Una Representación Vectorial para Contenido de Textos en Tratamiento de Información

Maya Carrillo Ruiz, Aurelio López López

Reporte Técnico No. CCC-08-004
25 de Abril de 2008

Coordinación de Ciencias Computacionales
INAOE

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Una Representación Vectorial para Contenido de Textos en Tratamiento de Información

Maya Carrillo Ruiz [1], Aurelio López López[2]
Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro 1, Sta.Ma. Tonantzintla
72840, Puebla, México
[1,2]{cmaya,allopez}@inaoep,.mx

Resumen. Este reporte presenta una representación para documentos que permite codificar relaciones textuales. Dicha representación no considera cada término como una entrada en un vector sino como un patrón, es decir un conjunto de entradas contiguas. Para tratar las variaciones inherentes al lenguaje natural, planeamos expresar las relaciones textuales (p. ej. frases nominales, entidades nombradas, sujeto-verbo, verbo-objeto, adjetivo-sustantivo y adverbio-verbo) como patrones compuestos. Se emplea un operador para unir términos, las relaciones, por lo tanto son codificadas como nuevos "términos", contando de esta manera con elementos descriptivos adicionales para indexar una colección. Los resultados de nuestros primeros experimentos, empleando la representación propuesta en recuperación de información, incorporando frases nominales de dos palabras, mostraron que la representación es factible, recupera y mejora la posición de los documentos relevantes y en consecuencia el valor medio de la precisión promedio.

Palabras clave: Recuperación de Información, Modelo de Recuperación, Modelo Vectorial, Relaciones Textuales, Frases Nominales.

Abstract. This paper presents a document representation to encode textual relations. This representation does not consider each term as one entry in a vector but rather as a pattern, i.e. a set of contiguous entries. To deal with variations inherent in natural language, we plan to express textual relations (such as noun phrases, named entities, subject-verb, verb-object, adjective-noun, and adverb-verb) as composed patterns. An operator is applied to form bindings between terms encoding relations as new "terms", thereby providing additional descriptive elements for indexing a document collection. The results of our first experiments, using the document representation to conduct information retrieval and incorporating two-word noun phrases, showed that the representation is feasible, retrieves, and improves the ranking of relevant documents, and consequently the values of mean average precision.

Key words: Information Retrieval, Retrieval Model, Vector Model, Text Relations, Noun Phrases.

1. Motivación

El aumento de información en formato digital, facilitado por las tecnologías de almacenamiento, impone nuevos retos a las tareas de tratamiento de información, entre las cuales pueden mencionarse: recuperación de información (RI), búsqueda de respuesta (QA), detección y seguimiento de tópico (TDT), agrupamiento y clasificación, entre otras.

La RI puede caracterizarse como el problema de seleccionar de una colección un subconjunto de documentos cuyo contenido es relevante para las necesidades de información, expresadas por un usuario en una consulta. Las técnicas clásicas 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) en común aumenta, entonces la relación es mayor. Bajo este acercamiento, la RI trata de determinar cuánto se parece la bolsa de palabras de la consulta a la bolsa de palabras de cada documento.

Los sistemas de RI basados en bolsa de palabras, que retornan una lista amplia de documentos, con todos o algunos de los términos expresados en la consulta, son incapaces de retornar una respuesta concisa a una solicitud de información específica, para esto se tienen los sistemas de búsqueda de respuesta. Estos últimos podrían aprovechar recursos construidos con documentos preprocesados en cuyas representaciones se almacene información importante ya caracterizada, más allá de la proporcionada por la bolsa de palabras. Un componente importante de estos sistemas es el módulo de selección de respuesta, emplear técnicas basadas en bolsa de palabras en esta actividad, descuida la identificación de relaciones cruciales entre términos, convirtiéndose en una fuente importante de falsos positivos, ya que muchos pasajes irrelevantes comparten los mismos términos de la pregunta, pero las relaciones entre sus términos son diferentes a las mantenidas por los términos de la pregunta [24,25].

La detección y seguimiento de tópico se interesa en la identificación de eventos y en dar continuidad a la información de los mismos. La entrada a este proceso es una secuencia de historias sobre las cuales pueden realizarse tres tareas: el seguimiento de eventos conocidos, la detección de eventos desconocidos y la segmentación de una fuente de noticias en historias. Las aproximaciones que abordan estas tareas comparan las palabras comunes entre historias, entre más palabras tengan en común dos historias, mayor es la probabilidad de que pertenezcan al mismo evento. Este método (bolsa de palabras) es la base para todas las aproximaciones, desde las vectoriales hasta las basadas en modelos de lenguaje [29].

Convencionalmente, las investigaciones en clasificación de documentos se centran en mejorar la capacidad de aprendizaje de los clasificadores. No obstante, la eficacia de la clasificación está limitada por la idoneidad de la representación de los documentos, que generalmente se realiza empleando bolsa de palabras.

En el agrupamiento de documentos, que tiene como finalidad analizar colecciones, dividiéndolas en grupos de documentos similares, tradicionalmente los documentos son representados como bolsas de palabras, sin aprovechar las relaciones existentes entre las mismas.

Así pues, la aproximación de bolsa de palabras se utiliza ampliamente dado que genera de manera rápida resultados aceptables; sin embargo, no considera variaciones lingüísticas como la morfológica, que origina palabras con diferente número, género, tiempo, modo; la léxica, en la que diferentes palabras tienen el mismo significado; la sintáctica, en la que el orden de las palabras cambia el significado y la semántica, en la que una palabra puede tener diferentes significados [18].

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 lingüísticas. 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. Se ha postulado que una representación más adecuada del texto debería incluir grupos de palabras, ya sea frases o expresiones que denoten entidades con significado, conceptos o relaciones dentro del dominio de búsqueda.

