haciendo un corte a 10 documentos, BoC coloca 2 documentos relevantes dentro de este rango y el VSM sólo 1. El último documento relevante, BoC lo coloca en la posición 447 y el VSM en la 793.

En cuanto a las consultas en las que se obtuvo porcentajes de cambio negativo, el cambio en posición de los relevantes no es drástico, por ejemplo para la consulta 8 con 3 documentos relevantes con el VSM, se ubican en las posiciones 3, 46 y 97 mientras que al agregar BoC en las 4, 45 y 83, mejorando las posiciones de dos de ellos. En la consulta 49 con 7 documentos relevantes, a 20 documentos el BoC sólo hace bajar un documento a la posición 22, y a 30 ambas aproximaciones tienen 4 documentos relevantes. Por último para la consulta 60 con 23 documento relevantes, a 20 y 30 documentos la diferencia entre las dos aproximaciones es de sólo 1 documento relevante.

Representaciones Holográficas Reducidas. Para ejemplificar las consultas que resultaron con MAP desfavorable al aumentar los CT, considérese la consulta 1, cuyos resultados se observan en la Tabla 5.8, y texto a continuación:

What articles exist which deal with TSS (Time Sharing System), an operating system for IBM computers?

Esta consulta tiene 5 documentos relevantes y dos CTs: *oper system e ibm comput*, La Tabla 5.9 muestra el identificador de los documentos relevantes y la posición alcanzada por los mismos para VSM y el VSM+CT. El cambio tan drástico en posiciones al agregar los CTs, se debió a que ninguno de los documentos relevantes tiene CTs en común con la consulta. Ilustrando lo anterior, en la Tabla 5.10 se muestran los 11 documentos colocados en las primeras posiciones por el VSM+CT. Puede observarse como los documentos que comparten CTs con la consulta son promovidos a posiciones superiores. Los documentos 1938, 2371, 1071 y 1572 mantienen posiciones altas por la colocación inicial buena que les da el VSM. Pero definitivamente, el hecho de que los documentos relevantes no contengan CTs en común con la consulta afecta la precisión.

En cuanto a las consultas que resultaron favorables, como ejemplo tenemos la número 48 cuya información aparece en la Tabla 5.8. Esta consulta tiene 12 documentos relevantes y un CT: *linear program*. En la Tabla 5.11 se presentan las posiciones de los documentos relevantes tanto en el VSM, como después de incorporar los CTs. Los 7 primeros documentos relevantes que coloca el VSM+CT son aquellos que comparten con la consulta el mismo CT. Mientras que los 5 últimos no tienen *linear program* entre sus CTs. Nótese como el documentos 2325, a pesar de tener una buena posición en el VSM, por no compartir ningún CT con la consulta es colocado hasta la posición 33.

Tabla 5.9: Posición de documentos relevantes para la consulta 1

VSM		VSM+CT	
Id.Rel.	Posición	Id.Rel.	Posición
1572	3	1572	11
1410	5	1410	21
1605	14	1605	58
2358	45	2358	60
2020	608	2020	642

Tabla 5.10: Primeros 11 documentos para la consulta 1 obtenidos por el VSM+CT

	VSM	VSM+CT	Término compuesto en
Id.Rel.	Posición	Posición	común con la consulta
1938	1	1	Ninguno
2319	6	2	Oper system
1680	15	3	Oper system
1591	31	4	Oper system
1519	30	5	Oper system
2371	4	6	Ninguno
1071	2	7	Ninguno
3068	50	8	Oper system
1472	146	9	Oper system
1264	209	10	Oper system
1572	3	11	Ninguno

Comparación con el trabajo relacionado. Debe tenerse presente que es difícil establecer conclusiones definitivas con respecto a la efectividad del indexado y métodos de recuperación que han sido probados en diferentes laboratorios y diferentes tiempos, ya que las condiciones experimentales pueden diferir significativamente. Las colecciones varían y aun los detalles de los procedimientos ampliamente aceptados

Tabla 5.11: Posición de documentos relevantes para la consulta 48

VSM		VSM+CT	
Id.Rel.	Posición	Id.Rel.	Posición
1797	6	1797	3
2325	8	1353	4
1729	41	1729	5
1863	42	1863	6
2223	43	2223	7
1353	49	1666	12
2589	58	2073	13
2285	128	2325	33
1666	135	2285	34
2073	136	2589	123
0149	164	0149	273
2226	360	2226	453

de indexación, recuperación y evaluación pueden cambiar. Sin embargo, cuando la información está disponible, es ilustrativo comparar resultados experimentales para obtener una idea de la efectividad relativa alcanzada por diferentes métodos de indexado y recuperación, sin perder de vista las limitaciones de este tipo de comparación.

Por otra parte, CACM es una colección creada con fines de investigación, utilizado en numerosos artículos de recuperación de información. Sin embargo, se considerada pequeña para publicaciones actuales. Con fines de comparación recurrimos a trabajos publicados hace ya varios aos.

En la Tabla 5.12 se muestran los resultados reportados por Salton en (Salton 1986) para CACM a 5 puntos de recuerdo indexada con términos simples y términos compuestos. Las dos últimas columnas muestras los resultados obtenidos al codificar los CTs como HRRs. Puede observarse una clara superioridad de la aproximación propuesta.

Fagan en (Fagan 1987) reporta una mejora del 7.6 % para CACM cuando considera los mismos parámetros para todas las colecciones. Utilizando los resultados de precisión promedio que reporta para la indexación con términos simples a 11 puntos

Tabla 5.12: Comparación de los resultados obtenidos con los reportados por Salton

	Resultados reportados por Salton			Resultados con HRRs	
Recuerdo	Términos simples	Términos compuestos	% Cambio	Término compuestos	% Cambio
0.1	0.5086	0.5427	6.70	0.5966	17.30
0.3	0.3672	0.3971	8.14	0.387	5.39
0.5	0.2398	0.2527	5.38	0.2601	8.47
0.7	0.1462	0.1462	0.00	0.1625	11.15
0.9	0.0711	0.0759	6.75	0.0729	2.53
Promedio	0.2666	0.2829	6.13	0.2958	10.97

de recuerdo con un valor de 0.2724, comparada con la obtenida por nuestra representación de 0.3004 obtenemos un 10.28 % de mejora.

Finalmente para estos experimentos, y los realizados con la colecciones de las siguientes secciones, se efectuó la *student's t-test* para medir cuán significativos eran los resultados, respecto al cambio en MAP obtenido.

Para CACM, el cambio global en MAP para VSM+BoC y VSM+BoC+CT resultó significativo en un intervalo de confianza de 99 %.

5.4 Colección NPL

NPL es una colección con documentos referentes a Ingeniería Eléctrica, desarrollada en el National Physical Laboratory en el Reino Unido. La Tabla 5.5 muestra información general de esta colección. De las 93 consultas, se tomaron 92 que son las que cuentan con documentos relevantes. Además NPL en promedio tiene 22 palabras por documento, 6 por consulta y 22 documentos relevantes por consulta.

Bolsa de Conceptos. La Tabla 5.13 muestra la comparación entre el VSM y los resultados al añadir la similitud obtenida con BoC. De las 92 consultas examinadas, 76 tuvieron un porcentaje de cambio en precisión promedio (AP) positivo, 15 negativo y 1 nulo. BoC evidentemente mejora la precisión tanto en MAP, R-Prec y los diferentes cortes desde 5 a 30 documentos. De las 76 consultas cuya AP se incrementó, 52 tuvieron un incremento mayor del 10 % y 13 del 5 %.

Representación Holográfica Reducida. En la Tabla 5.14 se muestra la variación

Tabla 5.13: Comparación del VSM y los resultados al sumarle BoC (NPL)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.2037	0.2278	0.3174	0.2652	0.2304	0.213	0.1812
VSM+BoC	0.2323	0.2597	0.3739	0.2935	0.2667	0.2397	0.2
% Cambio	14.04	14.00	17.80	10.67	15.76	12.54	10.38

al agregar los términos compuestos (CT) codificados como HRRs al VSM y posteriormente al VSM+BoC. El promedio de términos compuestos por documento es 2 y por consulta menos de 1 (0.7). De la tabla es evidente que los cambios observados son pequeños, aunque favorables en su mayoría. Así por ejemplo, el cambio global en MAP es de 0.2 % y a 5 documentos de 2.74 %. No obstante, al combinar las tres representaciones encontramos que el cambio en precisión en todas las métricas mostradas está arriba del 11 %.

Para esta colección y las restantes el incremento en precisión con la HRR fue menos significativa de los que se esperaba. Por tal razón, a partir de esta colección y las siguientes, se analizaron las consultas cuya precisión se afectó negativamente a fin de identificar las causas.

Analizando la repercusión de la HRR en la recuperación para NPL, encontramos que de las 92 consultas de la colección, sólo 48 tienen al menos un término compuesto. De esas 48, sólo 32 exhiben cambios en precisión al agregar los CT, pues su codificación HRR es similar en grado ponderable a algún documento de la colección. La Tabla 5.15 muestra los resultados para las 32 consultas mencionadas. Puede observarse que aún la aportación de los HRRs a la precisión es marginal. Observando las 32 consultas y analizando las afectadas más negativamente, se encontró que dos grupos de consultas disminuían la precisión. El primero grupo formado por 5 consultas, cuyo información pueden verse en la Tabla 5.16, tiene como característica que las consultas no tienen CTs en común con sus documentos relevantes(caso similar identificado en CACM). En consecuencia, documentos no relevantes que contienen un CT en común con la consulta son promovidos a posiciones superiores disminuyendo la precisión. Por ejemplo, para la consulta 90 que resultó la más afectada por la incorporación de los HRRs, se encontró, analizando sólo las primeras posiciones, que el VSM en los dos primeros lugares coloca a los documentos 6190 y 7011. Cuando los HRRs se agregan, el documento relevante 7011 desciende al lugar 4, y antes de él se colocan los documentos 986 y 4760 que comparten con la consulta el CT *"electrón densiti"*. Los resultados obtenidos al evaluar sólo a 27 consultas, se observan en la Tabla 5.15.

Tabla 5.14: Comparación del VSM y los resultados al añadirle BoC y CT (NPL)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.2037	0.2278	0.3174	0.2652	0.2304	0.213	0.1812
VSM+CT	0.2041	0.2264	0.3261	0.2707	0.2319	0.213	0.1837
% Cambio	0.20	-0.61	2.74	2.07	0.65	0.00	1.38
VSM+BoC+CT	0.2325	0.2611	0.3674	0.2967	0.2688	0.2391	0.204
%Cambio	14.14	14.62	15.75	11.88	16.67	12.25	12.58

Tabla 5.15: MAP en grupos de consultas seleccionadas (NPL)

Número Consultas	VSM	VSM+ BOC	% Cambio	VSM+ CT	% Cambio	VSM+BOC +CT	% Cambio
32	0.2062	0.2233	8.31	0.2073	0.55	0.2239	8.57
27	0.2288	0.2445	6.86	0.2366	3.40	0.2537	10.86
22	0.2316	0.2490	7.54	0.2524	9.01	0.2695	16.39

La aportación de los HRRs de 0.55 % se incrementa a 3.4 %. El segundo grupo, Tabla 5.17, lo constituyen también 5 consultas que comparten CTs con pocos documentos relevantes. En la Tabla 5.15 pueden observarse los resultados de la evaluación a 22 consultas. La contribución de los HRRs a la precisión de un 3.4 % se incrementa a un 9 %.

Para esta colección, el cambio en MAP resultó significativo en un intervalo de confianza del 99 %, tanto para VSM+BoC como para VSM+BoC+CT.

Tabla 5.16: Consultas que no comparten CTs con sus documentos relevantes (NPL)

ID	Términos compuestos	Número de relevantes	% Cambio MAP
52	circuit diagram	25	-38.04
59	digit comput, numer data	1	-31.33
64	electromagnet wave, diffract theori, electron stream	22	-41.34
79	high frequenc	29	-24.41
90	electron densiti, semiconductor compare	29	-56.42

Tabla 5.17: Consultas que comparten pocos CT con sus documentos relevantes (NPL)

ID	Términos compuestos	No. de relevantes	Relevantes con algún CT de la consulta	% Cambio MAP
20	electr conduct, upper atmospher	26	7	-23.70
27	lighth discharge, sourc spectra	28	2	-24.65
35	Ionosper drift	30	2	-24.05
44	digit comput, error check	8	2	-20.80
53	Frequenc rang, magnet amplifi	17	5	-20.68

5.5 Colección ADHOC (TREC)

La colección en inglés del CLEF (Cross Language Evaluation Forum) es una colección de noticias tomadas del Glasgow Herald (GB) 1995 y del LA Times (EUA) 1994. Las noticias cubren tanto eventos nacionales como internacionales. La Tabla 5.18 muestra la cantidad de documentos de cada fuente. Esta colección con 169477 documentos, 215278 términos simples y 245 palabras en promedio por documento, se utiliza en diferentes tareas del CLEF, variando la cantidad y contenido de las consultas. En la Tabla 5.19 se presenta información general de las consultas especificadas para ADHOC. Un ejemplo de las consultas para la colección en inglés del CLEF se muestra en la Figura 5.6, donde pueden observarse tres componentes: título, des-

```
<topic lang = ``en" >
  <identifier>251</identifier>
  <title>Alternative Medicine</title>
  <description>Find documents discussing any kind of alternative or natural medical treatment including
    specific therapies such as acupuncture, homeopathy, chiropractics, or others.</description>
  <narrative>Relevant documents will provide general or specific information on the use of natural or
    alternative medical treatments or practices.</narrative>
</topic>
```

Figura 5.6: Tópico 251 de ADHOC, colección CLEF para el idioma inglés

Tabla 5.18: Colección en inglés del CLEF

Colección	Origen	No. de documentos
GH95	The Glasgow Herald	56472
LAT94	The Los Angeles Times	113005
		Total: 169477

cripción y narrativa. Para ésta, y las siguientes colecciones se utilizaron el título y descripción de las consultas.

Bolsa de Conceptos. La Tabla 5.20 muestra la comparación entre el VSM y los

Tabla 5.19: Estadísticas de la colección CLEF para idioma inglés después de eliminar palabras vacías, truncar las restantes y eliminar términos compuestos de frecuencia 1

Característica	ADHOC	ROBUST	GEO2005	GEO2008
No. de consultas	50	160	25	25
No. de relevantes	2063	4327	1028	747
No. de relevantes recuperados	1573	3512	775	556
No. prom. de palabras por consulta	8	5	9	8
No. prom. de CT por consulta	0.8	0.6	1.4	1.2
No. promedio de relevantes por consulta	42	28	41	29

Tabla 5.20: Comparación del VSM y los resultados al agregarle BoC (ADHOC)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.1969	0.2181	0.284	0.278	0.268	0.265	0.2413
VSM+BoC	0.2075	0.2336	0.316	0.296	0.2733	0.282	0.254
% Cambio	5.38	7.11	11.27	6.47	1.98	6.42	5.26

resultados al añadir la similitud obtenida con BoC. De las 50 consultas examinadas, 36 tuvieron un porcentaje de cambio en AP positivo, 13 negativo y 1 nulo. BoC mejora la precisión de manera notable tanto en MAP, R-Prec y precisión a 5, 10, 20 y 30 documentos. De las 36 consultas cuya AP se incrementó, 20 tuvieron un incremento mayor del 10 % y 8 del 5 %.

Representación Holográfica Reducida. En la Tabla 5.21 se presenta la variación de la precisión al incorporar los términos compuestos. El número de términos compuestos para la colección del CLEF en inglés es de 308942 términos con frecuencia mayor a 1. El promedio de los mismos por documentos es de 20 y para las consulta de ADHOC menos de 1 (0.8). Directamente de la Tabla 5.21 es notorio que los cambios en precisión alcanzados con los HRRs son pequeños, con el de mayor magnitud a 10 documentos de 4.32 %. Al combinar las tres representaciones encontramos cambios notables en MAP, R-Prec, Precisión a 5 y 10 documentos.

Para los HRRs encontramos que de las 50 consultas de la colección sólo 31 tienen al menos un término compuesto. De esas 31, 24 muestran cambios en precisión al añadir los CTs. La Tabla 5.22 muestra los resultados para las 24 consultas aludidas. Puede observarse que los HRRs, de contribuir al cambio en precisión en 0.86 %, ahora lo hacen en un 2.4 %. Examinando las 24 consultas, encontramos que sólo 3 de ellas presentaban cambios desfavorables menores del -10 %. Estas consultas son la 265, 276 y 286, mostradas en la Tabla 5.23. La consulta 286 no comparte CTs con sus documentos relevantes, pertenece a uno de los grupos explicados al analizar NPL. En las otras consultas, se identificó un factor adicional que deteriora la precisión. Para ilustrarlo tomemos la consulta 265. El primer documento relevante *LA021394-0222*, es colocado por el VSM en la posición 14. Al agregar los CTs, este documento también es el primer relevante posicionado, pero ahora en el lugar 22. Revisando los documentos con posiciones superiores a la 22, se encontró que algunos evidentemente fueron promovidos por la presencia del CT: *deutsch bank*. Sin embargo, otros no lo tenían y fueron promovidos. Este hecho se explica porque los documentos mencionados tenían CTs con alguno de los términos simples de *deutsch bank*. Uno de estos documentos

Tabla 5.21: Comparación del VSM y los resultados al añadir BoC y CT (ADHOC)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.1969	0.2181	0.284	0.278	0.268	0.265	0.2413
VSM+CT	0.1986	0.2213	0.284	0.29	0.272	0.267	0.2413
% Cambio	0.86	1.47	0.00	4.32	1.49	0.75	0.00
VSM+BoC+CT	0.2068	0.2314	0.316	0.298	0.2747	0.277	0.2513
%Cambio	5.03	6.10	11.27	7.19	2.50	4.53	4.14

Tabla 5.22: MAP en grupos seleccionados de consultas para ADHOC

Número Consultas	VSM	VSM+ BOC	% Cambio	VSM+ CT	% Cambio	VSM+BOC +CT	% Cambio
24	0.1472	0.1569	6.57	0.1509	2.46	0.1553	5.50
21	0.1531	0.1623	6.02	0.1603	4.71	0.1658	8.33

fue el *GH950417-000066* que, de la posición 67 que le dio el VSM, fue promovido a la posición 21. Este documento tiene 48 CTs y entre ellos están: *jacobit bank, royal bank, scottish bank (2 veces), union bank, ayr bank y western bank*. Todos ellos comparten la segunda parte de su HRR con el CT de la consulta. El término *bank* es sobre enfatizado y en consecuencia, este documento es promovido a una posición superior, disminuyendo la precisión.