Investigadores trabajando en el campo han utilizados técnicas de procesamiento de lenguaje natural (NLP) para tareas de tratamiento de texto, suponiendo que un mejor entendimiento de las solicitudes de información y de los documentos es clave para mejorar la efectividad de dichas tareas. El NLP busca utilizar información semántica, además de información estadística, para mejorar el análisis de los textos y extraer información que será almacenada de manera permanente en un índice (indexar los documentos), con la información extraída, se producen los elementos procesados y la estructura sobre la cual se realizarán las tareas de tratamiento de información. Crear índices más adecuados para los documentos, mejora la precisión al compararlos. La

información semántica se obtiene procesando el lenguaje no tratando cada palabra de manera independiente. El resultado más simple de este procesamiento genera frases que pueden ser empleadas como elementos para indexar los documentos. Algunos métodos de extracción de frases emplean el análisis sintáctico e intentan capturar uniformidades semánticas a partir de la estructura superficial, acercándose un poco más al contenido. Las frases sintácticas parecen ser indicadores razonables de contenido, ya que permiten identificar cambios en el orden de las palabras y algunas otras variaciones en la estructura. Sin embargo, este análisis sintáctico está lejos de un análisis semántico real. Una ventaja del NPL es que puede producir frases de varios términos para representar un concepto más específico, p. ej. *Sistema de información geográfica*. El primer paso para determinar las frases es un análisis léxico. Posteriormente, para obtener ventajas mayores, pueden identificarse dependencias sintácticas y semánticas creando una jerarquía de conceptos semánticos. En el caso ideal, todas las variaciones de una frase deben reducirse a una forma canónica. Una aproximación para encontrar una forma común es transformar la frase a una forma operador-argumento o encabezado-modificador.

Análisis más complejos, pueden combinar conceptos (frases) para formar niveles superiores de conceptos que se conocen como representaciones temáticas. En cuanto a la identificación de conceptos, pueden emplearse vocabularios controlados, definidos por una organización para representar los conceptos que ellos consideran representaciones importantes de sus datos y hacer corresponder los términos identificados a cada uno de los conceptos especificados. De esta manera, una colección puede ser indexada en función de conceptos, reduciendo el número de términos empleados para crear el índice. En lugar de definir de manera anticipada los conceptos, estos pueden definirse de manera automática iniciando con clases no etiquetadas de conceptos y dejando que la información de los documentos vaya creando las clases de conceptos empleando p.ej. algoritmos basados en redes neuronales. El proceso de hacer corresponder un término a un concepto que lo represente es complejo, porque un término puede representar diferentes conceptos con diferentes grados de certeza (polisemia), entonces un término en un documento necesita ser representado por varios códigos de conceptos con diferentes pesos. Por ejemplo *automóvil* puede representar los conceptos *vehiculo, transportación, ambiente, combustible, dispositivo mecánico,* con pesos 0.65, 0.60, 0.35, 0.33, 0.15, respectivamente, entonces el término *automóvil* se representará como un vector de conceptos con estos pesos. Las frases o documentos se representarán entonces como el promedio de los pesos, de los vectores de conceptos correspondientes a los términos presentes en ellos [15].

En el presente trabajo, se buscará identificar relaciones entre los términos de un texto a fin de establecer una representación de contenido de texto, que permita describir documentos con más detalle que el permitido por la aproximación de bolsa de palabras. La representación se probará en RI, pero tendrá el potencial de ser explotada por otras tareas de tratamiento de información.

2. Problema

La efectividad de una representación de documentos, está directamente relacionada a la exactitud con la cual el conjunto de términos seleccionados representa el contenido de un documento y a cuán bien se puede contrastar el contenido de dicho documento con otro. Es decir, dados dos documentos D_1 y D_2 y sus representaciones R_1 y R_2, respectivamente, si R_1 es igual a R_2 esto significa que el contenido de R_1 es igual al contenido de R_2 a cierto nivel de abstracción.

En cuanto a la representación de contenido, la bolsa de palabras es insuficiente, ya que palabras aisladas son poco específicas para efectuar una discriminación adecuada entre documentos [23, 26]. Un método más apropiado, como se ha mencionado, sería identificar grupos de palabras con significado que denoten conceptos o relaciones, este será el trabajo de la presente investigación, la utilidad de la representación resultante será comprobada en RI, por tal motivo, a continuación se presenta una descripción de dicha tarea.

La recuperación de información (RI) se interesa en seleccionar, de grandes volúmenes de texto, aquellos documentos que contienen información relevante de acuerdo a las necesidades de información expresadas por un usuario.

La RI incluye dos actividades principales: indexar y buscar. La primera se refiere a representar el contenido de los documentos y la solicitud de información que emite 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 los que resulten de mayor relevancia.

La riqueza inherente al lenguaje, y la diversidad en la manera de expresar las necesidades de información por parte de los usuarios, 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.

Un modelo de RI se define con la representación que se da a los documentos y las consultas (objetivo de esta investigación), y la función de relevancia que se emplea para compararlos.

Ahora bien, dada una solicitud de información, por ejemplo:

"Estoy interesado en mecanismos de comunicación entre *procesos* disjuntos, posiblemente, pero no exclusivamente en ambientes distribuidos. Preferiría ver *descripciones* de *mecanismos* completos con o sin *implementación* en lugar de trabajos teóricos del *problema*. Llamada a procedimientos remotos y paso de mensajes, son ejemplos de mi interés" [10].

Podemos observar que existen palabras como *descripciones*, *mecanismos*, *implementación* y *problema,* que deberían estar presentes en los documentos recuperados pero manteniendo la relación expresada, ya que se buscan "descripciones de mecanismos de comunicación entre procesos". Si empleamos de manera aislada las palabras mencionadas, algunas serán ignoradas por su generalidad al crear el índice y si se incluyen en el índice conducirán a recuperar gran cantidad de documentos irrelevantes.