En la Tabla 5.22, a 21 consultas, se advierte que la contribución de los HRRs es ahora del 4.71 %, lo que permite incrementar el porcentaje de cambio al integrar las tres representaciones a un 8.33 %.

Para ADHOC el cambio en MAP, tanto para VSM+BoC como VSM+BoC+CT, fue significativo estadísticamente en un intervalo de confianza del 99 %.

Tabla 5.23: Consultas cuya precisión disminuye notablemente en ADHOC

ID	Términos compuestos	No. de relevantes	Relevantes con algún CT de la consulta	% Cambio MAP
265	underway takeov, deutsch bank	6	3	-33.58
276	agricultur subsidi, european union	45	14	-15.93
286	footbal soccer, profesion footbal	12	0	-10.87

5.6 Colección ROBUST (CLEF)

Para la tarea de ROBUST en el CLEF, se definieron 160 consultas, en promedio tienen 5 palabras y 28 documentos relevantes. Otros datos de las consultas pueden leerse en la Tabla 5.19. De las 160 consultas definidas para ROBUST, sólo 153 tienen documentos relevantes, con este grupo se probaron las representaciones propuestas.

Bolsa de Conceptos. La Tabla 5.24 muestra tanto los resultados para el VSM, como los obtenidos al adicionarle los generados con la representación BoC. De las 153 consultas investigadas, 109 tuvieron un porcentaje de cambio en AP positivo, 37 negativo y 7 nulo. De las 109 consultas cuya AP se incrementó, 65 tuvieron un incremento mayor del 10 % y 19 del 5 %. BoC mejora la precisión, demostrado por las métricas utilizadas, con porcentajes de cambio arriba de 5 % con excepción de precisión a 15 y 30 documentos.

De las tablas 5.20 y 5.24 se advierte que en colecciones mayores, la representación BoC, genera mayor porcentaje de cambio en R-Prec que en MAP. Una explicación que se da a estos resultados es que BoC modela las relaciones de sinonimia, y por lo tanto incrementa el recuerdo en la recuperación. Interpretar y dar seguimientos a los resultados producidos por BoC no es una tarea sencilla, principalmente por la poca información que se puede recabar de la estructura del espacio de palabras creado a través de la indexación aleatoria. Sin embargo, el cambio en MAP al incorporar BoC al VSM en todas las colecciones siempre está arriba del 5 %.

Representación Holográfica Reducida. En la Tabla 5.25 se muestran los resultados obtenidos al agregar los términos compuestos en forma de HRR al VSM y después al VSM+BoC. El promedio de términos compuestos por consulta es menor de 1 (0.6). En la Tabla 5.25 se observa que el cambio en precisión al adicionar los HRRs es nuevamente pequeño y al igual que en las colecciones anteriores, la integración

Tabla 5.24: Comparación del VSM y los resultados al agregarle BoC (ROBUST)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.2247	0.2278	0.2928	0.2464	0.2244	0.2127	0.1871
VSM+BoC	0.2397	0.2434	0.3137	0.2608	0.2344	0.2248	0.1943
% Cambio	6.68	6.85	7.14	5.84	4.46	5.69	3.85

de las tres representaciones produce resultados con cambios en precisión en varias métricas arriba del 5 %.

Las consultas que tienen al menos un término compuesto son 89 y de esas 89, únicamente 63 muestran cambio al agregar los CT. En estas 63 consultas se ubicaron aquellas con una disminución en precisión más baja del -10 %. Nueve fueron las consultas en el caso aludido. De estás 9 consultas 4: la 170, 182, 196 y 322 no comparten ningún CT con sus documentos relevantes, caso explicado en el análisis del primer grupo de consultas de NPL. La consulta 311, con 51 documentos relevantes, sólo con 3 comparte su único CT: *european country*; lo que la clasifica en el segundo grupo de consultas explicado para NPL. Las consultas 265, 276 y 286 son las mismas que en ADHOC, y en consecuencia bajan la precisión por las razones mencionadas en el análisis previo de esta colección. Finalmente, la consulta 339 con 19 documentos relevantes presentó características diferentes, pues todos sus relevantes tenían el único CT: *sinn fein* de la consulta. Analizando la posición de los documentos relevantes producida por el VSM y por el VSM+CT, encontramos que los documentos en el VSM+CT bajaron en promedio 6 posiciones. El primer documento relevante, ambas aproximaciones lo colocan en la posición 2. Pero el segundo, el VSM lo ubica en el lugar 4 y al agregar los CTs desciende a la posición 11. Los documentos relevantes de esta consulta tienen en promedio 3 veces el CT. Los ocho documentos colocados por el VSM+CT antes del segundo documento relevante, que también es el mismo en ambas aproximaciones, tienen en promedio 8 veces el CT. Está es la razón por la cual son trasladados a posiciones superiores, dañando en consecuencia la precisión. La situación descrita sucederá siempre que una consulta tenga algún CT, y la frecuencia de éste sea mayor, de manera sobresaliente, en documentos no relevantes que en los relevantes. En la Tabla 5.26 se muestra la precisión alcanzada considerando 54 consultas. La contribución de los HRRs al cambio en precisión de 0.18 % se incrementa al 2.58 %.

Para ROBUST nuevamente el cambio global en MAP resultó significativo en un intervalo de confianza de 99 %, tanto para VSM+BoC como para VSM+BoC+CT.

Tabla 5.25: Comparación del VSM y los resultados al añadir BoC y CT (ROBUST)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
VSM	0.2247	0.2278	0.2928	0.2464	0.2244	0.2127	0.1871
VSM+CT	0.2251	0.2294	0.2928	0.2529	0.2279	0.2157	0.1874
% Cambio	0.18	0.70	0.00	2.64	1.56	1.41	0.16
VSM+BoC+CT	0.2389	0.243	0.3111	0.2654	0.2344	0.2239	0.1943
%Cambio	6.32	6.67	6.25	7.71	4.46	5.27	3.85

Tabla 5.26: MAP en grupos seleccionados de consultas para ROBUST

Número Consultas	VSM	VSM+ BOC	% Cambio	VSM+ CT	% Cambio	VSM+BOC +CT	% Cambio
54	0.1584	0.1729	9.16	0.1625	2.58	0.1766	11.49

5.7 GeoCLEF 2005 y GeoCLEF 2008

La Recuperación de Información Geográfica (GIR) se interesa en información relacionada con localidades, tales como nombres de ríos, ciudades, lagos y países (Henrich y Ldecke 2007). La información que está ligada a espacios geográficos se conoce como relaciones espaciales o geo-referencias. Por ejemplo, para la consulta "ETA in France", la relación espacial estará representada por la preposición *in*. Ahora bien, las técnicas tradicionales de IR no serán capaces de producir una respuesta efectiva para esta consulta, ya que la información proporcionada por el usuario es muy general. Por lo tanto, los sistemas GIR tienen que interpretar información implícita, contenida en documentos y consultas, para proporcionar respuestas efectivas a las consultas. Dicha información implícita sería necesaria en el ejemplo para relacionar la palabra "France" con otras ciudades francesas como París, Marsella, Lyón, etc.

Desarrollos recientes han demostrado que el problema de GIR es parcialmente resuelto mediante pequeñas variaciones de técnicas tradicionales de IR. Es posible observar que las máquinas tradicionales de IR recuperan la mayoría de los documentos relevantes para la mayoría de las consultas geográficas, pero tienen dificultad para generar un ordenamiento adecuado de los documentos recuperados, lo que conduce a

precisión insatisfactoria (Villatoro-Tello et al. 2008).

Una fuente importante del problema de ordenar los documentos de manera apropiada, es la falta de información. Por lo tanto, diversas investigaciones en GIR han tratado de llenar esta carencia utilizando recursos geográficos robustos (por ejemplo ontologías geográficas o reconocedores de entidades nombradas geográficas), mientras que otras investigaciones han utilizado técnicas de retroalimentación de relevancia. Por otro lado, resultados recientes de evaluación en GIR indican que no hay una ventaja notable entre estrategias basadas en recursos y métodos que no dependen de recursos geográficos (Villatoro-Tello et al. 2008).

Esta panorámica motivó la curiosidad de saber cuál sería el efecto de aplicar las representaciones propuestas a colecciones de GIR, como son GeoCLEF 2005 y 2008. Por un lado, la representación de BoC podría claramente capturar la información implícita en documentos y consultas para mejorar la precisión y por otro los HRRs podrían utilizarse adicionalmente para representar relaciones espaciales como "in France".

Así que, con estas colecciones, se realizaron tres grupos de experimentos:

(a) Mismos experimentos que para las colecciones anteriores

(b) Generación inicial de resultados con Lemur y reordenamiento de estos resultados con BoC, para compararlos con la técnica de retroalimentación de relevancia

(c) Codificación de la relaciones espaciales con HRRs

Bolsa de Conceptos. La Tabla 5.27 muestra tanto los resultados obtenidos con el VSM, como los obtenidos al complementarlos con la representación BoC. De las 25 consultas investigadas para 2005, 19 tuvieron un porcentaje de cambio en AP positivo, 6 negativo. Para el 2008, fueron 17 positivo, 5 negativo y 3 nulo. Sin lugar a dudas BoC, como en todas las colecciones anteriores, mejora la precisión, con porcentajes de cambio en MAP siempre arriba del 5 %, en 2005 se obtiene hasta un 10 % de cambio. Las otras mediciones de precisión son también sobresalientes con mejoras de hasta un 15 % o 14 %, en el mejor de los casos.

Representación Holográfica Reducida En la Tabla 5.27 se muestran los resultados obtenidos al agregar los términos compuestos en forma de HRRs al VSM+BoC. En 2005, el promedio de términos compuestos por consulta fue 1.4, y en 2008 fue 1.2.

En la misma tabla, se observa que el cambio en MAP al adicionar los HRRs en el 2008 a nivel global es de casi 2 % adicional a lo ya incrementado por BoC. Sin embargo, en la colección del 2005 los HRRs decrecen la precisión. Efectuando el mismo análisis que en las colecciones previas, en 2005 sólo 21 consultas tienen al menos un CT, de esas 21 sólo 12 comparten algún CT con documentos de la colección y de ellas, dos consultas presentan una caída en precisión más allá del -10 %. Estás son la 13 y la 17.

Tabla 5.27: Comparación del VSM y los resultados al agregarle BoC y CT (GeoCLEF 2005 y 2008)

	MAP	R-PREC	P@5	P@10	P@15	P@20	P@30
GeoCLEF 2005							
VSM	0.171	0.2112	0.264	0.248	0.2293	0.22	0.1867
VSM+BoC	0.1883	0.2283	0.304	0.276	0.2613	0.236	0.2
% Cambio	10.12	8.10	15.15	11.29	13.96	7.27	7.12
VSM+BoC+CT	0.1814	0.221	0.32	0.28	0.2533	0.228	0.2027
% Cambio	6.08	4.64	21.21	12.90	10.47	3.64	8.57
GeoCLEF 2008							
VSM	0.1473	0.1665	0.216	0.164	0.152	0.14	0.124
VSM+BoC	0.1554	0.1839	0.224	0.188	0.1707	0.15	0.14
% Cambio	5.50	10.45	3.70	14.63	12.30	7.14	12.90
VSM+BoC+CT	0.1581	0.1856	0.232	0.188	0.1733	0.15	0.1413
% Cambio	7.33	11.47	7.41	14.63	14.01	7.14	13.95

Analizando su contenido encontramos que son casos semejantes a la consulta 339 de ADHOC. La consulta 13 tiene 7 documentos relevantes. Con tres de ellos comparte el CT: *president clinton*. La frecuencia de esta frase, en documentos no relevantes, es mucho mayor que en los documentos relevantes, en los cuales su frecuencia es igual a 1. Por otra parte, la consulta 17 tiene 129 documentos relevantes, con 77 comparte el CT: *bosnia herzegovina*, la frecuencia promedio de este CT en los documentos relevantes es 1, mientras que en otros documentos su frecuencia es mucho mayor. Este fenómeno ocasiona que documentos irrelevantes sean colocados en posiciones superiores a los documentos relevantes y por ende la precisión baje.

En el 2008, 20 consultas tienen al menos un término compuesto, de ellas sólo 16 tienen cambio en precisión al añadir los CTs, y de estas últimas, solamente dos consultas presentan decrementos más allá de -10 %. La consultas aludidas son la 78 y 99, con términos compuestos *sport event y european citi* respectivamente, que no aparecen en ninguno de sus documentos relevantes. La consulta 78 tiene 12 documentos relevantes y la 99 tiene 85.

Los resultados a 10 consultas en el 2005 y 14 en el 2008 se muestran en la Tabla

Tabla 5.28: MAP en grupos seleccionados de consultas para GeoCLEF

Número Consultas	VSM	VSM+ BOC	% Cambio	VSM+ CT	% Cambio	VSM+BOC +CT	% Cambio
GeoCLEF 2005							
10	0.1285	0.1474	14.72	0.1365	6.24	0.1520	18.29
GeoCLEF 2008							
14	0.1617	0.1750	8.24	0.1678	3.78	0.1801	11.40

5.28. Puede ahora observarse el buen desempeño de los HRRs, permitiendo porcentajes de cambio en MAP que se pueden calificar de materiales.

Para GeoCLEF 2005, las pruebas estadísticas mostraron que los cambios en MAP para VSM + BoC son significativos en un intervalo de confianza del 97 % y para VSM + BoC + CT en un intervalo del 90 %.

En cuanto a GeoCLEF 2008, los cambios en MAP para VSM+BoC y VSM + BoC + CT son estadísticamente significativos en un intervalo de confianza del 90 %.

Reordenamiento de documentos con BoC y HRRs. En este grupo de experimentos, como línea base de comparación se utilizaron los resultados generados con Lemur [3]. Dicha herramienta utiliza modelos probabilísticos del lenguaje, para mejorar la precisión del VSM. De los resultados de Lemur, para cada consulta, se tomaron los 1000 primeros documentos. Cada grupo de mil documentos se indexó con RI, lo que implicó la generación de 50 indizaciones diferentes. Después se generaron las representaciones BoC de documentos y consultas y se calculó su similitud. Las similitudes obtenidas se sumaron a las generadas por Lemur, no sin antes multiplicarlas por un factor de un cuarto, para producir Lemur + BoC. Posteriormente se construyeron las representaciones HRRs de cada grupo de 1000 documentos, utilizando los IV de las indexaciones anteriores. Las similitudes entre documentos y consultas HRR, se multiplicaron por un factor de un sexto y se agregaron a las ya obtenidos con Lemur + BoC. Las relaciones textuales que se codificaron como HRRs, fueron las relaciones espaciales (SRs) de las consultas y documentos. Para identificar la SRs, todos los documentos y consultas, se etiquetaron con el sistema para reconocer entidades nombradas de la Universidad de Stanford [4]. Hecho esto, las entidades identificadas como

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

[4]Named Entity Recognition System of Stanford University

http://npl.stanford.edu/software/CRF-NER.shtml

localidad precedidas por la preposición *in* se extrajeron. Esta restricción fue tomada después de analizar las consultas para cada año y notar que sólo el 12 % de ellos tenía alguna SR diferente (por ejemplo: around, across, near, etc.). Así los HRRs para documentos y consultas se produjeron, generando un HRR de dimensión 4096 para representar a la preposición *in*. El vector HRR de *in*, se unió a las localidades identificadas empleando la convolución circular. Por ejemplo, para representar la relación espacial $R = in\ Asia$ se utiliza el vector índice de *Asia*, digamos r_1 que se une a su rol de localidad, un HRR, rol_1, que representa la relación *in*. De esta manera, R será representada como el vector:

$$\mathbf{R} = (\mathbf{rol_1} \otimes \mathbf{r_1}) \tag{5.4}$$

Es importante mencionar que para estas consultas, además de considerar el título y descripción de las consultas, se incluyó la narrativa de las mismas. Los resultados generados con Lemur, fueron inferiores al incluir la narrativa de la consultas, así que para Lemur se incluyo sólo el título y descripción.

En la Tabla 5.29 se muestran los resultados obtenidos para ambos años, puede observarse que BoC incrementó el MAP teniendo cambios en precisión arriba del 5 %, sin embargo las relaciones espaciales, a pesar de que incrementan la precisión, no lo hacen de manera importante.

En la Tabla 5.30 se muestran 10 consultas para el 2008 cuyo MAP cambió favorablemente al agregar BoC. De estas consultas, aquellas que fueron mejoradas después de agregar los HRRs, en promedio tienen 2 relaciones espaciales. Nuestra conclusión es que la baja contribución de los HRRs para mejorar la precisión, se debió a la limitada cantidad de relaciones espaciales que aparecen en el conjunto de consultas (1 o menos), creemos que si se contara con un número mayor de SR para representar, la contribución de esa representación a la precisión será mayor. En el análisis realizado encontramos que los HRRs mejoran la precisión cuando existen SRs distintivas y específicas, por ejemplo *in Finland* en lugar de *in nothern Europe*. Por lo tanto, cuando la información geográfica es más precisa, los HRRs ayudan a mejorar la efectividad.

Efectuamos la *student's t-test* para medir la importancia estadística de los resultados, encontrando que para el 2005 los resultados son significativos, con respecto al cambio en MAP, en un intervalo de confianza del 99 %, tanto para Lemur+BoC como para Lemur+BoC+SR. Al comparar estos resultados con los obtenidos por los equipos participantes en el CLEF2005, encontramos que, nuestra propuesta está por arriba de la media en MAP que fue de 0.26 (Gey et al. 2006). Para el 2008 la media de los participantes fue de 0.2370, así que nuestra propuesta está por arriba de ella en 6.45 %. Sin embargo, el cambio en MAP no resultó significativo estadísticamente. En este año, el equipo participante del CLEF que ocupó la posición más alta obtuvo un MAP de 0.3040 (Mandl et al. 2008)(Wang y Neumann 2008). Ellos utilizaron

Tabla 5.29: MAP resultante del reordenamiento de documentos en GeoCLEF

	Lemur	Lemur+BOC	%Cambio	Lemur+BOC+SR	%Cambio
GeoCLEF2005	0.3191	0.3530	10.62	0.3529	10.59
GeoCLEF2008	0.2347	0.2514	7.12	0.2523	7.50

dos ontologías construidas manualmente, empleando información de las narrativas. Además utilizaron Wikipedia en el proceso de recuperación. En contraste, nosotros no utilizamos ningún recurso externo especializado.