Cuando se recupera un grupo de documentos se encontrará que algunos de ellos no son relevantes 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 recuperados y el recuerdo, la razón del número de documentos relevantes recuperados entre el total de documentos relevantes existentes en la colección.

En el área de recuperación de información se continúa experimentando con nuevos modelos y por ende nuevas representaciones de documentos, para intentar cubrir con mayor precisión las necesidades de información de los usuarios.

Como se ha especificado, nos centraremos en RI, donde si se desea obtener mejoras substanciales debe empezarse a pensar más allá de la aproximación de bolsa de palabras, con precisiones típicamente por debajo del 50% hay mucho margen para mejorar [27]. En la siguiente sección se presentan trabajos que describen modelos clásicos y nuevos de recuperación de información para concluir con métodos que explotan la obtención de relaciones, empleando técnicas de procesamiento de lenguaje natural, para realizar recuperación de información.

3. Trabajo Relacionado

Los modelos clásicos de RI consideran que los documentos están descritos por un conjunto de términos que son empleados para indexar y resumir su contenido. 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 asignando diferentes pesos a dichos términos. Existen tres modelos de RI clásicos: el booleano, el vectorial y el probabilístico [16,19, 28].

El modelo booleano ve el problema de RI desde la perspectiva de la teoría de conjuntos y algebra booleana. Dada su simplicidad este modelo fue adoptado por los primeros sistemas bibliográficos comerciales, sin embargo presenta inconvenientes: primero su estrategia de

recuperación está basada en un criterio de decisión binario, el documento es relevante o no lo es; segundo, dado que las expresiones booleanas tienen una semántica precisa y limitada, con frecuencia es difícil para el usuario traducir sus necesidades de información a expresiones booleanas. Este modelo considera un término como presente o ausente en un documento y como resultado de esta consideración el peso de un término será 1 "o" 0. Una consulta está compuesta por términos unidos por los conectores: not, and, or. La similitud de una consulta con un documento será 1 si dada la forma disyuntiva de la consulta, alguna de sus partes está presente en el documento y 0 en caso contrario. Un modelo alternativo al booleano es el modelo de conjuntos difusos, que considera a la consulta y a los documentos como conjuntos difusos. El conjunto difuso, que incluye los conceptos $C = \{c_1, c_2, ..., c_n\}$ se representa como: $A = \{(c_1, f_{A(c_1)}), (c_2, f_{A(c_2)}), ..., (c_n, f_{A(c_n)})\}$ donde $f_A : C \rightarrow [0,1]$ es la función de membresía que indica el grado de pertenencia de un elemento al conjunto. Si se considera el conjunto D de todos los documentos de la colección. El conjunto difuso D_t será el conjunto de todos los documentos que contienen el término t: $D_t = \{(d_1, f_{t1}), (d_2, f_{t2}), ... (d_n, f_{tn})\}$ que indica que el documento d_1 contiene el termino t con confianza f_{t1}. De igual manera el conjunto D_s de todos los documentos que contienen el término s puede definirse como $D_s = \{(d_1, f_{s1}), (d_2, f_{s2}), ... (d_n, f_{sn})\}$. Calcular $s \vee t$ requiere de $D_s \cup D_t$ y $s \wedge t$ de $D_s \cap D_t$ que se calculan tomando el máximo valor para la unión y el mínimo para la intersección para cada documento. Con estas dos operaciones y el complemento $1 - f_{A(ci)}$ pueden construirse expresiones más complejas, al final se tendrá un conjunto con los documentos y sus coeficientes de similitud. Esta aproximación puede utilizarse en redes conceptuales para determinar con cuanta certeza los conceptos de una consulta se presentan en cada documento [17].

El modelo vectorial, que ve el problema de RI desde la perspectiva del algebra lineal, considera similitud parcial entre consultas y documentos asignando pesos (no binarios) a los términos. Estos pesos se emplean para calcular la similitud. Este modelo ordena los documentos recuperados de manera decreciente, de acuerdo al grado de similitud. Los documentos y consultas son representados como vectores de dimensión t (número total de términos en la colección) y la similitud se calcula como el coseno del ángulo entre el vector que representa a la consulta y el que representa al documento. Dentro de esta clasificación, como modelo alternativo, está el modelo de indexación semántica latente (Latent Semantic Indexing, LSI) y el de redes neuronales.

El modelo de LSI, considera la asociación de documentos a consultas en función de conceptos. La idea principal es hacer corresponder los vectores de documentos y consultas a un espacio de menor dimensión asociado con conceptos [16]. El proceso es relativamente sencillo. Una matriz A de términos se construye de tal manera que la posición (i, j) indique el número de veces que el término i aparece en el documento j, es decir A es una matriz cuyos renglones representan términos y las columnas documentos. Una descomposición en valores singulares (SVD) de esta matriz resulta en matrices USV^T tal que S es una matriz diagonal. Los valores en S son los valores singulares que se ordenan por magnitud y se eligen los k valores superiores. Los valores singulares restantes son igualados a 0. Sólo las primeras k columnas se mantienen en U_K; sólo los primeros k renglones se conservan en V_K^T. Una matriz nueva A' se genera para aproximar $A=USV^T$. La comparación entre dos términos se realiza con el producto punto de los renglones correspondientes en U_K y la comparación de dos documentos con el producto punto de los correspondientes renglones en V_K^T. Para calcular la similitud entre los documentos y una consulta, ésta última se trata como un documento adicional y se calcula SVD [17].

El modelo de redes neuronales representa las conexiones entre las neuronas cerebrales por un grafo simplificado. Los nodos en el grafo son las unidades de proceso y las aristas son consideradas las conexiones sinápticas del cerebro. Para simular que la fuerza de las conexiones sinápticas cambia a los largo del tiempo se asignan pesos a las aristas. En cada instante el estado de los nodos se define por su nivel de activación (que está en función de su estado inicial y las señales que recibe de entrada). De acuerdo al nivel de activación, un nodo N puede enviar una señal al nodo M y la fuerza de dicha señal está determinada por el peso de la arista que une a los nodos. Para RI

una red neuronal está compuesta por tres capas de nodos: una para los términos de las consultas, una para los términos de los documentos y una para los documentos. Los nodos de los términos de la consulta son los que inician el proceso de inferencia enviando señales a los nodos de los términos de los documentos, estos a su vez generan señales para los nodos de los documentos, lo que comprende la primera etapa. Posteriormente los nodos de los documentos pueden generar nuevas señales que regresen a los nodos de los términos de los documentos que a su vez al recibir este estímulo, disparar nuevas señales a los nodos de documentos repitiéndose el proceso. Las señales se debilitan en cada iteración y la activación de nodos eventualmente se detiene. A los nodos de los términos de la consulta se les puede asigna inicialmente un valor que corresponda al peso normalizado asociado a este término en el modelo vectorial, una vez que los nodos de los términos de los documentos reciben estas señales envían una señal igual al peso normalizado de dicho término en el documento de acuerdo al modelo vectorial. Los valores recibidos en los nodos de los documentos serán igual a la suma de los pesos de los términos en consultas y documentos. Para mejorar el proceso de recuperación, la red continúa con la propagación del proceso de activación lo que modifica el ordenamiento vectorial inicial, en un proceso análogo al de retroalimentación de relevancia por el usuario. Para hacer el proceso de activación más efectivo puede definirse una vecindad mínima, de tal manera que los documentos con valores inferiores a esta vecindad dejan de enviar señales.

Por otra parte, el modelo probabilístico, dada una consulta q y un documento d en una colección, trata de estimar la probabilidad de que el usuario encuentre el documento interesante (relevante). El modelo supone que la probabilidad de relevancia depende solo de la representación de la consulta y el documento. Además supone que existe un subconjunto de todos los documentos preferidos por el usuario como conjunto respuesta a la consulta q. Tal conjunto respuesta ideal (R) debe maximizar la probabilidad de relevancia para el usuario. Los documentos en R son considerados relevantes e irrelevantes si no están en R. El principal inconveniente de este modelo es que no se establece explícitamente cómo calcular las probabilidades de relevancia, aun más no se establece el espacio de muestreo que debe emplearse para definir tales probabilidades. La similitud entre un documento y una consulta se calcula como la razón entre dos probabilidades: P (d sea relevante para q) / P (d no sea relevante para q). Las probabilidades de relevancia obtenidas, permiten ordenar los documentos de acuerdo a la importancia que tienen para la consulta [16]. Asociados al modelo probabilístico están los modelos de lenguaje cuya idea central es que los documentos pueden ordenarse de acuerdo a la verosimilitud que tienen de generar la consulta. Formalmente, la similitud se calcula como $Sim\ (q,d_i) = P(q|M_{Di})$ donde M_{Di} es el modelo de lenguaje implícito en d_i . El significado de "generar la consulta" es contar con un modelo probabilístico para las consultas, para esto puede modelarse la presencia o ausencia de términos como eventos independientes de Bernoulli y ver la generación completa de la consulta como la unión de los eventos de observar todos los términos de la consulta y no observar ningún término no presente en la consulta. En este caso la similitud será calculada como:

$$Sim(q,d_i) = \prod_{t_j \in q} P(t_j \mid M_{D_i}) \prod_{t_j \notin q} 1 - P(t_j \mid M_{D_i})$$

es decir el producto de las probabilidades de los términos presentes en la consulta y los términos ausentes en ella. La manera más simple de calcular $P(t_j| M_{Di})$, la probabilidad del término j dado M_{Di}, es calcularla como la frecuencia relativa del término, es decir, la razón de la frecuencia del término t_j en el documento d_i dividida por la longitud del documento d_i. Esta medida tiene problemas cuando un término de la consulta no aparece en el documento, pues el valor de la probabilidad se hace 0. Para evitar dicho problema existen varios enfoques de suavizado [17].

Otro modelo dentro de los probabilísticos es el de redes de inferencia, que asocia variables aleatorias a los documentos, consultas y términos de la colección (un vector $\vec{k} = (k_1, k_2, ..., k_t)$ de variables aleatorias para los términos). La variable aleatoria asociada al documento d_j representa el evento de observar este documento. Las variables de documentos y términos se representan como

nodos en una red. Se utilizan arcos dirigidos del nodo de un documento a los nodos de sus términos, indicando que la observación de un documento incrementa la creencia en los nodos de sus términos. La variable aleatoria asociada con la consulta modela el evento de que la solicitud de información especificada por la consulta se cumpla. Esta variable aleatoria también se representa como un nodo de la red. La creencia en el nodo de esta consulta está en función de la creencia en los nodos asociados a los términos de la consulta. Así se tienen arcos dirigidos de los nodos de los términos al nodo de la consulta. La red puede incluir nodos de conceptos para documentos y consultas. Dado un conjunto de documentos ordenados de acuerdo a la importancia que tienen para la consulta *q*, la posición del documento d_j (ranking), es una medida de cuánto la observación del documento d_j soporta la evidencia de la consulta y se calcula como: $P(q \wedge d_j) = \sum_{\forall \vec{k}} P(q \mid \vec{k}) \times P(\vec{k} \mid d_j) \times P(d_j)$ [16].

Los trabajos que se presentan a continuación muestran el interés de los investigadores por definir nuevos modelos de recuperación de información y por ende, de contar con nuevas representaciones de documentos que repercutan en dicha tarea.