Finalmente comparamos los resultados de Lemur+BoC+SR con un método de reordenamiento tradicional como es la retroalimentación de relevancia (PRF). Para generar el ordenamiento inicial de los documentos, utilizamos el VSM, representando las consultas como vectores *tf.idf* y calculando su similitud con el coseno. La Tabla 5.31 presenta los resultados cuando los 2, 5 y 10 documentos, en las posiciones más altas del ordenamiento, son considerados como relevantes. Las consultas se construyeron con los campos de título y descripción.

Los valores que mejoran Lemur se muestran en negritas y los obtenidos con nuestra propuesta en cursivas. La diferencia en MAP entre la técnica de PRF y nuestros resultados con Lemur+BoC+SR es 9.9 % o más alta en el 2005 y 1.24 o mucho más alta en el 2008.

Tabla 5.30: MAP para consultas mejoradas por BoC en el 2008 y sus relaciones espaciales

ID consulta	Lemur	Lemur +BoC	% Cambio	SR	Lemur +BoC+SR	% Cambio adicional
76	0.4400	0.4857	10.39	12	0.5000	2.94
80	0.2518	0.2555	1.47	1	0.2555	0
82	0.0005	0.0015	200	3	0.0018	20
84	0.1385	0.2183	57.62	0	0.2183	0
85	0.4554	0.4767	4.68	0	0.4767	0
86	0.0592	0.1101	85.98	2	0.1130	2.63
91	0.0625	0.1667	166.72	1	0.1667	0
93	0.7375	0.8340	13.08	1	0.8340	0
95	0.4910	0.5320	8.41	6	0.5337	0.26
96	0.2232	0.2418	8.33	11	0.2454	1.49

Tabla 5.31: Diferencia entre el MAP de PRF y de Lemur+BoC+SR

	2 documentos			5 documentos			10 documentos		
	5 term	10term	15term	5 term	10term	15term	5 term	10term	15term
GeoCLEF2005									
0.3529	0.3034	0.2693	0.2502	0.2943	0.2989	0.2937	**0.3211**	0.3050	0.3082
% Dif.	16.32	31.04	41.05	19.91	18.07	20.16	9.90	15.70	14.50
GeoCLEF2005									
0.2523	**0.2437**	**0.2492**	**0.2405**	0.2214	0.2152	0.2017	0.2064	0.2025	0.2209
% Dif.	3.53	1.24	4.91	13.96	17.24	25.09	22.24	24.59	14.21

5.8 Reordenamiento en ADHOC y ROBUST

Los resultados del reordenamiento con BoC en GeoCLEF 2005 fueron muy semejantes a los obtenidos indexando todos los documentos de la colección a la vez,

teniendo en el 2008, sin embargo, una diferencia aproximada del 2 %. Para obtener una mejor idea de la utilidad de la representación de BoC como herramienta de reordenamiento, realizamos experimentos con los conjuntos de consultas más amplios que teníamos: ADHOC y ROBUST.

El proceso para efectuar el reordenamiento fue el mismo que el descrito para las consultas de GeoCLEF. Para estas colecciones, las similitudes generadas por Lemur estaban en los rangos de [81.33, 5.24] para ADHOC y [227,1.55] para ROBUST. El valor de las similitudes generadas por BoC o HRR siempre está entre 0 y 1. Así que para observar su efecto, se normalizaron los resultados obtenidos por Lemur. Al analizar la dispersión de los resultados generados por Lemur, nuestra conclusión fue que siempre que la desviación estándar de los mismos sea mayor de 1, los resultados deben normalizarse. Esta conclusión se deduce de la información presentada en la Tabla 5.32. Normalizar los resultados de las colecciones cuya desviación estándar no es mayor de uno, decrece ligeramente la precisión (menos del 1 %). En la Tabla 5.33 se muestran los resultados obtenidos para ADHOC y ROBUST donde se observan cambios importantes en MAP, ambos por arriba del 10 %. Este cambio resultó estadísticamente significativo en un intervalo de confianza del 99 %.

En la Figura 5.7 se muestran las 4 colecciones cuyos documentos se reordenaron con BoC, se observa el incremento en precisión obtenido para las cuatro colecciones, comprobándose que BoC junto con la indexación aleatoria son una herramienta útil para reordenar documentos. Además estos experimentos dan evidencia que RI y BoC pueden escalarse. Otros métodos que reducen el espacio vectorial original, sólo son útiles para colecciones pequeñas, por el alto costo de procesamiento que requieren. El tiempo necesario para indexar los 169477 documentos de la colección fue de 72.13 minutos. Guardar los datos en disco tomó 2 horas y la verificación de la generación de vectores índice 63 segundos. El proceso completo tomó aproximadamente tres horas. La generación de los vectores de documentos para toda la colección con BoC tomó 19 horas y para los HRRs tomó 28 horas. La recuperación de los documentos, por ejemplo para todas las consultas de ADHOC tomó aproximadamente 3.5 minutos, es decir 4.2 segundos en promedio por consulta.

Contestando a las preguntas formuladas al inicio de este capítulo tenemos que: El efecto de incluir representaciones BoC y HRR en la recuperación fue el incremento de la precisión con cambios positivos en MAP, que van de 5 % al 14 %. Estos cambios fueron estadísticamente significativos en un intervalo de confianza del 99 % para CACM, NPL, ADHOC y ROBUST. Para GeoCLEF 2005, lo fueron en un intervalo del 97 % y para Geo2008 en un intervalo del 90 %.

Encontramos varias situaciones en las que incorporar HRRs a la recuperación afectó negativamente a las consultas, estas situaciones fueron: pocas relaciones a ser

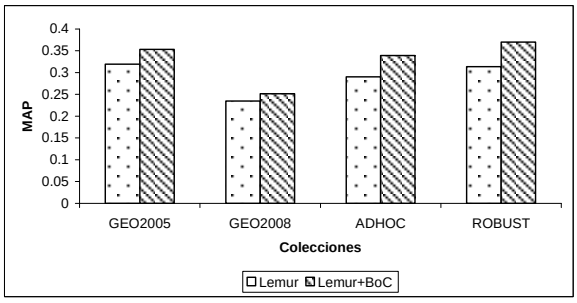

Figura 5.7: Lemur+BoC como herramienta de reordenamiento

Tabla 5.32: Dispersión de los resultados obtenidos por Lemur

Colección	Media	Desviación estándar	Normalizar con $\sigma > 1$
GeoCLEF 2005	-5.3779	0.6895	No
GeoCLEF 2008	-5.9149	0.7891	No
ADHOC	13.3544	6.2166	Si
ROBUST	15.6093	7.0729	Si

representadas, falta de relaciones en los documentos relevantes, mayor porcentaje de relaciones en los documentos no relevantes que en los relevantes, y sobre-acentuación de un sólo aspecto de las consultas.

Por otra parte, al analizar las consultas y omitir las que presentaban alguna situación problemática de las descritas en el párrafo anterior, la contribución de los HRRs a la precisión fue notoria. Por lo tanto, bajo condiciones adecuadas los HRRs contribuyen a mejorar la precisión de la recuperación.

Tabla 5.33: MAP resultante del reordenamiento de documentos con BoC en ADHOC y ROBUST

Colección	Lemur	Lemur+BoC	% Diferencia
ADHOC	0.2903	0.3392	16.84
ROBUST	0.3133	0.3701	18.13

5.9 Midiendo la Robustez

Un sistema de IR requiere de la interacción de diferentes componentes como esquemas de ponderación y métodos de procesamiento de lenguaje natural. Así, la efectividad para una consulta no puede predecirse, dado que cada componente tienen un impacto diferente de acuerdo a la colección, el lenguaje o el contexto del usuario (Mandl 2009). Sin embargo es deseable que un sistema de IR sea capaz de funcionar correctamente bajo diferentes condiciones (consultas, colecciones usuarios, idiomas, etc.) produciendo resultados aceptables para el usuario. Cuando se trabaja con un conjunto de consultas con diferente grado de dificultad, lo óptimo sería que cada consulta contribuyera en igual manera a la efectividad global del sistema. Desde esta perspectiva el inconveniente de utilizar el MAP como medida para las consultas que tienen desempeños bajos, es que cambios en los valores de las consultas que tienen desempeños altos, ocultan los cambios en ellas (Voorhees 2005). En el 2003 inició la tarea de Robust en el TREC cuyo objetivo es el desarrollo de métodos para mejorar la consistencia de la recuperación y una de las métricas utilizadas para esta tarea es la media geométrica de la precisión promedio obtenida para cada consulta (GMAP) (Ravana y Moffat 2008).

La tabla 5.34 muestra el GMAP para el VSM y la obtenida al utilizar las representaciones propuestas. Como puede observarse en todos los casos los porcentajes de cambio son favorables a nuestra propuesta, al incluir BoC+CT siempre por arriba del 10 % y con BoC sólo en dos casos por debajo del 10 % pero arriba del 5 %.