Shi et al. en [1] proponen Gravitacional-Based Model (GBM) un modelo de RI inspirado en la Teoría de la Gravedad de Newton. En este modelo se define un término como un objeto físico compuesto de partículas con forma específica (esfera o cilindro ideal) que tiene tres atributos: tipo, masa y diámetro. Una partícula tiene dos atributos tipo y masa. Dos partículas del mismo tipo se atraen mutuamente. Un documento es una lista de términos. La masa del documento será la suma de las masas de sus términos explícitos (presentes en el documento) e implícitos (no presentes). El diámetro será igualmente la suma de los diámetros. La consulta se modela como un objeto compuesto de términos únicamente explícitos. La relevancia de un documento, dada una consulta, se calcula como la fuerza de atracción entre los objetos presentes en él y los presentes en la consulta empleando la fórmula de gravedad de Newton. Se presenta una derivación de la fórmula BM25 (GBM-Std) y se plantea la posibilidad de utilizar esta aproximación para derivar nuevas funciones de relevancia. Los autores presentan experimentos realizados sobre extractos del TREC 2000-2004 y comparan el rendimiento de su formula GBM-Std con la de ponderación normalizada con pivote (pivoted normalization weighting), y la fórmula Dirichlet empleada en modelos de lenguaje. La métrica que emplean para dicha comparación, es la media de la precisión promedio (MAP). Los resultados que reportan para las diferentes fórmulas, son equiparables con un desempeño ligeramente mejor para GBM-Std.

Gonçalves, et al. en [2] presentan Latent Relation Discovery (LRD), un método que agrega información al modelo vectorial con base en el establecimiento de relaciones entre entidades nombradas (personas, organizaciones, localidades) que aparecen conectadas unas a las otras (coocurrencia), obtenidas de los textos. Los autores identifican las entidades nombradas y determinan la fuerza de la relación de coocurrencia entre ellas, en función de la distancia que las separa y frecuencia de cada coocurrencia. Dado un documento D, en el que aparecen las entidades e_1, e_3, e_4, e_5, si por el análisis del corpus se sabe que e_1 tiene una relación de coocurrencia fuerte con la entidad e_2, entonces al formar el vector de D, se agrega la entidad e_2. La similitud entre el documento y la consulta la determinan mediante el coseno. El método es comparado con el de información mutua, información mutua de Vechtomova et al., Phi cuadrada, score Z y LSI. Los experimentos se realizaron con CISI (Glasgow Information Retrieval benchmark dataset) que tiene 1460 documentos y 112 queries. Los autores eligen 20 consultas de manera aleatoria de las 112, establecen una vecindad (0.54) para el valor del coseno obtenido al comparar las consultas con los documentos y sólo consideraron los valores dentro de esta vecindad para calcular la medida $F = 2 \times precisión \times recuerdo / precisión + recuerdo$, que es la que emplean para compararse con los métodos mencionados anteriormente, utilizan diferentes ventanas de texto y agregan un número diferente de entidades. En todos los casos los resultados obtenidos con el modelo planteado, tomando el promedio de la medida F para las 20 consultas, son mejores a los obtenidos con los métodos seleccionados para compararse, obteniendo como mayor promedio de F un 19.3 % sin ninguna ventana y agregando 30 entidades a los documentos, seguido por LSI con 16.6 %.

Tabla 3.1 Modelos de RI

Modelo	**Representación**	**Función de relevancia**	**Representación de conceptos y/o relaciones**
Booleano	Conjuntos	Función booleana	NA
Conjuntos difusos	Conjuntos difusos	Función con valores asociados	Redes conceptuales
Vectorial	Vectores	Coseno	NA
LSI	Matriz	Producto punto	Conceptos y relaciones implícitos en una descomposición matricial
Redes Neuronales	Nodos	Nivel de activación	Conceptos expresado en nodos
Probabilístico	Conjuntos de probabilidades	Suma de probabilidades	NA
Modelo de lenguaje	Conjuntos de probabilidades	Producto de probabilidades	NA
Redes de inferencia	Nodos	Función de probabilidad	Conceptos
GBM	Objetos	Fórmula de gravedad	NA
LRD	Vectores extendidos	Coseno	Sólo relaciones de concurrencia
TVMS	Vectores	Producto punto	Conceptos representados como vectores y ángulos

Becker, et al. en [3] presentan Topic-based Vector Space Model (TVSM), una nueva aproximación vectorial para comparar documentos. Los autores consideran un espacio vectorial positivo R de dimensión *d,* donde cada dimensión representa un tópico ortogonal con respecto a los otros (p. ej. literatura, computación). Los términos son vectores con pesos entre uno y cero. El peso de un término se define como la longitud del vector que lo representa. El vector de un término (p.ej. software, program) relacionado con un tópico (p. ej. computación) apunta a la misma dirección que el vector de dicho tópico. Los vectores de los términos con ninguna relación a los tópicos definidos (p. ej. is, the) tienen un ángulo de 45° con respecto a los vectores de los tópicos representados en el espacio. Un documento se representa como la suma de sus términos. La similitud entre documentos se calcula con el producto punto, un valor cercano a uno indica similitud; mientras uno cerca a cero indica que no hay relación. Los autores argumentan que su modelo permite la incorporación de diferentes recursos de procesamiento de lenguaje natural, lo que facilita explorar la dependencia entre algoritmos y mejorar los modelos de NLP, sin embargo su trabajo es meramente teórico. La tabla 3.1 muestra la representación dada a los documentos, la función de relevancia empleada y si pueden representarse conceptos y/o relaciones los modelos presentados.

Por otra parte, existen trabajos previos que sugieren el uso de más que términos simples para indexar y recuperar documentos. Por ejemplo, Lewis et al. en [4] se cuestionan con respecto a: ¿Cómo deberían ser las unidades lingüísticas empleadas para indexar los documentos? ¿Cuál debería ser el tamaño del texto que se explora para obtenerlas y cómo deben representarse? Ellos proponen utilizar unidades de lenguaje natural bien formadas tomadas del texto y afirman que dada la utilidad comprobada del ponderado estadístico, cualquier unidad que el NLP produzca debe filtrarse y ponderarse empleando estadística para determinar su importancia en la colección bajo estudio y tal vez para otras colecciones. Para tratar con colecciones muy grandes proponen construir resúmenes que representen de manera adecuada los documentos, determinando experimentalmente

la forma de reducción que sea útil y no muy costosa. En cuanto a la representación de los textos dicen que las palabras y los términos compuestos deben permitir representar conceptos con un rango aceptable de complejidad y de tal manera que la pérdida de acoplamiento entre ellos se mantenga eficiente y flexible para favorecer la recuperación. Los autores enfatizan la necesidad de múltiples experimentos para precisar la forma adecuada de los términos compuestos y de cómo deben seleccionarse y ponderarse. En cuanto al papel del NLP, afirman que debe emplearse para justificar la selección de términos con respecto a la estructura gramatical del texto y tal vez para caracterizar la estructura interna de los términos. Ellos mencionan que el NLP podría aplicarse solamente a las consultas con la evidencia de que los términos compuestos resultantes aplican a un documento, determinado esto último, sólo por pruebas de proximidad de palabras. Otra alternativa que plantean es aplicar NLP sólo a los documentos que ocupen los rangos más altos después de efectuar recuperación empleando únicamente términos. Concluyen diciendo que se necesita un diseño cuidadoso del sistema de RI como un todo para optimizar todos los factores involucrados, dada su independencia y que existen dos retos principales para las tecnologías de NLP en RI: primero, lograr que esta tecnología opere eficiente y efectivamente a la escala necesaria, y segundo, conducir la evaluación de pruebas necesarias para descubrir si una aproximación funciona de manera adecuada.

Mitra, et al. en [5] presentan un estudio que compara la utilidad de el reconocimiento de frases nominales empleando métodos estadísticos y lingüísticos. Los autores definen frase estadística como un par de palabras contiguas que aparecen en al menos 25 documentos, 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). El sistema de 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. Dado que la identificación de NP en documentos es cara, no procesan todos los documentos y aproximan estadísticamente la frecuencia de las frases. Los experimentos los realizan con las secciones de Wall Street Journal, AP Newswire, and Ziff-Davis del disco 2 del TREC con un total de 211,395 documentos. Con las consultas 151 a 200, para este conjunto de consultas se tienen 4273 documentos relevantes. Los 100 documentos con rango mayor, obtenidos para cada consulta, empleando términos simples, se procesan para obtener las frases nominales sintácticas e indexan el subconjunto de documentos con dichas frases. Obtienen una mejora de 0.1 % con frases estadísticas y de 1.1% con frases sintácticas. Ellos concluyen que las frases nominales son útiles a niveles de precisión bajos, cuando la conexión entre documentos relevantes es mínima, siempre y cuando se cuente con un esquema de ponderación adecuado.

Evans, et al. en [6] presentan una aproximación para indexar frases nominales para RI, ellos describen un método híbrido para extraer subcomponentes (continuos o discontinuos) de frases nominales complejas. Sus resultados mejoran tanto el recuerdo como la precisión.

Recientemente Vilares, et al. en [7], [8] y [9] presentan trabajos en los que la extracción de relaciones ha mejorado la precisión de la RI. 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, realizan pruebas con el corpus del CLEF 2003, sin éxito. Los autores posteriormente utilizan la colección del CLEF 2001/02 formada por 215,738 documentos (509 MB), empleando solo el titulo y la descripción de los mismos. En los experimentos consideran indexar tanto los documentos como el query por términos simples (palabras) y términos compuestos (pares de dependencias sintácticas). Primero aplican lematización al query, realizan la consulta y extraen los *t* mejores términos de los *n* documentos evaluados como más aproximados al query. Los *t* términos son empleados para expandir el query y realizan una nueva consulta para obtener el conjunto final de documentos recuperados. Esta vez sus resultados muestran mejora. En los documentos se presentan resultados de experimentos posteriores que permiten observar que la mejora, aunque en menor grado, se mantiene, utilizando sólo la relación nombre-modificador. Los mejores resultados los obtienen considerando precisión a 10 documentos, presentando una mejora del 13.18% al comparar el modelo vectorial con truncamiento y el modelo vectorial considerando las dependencias binarias que proponen y de 13.08 % considerando solo la dependencia nombre-modificador.

4. Investigación propuesta

En esta sección se describe la investigación a desarrollar. Se presenta la pregunta de investigación que nos ocupa, los objetivos a alcanzar, la metodología a seguir y las contribuciones esperadas.

4.1. Pregunta de Investigación

De lo expuesto en las secciones anteriores, surge la pregunta de investigación que este trabajo busca contestar:

¿Cuál será el impacto en tareas de tratamiento de información, si se consideran relaciones entre términos, que permitan asociarlos y emplear dichas asociaciones como unidades para evaluar la similitud entre documentos?

4.2 Objetivos

Los objetivos de este trabajo son:

4.2.1 Objetivo general

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

4.2.2 Objetivos particulares

- Determinar la representación adecuada para documentos que permita capturar asociaciones entre términos
- Establecer cómo asociar los términos para que constituyan unidades de comparación
- Plantear un esquema de pesado adecuado para las relaciones identificadas, de manera que contribuyan a mejorar la precisión.
- Especificar la función de similitud adecuada para comparar documentos.
- Evaluar la representación en recuperación de información y explorar las ventajas en otras.