Tabla 5.34: GMAP para el VSM y las representaciones propuestas

Colección	VSM	VSM+BoC	% Cambio	VSM+BoC+HRR	% Cambio
NPL	0.142	0.1700	19.72	0.1674	17.89
CACM	0.203	0.2300	13.30	0.2347	15.62
ADHOC	0.0945	0.1066	12.80	0.1063	12.49
ROBUST	0.098	0.1068	8.98	0.1088	11.02
GeoCLEF 2005	0.0908	0.1058	16.52	0.1043	14.87
GeoCLEF 2008	0.0329	0.0358	8.81	0.0374	13.68

5.10 Resumen

En este capítulo se explicaron dos propuestas de representación para documentos, previas a la establecida como definitiva en este trabajo. Después se mostró como se determinaron los valores de los parámetros para RI y para el esquema de ponderación.

Posteriormente se presentaron los experimentos realizados en seis colecciones al complementar el VSM con las representaciones BoC y HRR. La representación BoC resultó sin lugar a dudas apropiada para mejorar la precisión en recuperación de información con porcentajes de cambio del 5 % al 14 % con respecto al VSM. Los resultados obtenidos al utilizar la HRR resultaron prometedores, especialmente en grupos de consultas que cuentan con las condiciones apropiadas para beneficiarse por del uso de esta representación. Sin embargo, aunque su utilidad se mostró en grupos selectos de consultas, el escaso número de CTs a ser representados (menos de uno por consulta en varias colecciones), no permitió mostrar su utilidad de manera global.

Además se presentaron experimentos de reordenamiento de documentos con RI y BoC, obteniendo resultados por demás satisfactorios para ADHOC, ROBUST y GeoCLEF. Los cambios en precisión con respecto al MAP del VSM van de un 7 % obtenido para GeoCLEF 2008 a un 18 % obtenido para ROBUST. Los resultados obtenidos para las colecciones de GeoCLEF, se compararon con uno de los métodos tradicionales para reordenar documentos como es el PRF. Para GeoCLEF 2005, dependiendo de los n documentos y k términos considerados, los porcentajes de cambio en MAP, favorables a nuestra propuesta, van del 9 al 41 % y para GeoCLEF 2008 del 1.24 al 25 %.

Finalmente, BoC mejora el MAP ya sea utilizada de manera directa o como método de re-ordenamiento. Los cambios en MAP que produjo, van del 5 % al 10 % utilizada de manera directa, y de 7 % a 18 % utilizada como método de reordenamiento en las cuatro colecciones consideradas. Se observó que su desempeño como técnica de reordenamiento mejora entre más grande sea la colección y el número de consultas.

6

Conclusiones y Trabajo Futuro

En este capítulo se enuncian las conclusiones de la presente investigación, sus principales aportaciones, las publicaciones obtenidas y finalmente el trabajo futuro y líneas de investigación abiertas.

6.1 Conclusiones

- Los términos compuestos parecen ser de mayor utilidad en colecciones con terminología común es decir con documentos referentes a una área única del conocimiento. En CACM con documentos exclusivamente del área de computación, la aportación de los términos compuestos es mayor que en la colección del CLEF con noticias de diversos temas.

- El beneficio de incorporar términos compuestos a la recuperación decrece en la medida en que mejoran los resultados obtenidos con términos simples. Los términos compuestos enfatizan aspectos particulares de la consulta, si estos aspectos resultan ser los más importantes entonces la incorporación de términos compuestos contribuye a mejorar la precisión. Por el contrario al enfatizar aspectos secundarios (consultas más generales), los términos compuestos pueden tender a afectar la precisión.

- La representación de BoC demostró ser un mecanismo satisfactorio para mejorar la recuperación de información. Empleada como medio de reordenamiento produjo resultados por arriba de los generados con un método tradicional, como

el de retroalimentación de relevancia. Aunque establecer de manera precisa la manera en que captura la semántica implícita de los documentos se complica por el desconocimiento de las propiedades del espacio vectorial que genera RI, el evento reportado en CACM donde la precisión de una consulta aumenta de manera notable debido a que tres documentos relevantes tienen el mismo contenido, permiten comprobar que la información de los contextos está reflejada en la caracterización de los términos.

- La representación BoC empleada como medio de reordenamiento, entre más grande es la colección, produce mayores incrementos en MAP que utilizada de manera directa.

- La HRR sirve para representar relaciones textuales, como se demostró en la parte experimental. Sin embargo el inconveniente principal para establecer de manera concluyente su utilidad fue la falta de relaciones en las consultas utilizadas. Evidentemente, si el número de relaciones es bajo no sólo en las consultas, sino en el conjunto de documentos, la representación no podrá aportar mejoras a la búsqueda.

- Los HRR podían ser de utilidad de mayor utilidad en colecciones de dominio restringido como por ejemplo colecciones de documentos médicos en los que se necesite buscar medicamentos apropiados para ciertos padecimientos, este tipo de relaciones podrían codificarse como HRR, pudiendo mostrar mejor su beneficio en la recuperación de información.

- Los vectores índice mostraron un mejor acoplamiento que los vectores de contexto para generar los HRRs orientados a la tarea de recuperación de información (apéndice A).

- La utilización de espacios vectoriales separados para representar diferentes tipos de información textual: léxica, sintáctica y semántica y posteriormente integrar los resultados obtenidos de manera separada, resultó de mayor beneficio para recuperación de información, que representar en un sólo espacio información de más de un tipo (apéndice A). También resultó más ventajoso que hacer crecer la dimensión de un sólo espacio vectorial (Tabla 5.1).

- Las representaciones propuestas incrementan la robustez del VSM, según la evaluación realizada en términos de GMAP, sin embargo se necesita un análisis detallado para establecerlo como definitivo.

- BoC como método para capturar semántica implícita tiene ventajas sobre LSI, menor tiempo de procesamiento e indexación progresiva (la indexación se realiza a medida que crece la colección). Por lo tanto estás ventajas mejoraran el proceso de recuperación en general cuando se trabaje con colecciones grandes y cambiantes.

6.2 Aportaciones

(a) Se propuso la utilización de dos técnicas como son la indexación aleatoria (Sahlgren 2005) y la representación de bolsa de conceptos (Sahlgren y Cöster 2004) en recuperación de información. Se comprobó que la indexación aleatoria, a diferencia de otros métodos que reducen el espacio vectorial, puede escalarse y en combinación con la representación de bolsa de conceptos mejoran la precisión de la IR, cuando se utiliza para complementar la representación de BoW.

(b) Utilizando la representación holográfica reducida (HRR) (Plate 2003), perteneciente al área de Ciencia Cognitiva, se estableció un esquema de representación novedosa que es capaz de capturar relaciones sintácticas entre términos (sustantivos compuestos, sujeto-verbo, verbo-complemento, relaciones espaciales), sin incrementar la dimensión vectorial. Esta representación necesita que cada término esté representado como un vector, para tal propósito se utilizó la indexación aleatoria. Para definir los HRRs se utilizaron los vectores índice, a diferencias de la propuesta reportada en (Fishbein y Eliasmith 2008a) que utilizan los vectores de contexto (apéndice A).

(c) Se propuso una modelo de recuperación de información que maneja espacios vectoriales separados para representar información léxica, sintáctica y "semántica" implícita de los textos. Al integrar los resultados de los tres espacios vectoriales la precisión de la recuperación mejora con respecto a la representación de BoW.

6.3 Trabajo futuro

A partir del trabajo experimental realizado se hace evidente que:

- La ponderación dada a la similitud obtenida en los diferentes espacios vectoriales, debe estudiarse de tal manera que pueda determinarse en función de parámetros de la colección como son: vocabulario, número de relaciones, consultas, tipo de documentos.

- La indexación aleatoria es un método escalable que permite reducir el espacio vectorial y de esta manera aplicable a grandes volúmenes de información, sin embargo se necesita una investigación formal para establecer una relación entre los parámetros iniciales para su funcionamiento y las características de la colección sobre la que actuará, los parámetros utilizados en la parte experimental pueden ser no óptimos para otras colecciones.

La propuesta de utilizar los HRRs para codificar relaciones textuales abre un sinfín de posibilidades en el áreas de procesamiento de lenguaje natural, donde su utilidad podría ser de mayor impacto, por mencionar algunas:

- Búsqueda de respuesta: Supongamos que nos interesa contestar la pregunta: *Who was Pilates?*, después de procesarla con un etiquetador de entidades nombradas sabremos que Pilates es una entidad nombrada de tipo persona, información que podrá representarse como un HRR: $\mathbf{C} = \langle \mathbf{per} \otimes \mathbf{Pilates} \rangle$. De esta manera podremos elegir como respuesta a la pregunta textos donde se hable de Pilates como persona y no como método de ejercicios físicos.

- Agrupamiento de documentos, donde incluso la utilización de términos compuestos, codificados como HRR, podría producir mejores resultados al enfatizar aspectos representativos de un grupo de documentos.

- Exploración de recuperación de información en corpus de idiomas diferentes al inglés, ya que la representación HRR es independiente del idioma, sin embargo está ligada a la exactitud con que se identifiquen las relaciones entre términos.