4.3 Metodología

Los pasos a seguir para definir la representación de contenido de textos son:

1. Determinar la representación que se dará a los documentos y realizar experimentos para comprobar que dicha representación puede alcanzarse
2. Establecer las relaciones a considerar para la representación. Podría empezarse por las mencionadas en la bibliografía (frases nominales, entidad nombrada, sujeto-verbo, verbo-objeto, adjetivo-sustantivo, adverbio-verbo).
3. Especificar el mecanismo para asociar términos. a fin de representar las relaciones identificadas y realizar experimentos para comprobar su funcionamiento.
4. Extraer las relaciones determinadas de los textos mediante herramientas como analizadores sintácticos, etiquetadores de partes de la oración (POS) y etiquetadores de entidades nombradas, utilizando las herramientas disponibles que proporcionen la mejor precisión posible.
5. Agregar a la representación un tipo de relación a la vez para definir términos compuestos e indexar las colecciones, observar experimentalmente el comportamiento de la representación y determinar la utilidad de dicha relación para mejorar la tarea de RI en cuanto a precisión, aunque tratando de que el recuerdo también sea aceptable.
6. Validar experimentalmente la representación para contenido de textos resultante, aplicada a RI, comparándola con la representación vectorial clásica inicialmente y con otra, si se identifica alguna durante la revisión del estado del arte, que continuará hasta el primer tercio del último año de trabajo.
7. En función de los resultados obtenidos se podrá explorar el funcionamiento de la representación en otras tareas de tratamiento de información (p.ej. agrupamiento o búsqueda de respuestas).

4.4 Contribuciones

Las contribuciones de la presente investigación serán:

- Obtener un esquema de representación para contenido de textos sencillo y efectivo para expresar conceptos y sus relaciones. Algunas aproximaciones emplean árboles sintácticos sin embargo con esta representación comparar textos no es sencillo. Con una representación vectorial la comparación es más directa. La efectividad del esquema estará determinada por los resultados en cuanto a precisión que se alcancen en la tarea a la que se aplique.
- Un método para llevar las relaciones extraídas a la representación propuesta
- Definir una medida de similitud entre documentos
- Un modelo de recuperación de información a partir de la representación propuesta

4.5 Resultados Preliminares

En esta sección se presentan los resultados alcanzados hasta el momento: una representación de contenido de textos tentativa incorporada en un prototipo de sistema de recuperación de información y la evaluación en tres colecciones, considerando en ellas únicamente frases nominales de dos palabras. Los resultados se enuncian de acuerdo a los objetivos planteados para la investigación.

4.5.1 Determinar la representación adecuada para documentos que permita capturar relaciones entre términos

Se propone representar cada término por un patrón pequeño formado por dígitos binarios y un conjunto de índices, *index*, que indican la posición de los dígitos del patrón dentro de un vector. Si *pt* es el patrón de un término $\vec{pt}$ representa el vector que se obtiene al colocar *pt* en los índices indicados por *index.*

Los documentos se representan como conjuntos de atributos, es decir, vectores resultantes de combinar los vectores de sus términos mediante la operación de suma vectorial. Si *D* es un documento con términos $t_1, t_2, \ldots, t_n$, su representación estará dada por el vector:

$$\vec{D} = \langle \vec{t}_1 + \vec{t}_2 + \ldots + \vec{t}_n \rangle$$

donde la flecha colocada sobre las literales indica que se trata de vectores. Después de sumar los términos, el vector se normaliza, esto se denota por $\langle\ \rangle$. Supongamos que el documento *D* tiene los términos t_1, t_2 y t_3 cuyos vectores son: $[v_1, v_2, 0, 0, 0, 0]$, $[0, 0, v_3, v_4, 0, 0]$ y $[0, 0, 0, 0, v_5, v_6]$ respectivamente, considerando patrones de dos posiciones, entonces el vector de *D* será: $[v_1, v_2, v_3, v_4, v_5, v_6]$ que como último paso deberá normalizarse, es decir, dividir cada componente del vector entre $\sqrt{v_1^2 + v_2^2 + v_3^2 + v_4^2 + v_5^2 + v_6^2}$. Cada v_i puede ponderarse para reflejar la importancia del término, p. ej., empleando el esquema tf.idf [21].

Esta representación permitirá asociar los vectores de los términos para generar vectores de relaciones que se agreguen a la representación de los documentos, esto difícilmente puede realizarse con la representación vectorial clásica en la que cada término está representado por un solo número (comúnmente expresando la frecuencia del término)

4.5.2 Establecer cómo asociar los términos para que constituyan unidades de comparación

Como se ha mencionado, buscamos localizar y representar de manera adecuada las relaciones identificadas entre términos para mejorar la expresividad y en consecuencia la precisión. La representación de dichas relaciones permitirá expresar cierta estructura superficial, haciendo más detalladas las descripciones de los textos.

La codificación de estructura requiere una forma de vincular atributos particulares. Para este propósito, proponemos emplear el operador de convolución circular como operador de vinculación, para codificar asociaciones entre términos y de esta manera representar cierta estructura (propuesta inspirada en trabajos previos de ciencia cognitiva que buscan explicar cómo el cerebro humano procesa analogías [11,14]). La convolución circular hace corresponder dos vectores de valores reales de dimensión *n* dentro de otro. Si *x* e *y* son vectores *n*-dimensionales (referidos con subíndices de *0* a *n-1*), entonces los elementos de $z = x \otimes y$ son:

$$z_i = \sum_{k=0}^{n-1} x_k y_{i-k}$$

donde los subíndices son tomados modulo-*n* y $\otimes$ denota la convolución circular. Este operador de vinculación mantiene el mismo tamaño de los vectores, puede ser decodificado, preserva la estructura y puede aplicarse recursivamente [11,14].