- Búsqueda en información genómica donde relacionar un gen no conocido o una secuencia de proteínas no conocida con secuencias conocidas necesita de búsquedas en estructuras de información organizadas jerárquicamente, aquí sería interesante explorar la utilidad de los HRRs.

6.4 Publicaciones derivadas de la investigación

- Carrilo, M., Villatoro-Tello, E., López-López, A., Eliasmith, C., Montes-y-Gómez, M., Villaseñor-Pineda, L., *Concept Based Representations for Ranking in Geographic Information Retrieval*. In: Advances in Natural Language Processing, 7th International Conference on Natural Language Processing, IceTAL (2010), pp. 85-96 Reykjavik, Iceland, 2010.

- Carrillo, M., López-López, A., *Concept Based Representations as Complement of Bag of Words in Information Retrieval*. In: 6th IFIP Conference on Artificial

Intelligence Applications and Innovations, Larnaca, Cyprus, Octuber 2010 (por aparecer)

- Carrillo, M., Villatoro-Tello, E., López-López, A., Eliasmith, C., Montes-y-Gómez, M., Villaseñor-Pineda, L. *Representing Context Information for Document Retrieval*. In : T. Andreasen et al. (Ed.), Proceedings of the 8th International Conference on Flexible Query Answering Systems, LNAI vol. 5822, pp. 239-250, Roskilde, Denmark, 2009.

- Carrillo, M., Eliasmith, C., and López-López, A.,*Combining Text Vector Representations for Information Retrieval*.In: V. Matousek, P. Mautner (eds) Text, Speech and Dialogue. Proceedings of the 12th International Conference Text, Speech and Dialogue, LNAI, vol. 5729 pp. 24-31 Pilsen, Czech Republic, 2009.

- Carrillo, M., López-López, A., *Towards an Enhanced Vector Model to Encode Textual Relations: Experiments Retrieving Information*. In: IFIP International Federation for Information Processing,vol. 276/2008, pp. 383-392, Springer 2008.

- Carrillo, M., López-López, A., *Incorporación de aspectos semánticos y sintácticos para recuperar información*, Noveno encuentro de investigación, INAOE, Tonantzintla, Pue. pp. 135-138, (2008).

Bibliografía

Achlioptas, D.: 2001, Database-friendly random projections, *In: Procs. of the 20 ACM SIGMOD-SIGCAT-SIGART Symposium on Principles of Database Systems*, pp. 274–281.

Alonso, M. A., Vilares, J. y Darriba, V. M.: 2002, On the usefulness of extracting syntactic dependencies for text indexing, *In: Artificial Intelligence and Cognitive Science*, Vol. 2464, LNA, Springer, pp. 3–11.

Arampatzis, A., der Weide, T. P., van Bommel, P. y Koster, C.: 1999, Linguistically-motivated information retrieval. technical report csi-r9918, *Technical report*, Dept. of Information system, Faculty of Mathematics and Computer Science, University of Nijmegen, Netherlands.

Baeza-Yates, R. y Ribeiro-Neto, B.: 1999, *Modern Information Retrieval*, Addison Wesley.

Brants, T.: 2004, Natural language processing in information retrieval, *In: Procs. of the 14th Meeting of Computational Linguistics in the Net*, pp. 1–13.

Brill, E.: 1994, Some advances in rule-based part of speech, tagging., *In: Procs. of the 12th AAAl*, pp. 722–727.

Buckley, C. y Voorhees, E. V.: 2000, Evaluating evaluation measure stability, *In: Procs. 23rd Ann. Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40.

Chisholm, E. y Kolda, T.: 1999, New term weighting formulas for the vector space method in information retrieval, *Technical report*, Technical report, Oak Ridge Nacional Laboratory.

Croft, W., Turtle, H. y Lewist, D. D.: 1991, The use of phrases and structured queries in information retrieval, *In: Procs. of the 14th Annual International ACM/SIGIR Conference*, pp. 32–45.

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

Doucet, A. y Ahonen-Myka, H.: 2004, Non-contiguous word sequences for information retrieval, *In: Procs. of the 42nd annual meeting of the Association for Computational Linguistics (ACL), Workshop on Multiword Expressions: Integrating Processing*, pp. 88–95.

Eliasmith, C.: 2005, Cognition with neurons: A large-scale, biologically realistic model of the wason task, *In: Procs. of the 27 Annual Conference of the Cognitive Science Society.*

Eliasmith, C. y Thagard, P.: 2001, Integrating structure and meaning: A distributed model of analogical mapping, *Cognitive Science* **25**(2), 245–286.

Evans, D. A. y Zhai, C.: 1996, Noun-phrase analysis in unrestricted text for information retrieval, *In: Procs. of the 34th annual meeting on Association for Computational Linguistics*, pp. 17–24.

Fagan, J. L.: 1987, Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods, *In: Procs. Tenth Annual ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 91–101.

Fishbein, J. y Eliasmith, C.: 2008a, 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, LNCS,Springer, pp. 514–521.

Fishbein, J. y Eliasmith, C.: 2008b, Methods for augmenting semantic models with structural information for text classification, *In: Advances in Information Retrieval: Procs. of the 30th European Conference on IR Research*, Vol. 4956, LNCS,Springer, pp. 575–579.

Gallant, S.: 2000, Context vectors: A step toward a "grand unified representation", *In: Hybrid Neural Systems*, Vol. 1778, LNCS, Springer, pp. 204–210.

Gey, F., Larson, R., Sanderson, M., Joho, H., Clough, P. y Petras, V.: 2006, Geoclef: the clef 2005 cross language geographic information retrieval track overview, *In: 6th Workshop of the Cross-Language Evaluation Forum: CLEF 2005*, pp. 908–919.

Henrich, A. y Ldecke, V.: 2007, Characteristics of geographic information needs, *In: Procs. of Workshop on Geographic Information Retrieval*.

Hinton, G. E., McClelland, J. L. y Rumelhart, D. E.: 1986, Distributed representations, *In: Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, MA., pp. 77–109.

Hofmann, T.: 1999, Probabilistic latent semantic indexing, *In: Procs. of the 22st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57.

Kanerva, P.: 1996, Binary spatter-coding of ordered k-tuples, *In: Procs. of Artificial Neural NetworksICANN*, Vol. 1112, LNCS, Springer, pp. 869–873.

Kanerva, P., Kristofersson, J. y Holst, A.: 2000, Random indexing of text samples for latent semantic analysis, *In: Procs. of the 22nd annual conference of the Cognitive Science Society*, Erlbaum, pp. 103–106.

Karlgren, J. y Sahlgren, M.: 2001, From words to understanding, *In: Uesaka, Y., Kanerva P., Asoh H.: Foundations of Real-World intelligence*, Stanford: CSLI Publications, pp. 294–308.

Kaski, S.: 1998, Dimensionality reduction by random mapping: Fast similarity computation for clustering, *In: Procs. of the IJC on Neural Networks*, pp. 413–418.

Kvasnicka, V.: 2004, Holographic reduced representation in artificial intelligence and cognitive science, *Neural Network World* **14**, 521–532.

Lavelli, A., Sebastiani, F. y Zanoli, R.: 2004, Distributional term representations: an experimental comparison, *In: CIKM 04: Procs. of the thirteenth ACM conference on information and knowledge management*, ACM Press, pp. 615–624.

Lease, M.: 2007, Natural language processing for information retrieval: the time is ripe (again), *In: Procs. of the 1st Ph.D. Workshop at the ACM Conference on Information and Knowledge Management (PIKM)*, pp. 1–8.

Lewis, D. y Sparck, K.: 1996, Natural language processing for information retrieval, *Communications ACM* **39**, 92–101.

Mandl, T.: 2009, Easy task dominate information retrieval evaluation results, *In: Procs. Datenbanksysteme in Business, Technologie und Web (BTW 2009)*, pp. 107–116.

Mandl, T., Carvalho, P., Nunzio, G. D., Gey, F., Larson, R. R., Santos, D. y Womser-Hacker, C.: 2008, Geoclef 2008: The clef 2008 cross-language geographic information retrieval track overview, *In: Proceedings of the 9th CLEF conference on Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 808–821.

Mitra, M., Buckley, C., Singhal, A. y Cardie, C.: 1997, An analysis of statistical and syntactic phrases, *In: Procs. of RIAO-97, 5th International Conference*, pp. 200–214.

Papadimitriou, C. H., Raghavan, P., Tamaki, H. y S.Vempala: 1998, Latent semantic indexing: A probabilistic analysis, *In: Procs. of the 17th ACM Symposium on the Principles of Database Systems*, pp. 159–169.

Peshkin, L. y Savova, V.: 2003, Why build another part-of-speech tagger? a minimalist approach, *In: Recent Advances in Natural Language Processing RANLP*.

Plate, T. A.: 1998, Technical report cs-tr-98-4, *Technical report*, Victoria University of Wellington, Computer Science.

Plate, T. A.: 2002, *Distributed Representation*, Encyclopedia of Cognitive Science, Macmillan Reference Ltd.

Plate, T. A.: 2003, *Holographic Reduced Representation, Distributed Representation for Cognitive Structures*, CSLI Publications.

Pollack, J. B.: 1990, Recursive distributed representations, *Artificial Intelligence* **46**(1-2), 77–105.

Porter, M. F.: 1980, An algorithm for suffix stripping, *Program* **14**(3), 130–137.

Ravana, S. D. y Moffat, A.: 2008, Exploring evaluation metrics: GMAP versus MAP, *In: Procs. of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 687–688.

Rijsbergen, C. J. V.: 1979, *Information Retrieval*, 2nd edition, Department of Computer Science, University of Glasgow.

Sahlgren, M.: 2001, Vector-based semantic analysis: Representing word meanings based on random labels, *In: Procs. of the ESSLLI Workshop on Semantic Knowledge Acquisition and Categorization*, pp. 13–17.

Sahlgren, M.: 2005, An introduction to random indexing, *In: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*.

Sahlgren, M.: 2006, *The Word-Space Model: Using distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High Dimensional Vector Spaces*, PhD thesis, Stockholm University, Stockholm, Sweden.

Sahlgren, M. y Cöster, R.: 2004, 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.

Sahlgren, M. y Karlgren, J.: 2005, Automatic bilingual lexicon acquisition using random indexing of parallel corpora, *Journal of Natural Language Engineering Special Issue on Parallel Texts* pp. 327–341.

Salton, G.: 1986, On the use of term associations, in automatic information retrieval, *In: Procs. of l lth COLING*, pp. 380–386.

Salton, G., Wong, A. y Yang, C. S.: 1975, A vector space model for automatic indexing, *Communications of the ACM* **11**(18), 613–620.

Smeaton, A. F.: 1995, Low level language processing for large scale information retrieval: What techniques actually work, *In: Procs. of a workshop Terminology, Information Retrieval and Linguistics*.

Smeaton, A. F. y Kelledy, F.: 1998, User-chosen phrases in interactive query formulation for information retrieval, *In: of the 20th BCS-IRSG Colloquium, Springer-Verlag Electronic Workshops in Computing*.

Smolensky, P.: 1990, Tensor product variable binding and the representation of symbolic structures in connectionist systems, *Artificial Intelligence* **46**, 159–216.

Sparck, K.: 1999, What is the role of nlp in text retrieval?, *Natural language information retrieval*, T. Strzalkowski, Kluwer Academic Publishers, pp. 1–24.

Stein, B.: 2007, Principles of hash-based text retrieval, *In: Procs. of the 30th annual ternational ACM SIGIR conference on Research and development in information retrieval*, pp. 527–534.

Strzalkowski, T. y Vauthey, B.: 1992, Information retrieval using robust natural language processing, *In: Proc. of the 30th ACL Meeting*, pp. 104–111.

Turpin, A. y Moffat, A.: 1999, Statistical phrases for vector-space information retrieval, *In: Procs. of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 309–310.

Vilares, J., Alonso, M. A. y Vilares, M.: 2004, Morphological and syntactic processing for text retrieval, *In: Fernando Galindo, Makoto Takizawa and Roland Traunmller (Eds.), Database and Expert Systems Applications*, Vol. 3180, LNCS, Springer, pp. 371–380.

Vilares, J., Gómez-Rodrguez, C. y Alonso, M.: 2005, Managing syntactic variation in text retrieval, *In: Procs. of the ACM Symposium on Document Engineering*, ACM Press, pp. 162–164.

Villatoro-Tello, E., Montes-y-Gómez, M. y Villaseñor-Pineda, L.: 2008, Inaoe at geoclef 2008: A ranking approach based on sample documents, *In: Working notes for the CLEF 2008 Workshop*.

Voorhees, E. M.: 2005, The trec robust retrieval track, *In: ACM SIGIR Forum*, pp. 11–20.

Wang, R. y Neumann, G.: 2008, Ontology-based query construction for geoclef, *In: Proceedings of the 9th CLEF conference on Evaluating systems for multilingual and multimodal information access*, pp. 880–884.

Wong, S. K. M., Ziarko, W. y Wong, P. C.: 1985, Generalized vector space model in information retrieval, *In: Procs. of the 8th Annual International ACM-SIGIR Conference*, pp. 18–25.

Zobel, J. y Moffat, A.: 1998, Exploring the similarity space, *SIGIR Forum*, Vol. 32, pp. 18–34.

A

Experimentos previos

En este apéndice se muestran algunos de los experimentos iniciales, que condujeron a la definición de la representación propuesta.

Uno de los primeros experimentos realizados, tuvo como objetivo utilizar la plataforma del Computational Neuroscience Research Group (CNRG) para recuperación de información con la colección CACM. Los parámetros empleados fueron los identificados como óptimos por el CNRG para la tarea de clasificación de textos:

- Etiquetas únicas para cada palabra del vocabulario definidas con Spatter Codes (Kanerva 1996)
- DOR para construir las representaciones de BoC.
- HRR de dimensión 1024 generados con los vectores de contexto
- Etiquetado con partes de la oración de todas las palabras
- Las palabras cerradas se mantuvieron
- Medida de similitud: Distancia Euclidiana (DE).

La Figura A.1 muestra los resultados obtenidos empleando los parámetros anteriores y como medida de similitud entre consultas y documentos distancia euclidiana y coseno.

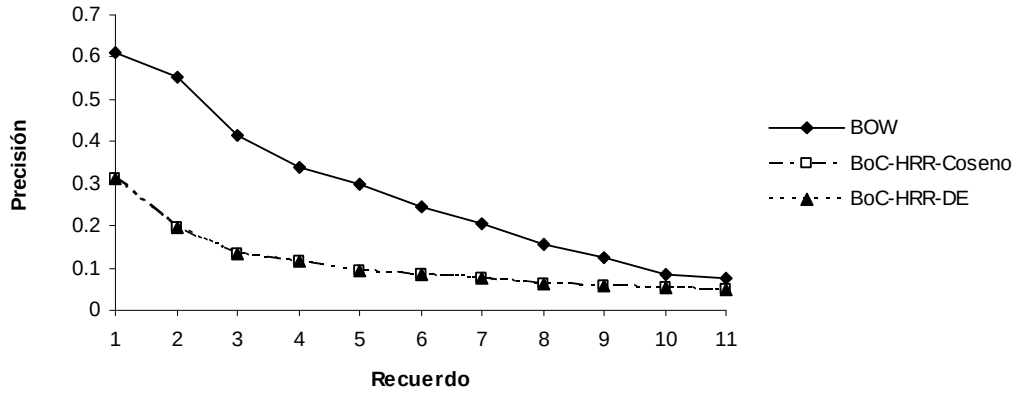

Figura A.1: Resultados empleando los parámetros utilizados para clasificación.

Como puede observarse en la gráfica los resultados están por debajo de los obtenidos con BoW. Los resultados con coseno y distancia euclidiana fueron equivalentes.

Otro experimento que se realizó con la colección CACM fue la generación de HRR con vectores de contexto (CV) y vectores índices (IV) construyendo las representaciones BoC-HRR en un sólo espacio vectorial. Como puede observarse en la tabla A.1 se obtuvieron mejores resultados utilizando los vectores índices. Las últimas dos columnas de la tabla, muestran los resultados obtenidos al construir la representación BoC+HRR con vectores índice en espacios vectoriales separados con resultados mejores, incluso que el BoW.

También se realizaron experimentos orientados a utilizar BoC+HRR de manera independiente de BoW. La Tabla A.2 muestra los resultados obtenidos. Puede observarse que el MAP para todas las colecciones, con excepción de CACM, decrece con respecto a BoW. Estos resultados nos guiaron a la decisión de utilizar BoC y HRR como representaciones complementarias a BoW.

Tabla A.1: Resultados del VSM y BoC-HRR con opciones diferentes

		Un sólo espacio vectorial				Dos espacios vectoriales	
Recuerdo	Precisión VSM	Precisión BoC -HRR-CV	% Cambio	Precisión BoC -HRR-IV	% Cambio	Precisión BoC+HRR	% Cambio
0	0.6129	0.3121	-49.08	0.5432	-11.37	0.6989	14.03
0.1	0.5522	0.2037	-63.11	0.4324	-21.70	0.5978	8.26
0.2	0.4149	0.1273	-69.32	0.3560	-14.20	0.4694	13.14
0.3	0.3407	0.1007	-70.44	0.2581	-24.24	0.3862	13.35
0.4	0.2968	0.0844	-71.56	0.1955	-34.13	0.3092	4.18
0.5	0.2432	0.0809	-66.74	0.1482	-39.06	0.2739	12.62
0.6	0.2058	0.0720	-65.01	0.1148	-44.22	0.2060	0.10
0.7	0.1580	0.0634	-59.87	0.0896	-43.29	0.1527	-3.35
0.8	0.125	0.0593	-52.56	0.0783	-37.36	0.1264	1.12
0.9	0.0862	0.0510	-40.84	0.0695	-19.37	0.0876	1.62
1.0	0.0746	0.0490	-34.32	0.0676	-9.38	0.0762	2.14
Promedio	0.2828	0.1094	-61.30	0.2139	-24.34	0.3077	8.81

Tabla A.2: MAP de BoW y BoC+HRR

Coleccin	BoW	BOC+PHR	% Cambio
ADHOC	0.1969	0.1234	-37.33
ROBUST	0.2247	0.1280	-43.04
GeoCLEF 2005	0.1710	0.1018	-40.47
GeoCLEF 2008	0.1473	0.0420	-71.49
NPL	0.2037	0.1708	-16.15
CACM	0.2727	0.2849	4.47