En la codificación de relaciones, además de emplear los patrones de los términos que intervienen en ellas, también se utilizan patrones especiales para identificar el papel (rol) de los términos (p. ej., parte derecha de frase nominal, parte izquierda de frase nominal, sujeto, verbo, objeto directo, objeto indirecto, adjetivo, adverbio, persona, localidad, organización). Los vectores de estos patrones especiales, junto con los vectores 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, sí estos desempeñan un papel diferente, para codificar la relación se necesitarán dos patrones especiales: *izq* y *der* . El vector *R* de la relación será:

$$\vec{R} = \left(izq \otimes \vec{r}_1 + der \otimes \vec{r}_2\right)$$

Dado un documento *D* con términos $t_1, t_2, \ldots, t_{x1}, t_{y1}, \ldots, t_{x2}, t_{y2}, \ldots, t_{xn}, t_{ym}, \ldots, t_n$ y relaciones $R_1, R_2, \ldots, R_n$ entre los términos $t_{x1}, t_{y1}; t_{x2}, t_{y2}; \ldots; t_{xn}, t_{yn}$, respectivamente su vector será construido como:

$$D = \left\langle \begin{array}{l} \vec{t}_1 + \vec{t}_2 + ... + \vec{t}_n + (izq \otimes \vec{t}_{x1} + der \otimes \vec{t}_{y1}) + (izq \otimes \vec{t}_{x2} + der \otimes \vec{t}_{y2}) + ... \\ + (izq \otimes \vec{t}_{xn} + der \otimes \vec{t}_{yn}) \end{array} \right\rangle$$

Si el documento *D* tiene los términos t_1, t_2 y t_3 cuyos vectores son: *[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],* respectivamente, (empleando nuevamente dos posiciones para representar el patrón de cada término), una relación de colocación entre t_2 y t_3 y además suponemos los vectores especiales *izq* y *der* como *[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]* entonces la convolución circular para dos vectores de dimensión diez se define como:

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

y por lo tanto:

$$izq \otimes t_2 = [s_7 v_3, 0,0,0,0,0,0,0, s_6 v_2, s_6 v_3 + s_7 v_2]$$

$$der \otimes t_3 = [0,0, s_8 v_4, s_8 v_5 + s_9 v_4, s_9 v_5, 0,0,0,0,0]$$

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

$t_1+t_2+t_3$ = *[v_0, v_1, v_2, v_3, v_4, v_5, 0, 0, 0,0]*

de donde el vector de D será:

$$[v_0 + s_7 v_3,\ v_1,\ v_2 + s_8 v_4,\ v_3 + s_8 v_5 + s_9 v_4,\ v_4 + s_9 v_5,\ v_5, 0, 0,\ s_6 v_2,\ s_6 v_3 + s_7 v_2]$$

4.5.3 Plantear un esquema de pesado adecuado para las relaciones identificadas.

Se ha trabajado sólo con relaciones de colocación de dos palabras contiguas, extraídas mediante un analizador sintáctico, mismas que fueron ponderadas empleando el esquema tf.idf, es decir, si f_{ik} es la frecuencia de la relación *i* en el documento *k*, *N* es el número de documentos en la colección y n_i el número de documentos donde la relación *i* ocurre, el peso de la relación *i* en el documento *k*, w_{ik} está dado por:

$$w_{ik} = f_{ik} * \log(N / n_i) \tag{1}$$

Si el vector de la relación *i* en el documento *k* es *[r1, 0, r3, r4, 0, 0, r5]* entonces este vector deberá multiplicarse por el w_{ik} correspondiente para obtener el vector ponderado de la relación

*[w_{ik}*r_1,0,w_{ik}*r_3,w_{ik}*r_4,0,0,w_{ik}*r_5].*

que será el que se sume al vector del documento *k*.

4.5.4 Especificar la función de similitud adecuada para comparar documentos

La similitud empleando sólo términos, se calcula con el producto punto entre vectores. Sea $D_1 = (t_1,...,t_n)$ y $D_2 = (q_1,...,q_n)$ los vectores de dos documentos distintos, se define el producto punto $D_1 \cdot D_2$ de la siguiente manera:

$$D_1 \cdot D_2 = t_1 q_1 + t_2 q_2 + ... + t_n q_n \tag{2}$$

Si los documentos contienen relaciones también empleamos el producto punto:

$$Sim = \left\langle \vec{d} + \delta * \sum_{j=1}^{m} f_j w_j \right\rangle \cdot \left\langle \vec{q} + \delta * \sum_{i=1}^{n} f_i w_i \right\rangle \tag{3}$$

es decir, la similitud está dada por el producto punto de los vectores construidos como el vector de términos ponderado ($\vec{d}, \vec{q}$), más el vector de las frases ponderado multiplicada por un factor (δ) y normalizados.

Tabla 4.1 Términos obtenidos para las colecciones de prueba

Colección	Vocabulario
CISI	5570
CACM	5073
NPL	7754

4.5.5 Evaluar la representación en recuperación de información y explorar las ventajas en otras.

Para RI las consultas se representaron igual que los documentos como se describe en 4.5.1 y una consulta con relaciones como se describe en 4.5.2. De esta forma los documentos con términos compuestos por relaciones, pueden ser evaluados y tener posiciones más altas dentro de la lista de documentos relevantes (ranking). Se realizaron dos experimentos:

- El primero, buscaba comprobar la factibilidad de la representación propuesta, así que sólo se emplearon términos.
- El segundo, empleando frases nominales, orientado a ir incorporando relaciones a la representación evaluando su impacto inicial.

Experimentos

La representación propuesta, se aplicó a tres colecciones tradicionales: CISI, CACM y NPL (http://ir.dcs.gla.ac.uk/resources/test_collections/). CISI tiene 1460 documentos y 112 consultas, CACM 3204 documentos y 64 consultas, y NPL 11,429 documentos y 93 consultas. Se seleccionaron estas colecciones porque son bien conocidas y relativamente pequeñas para realizar las primeras pruebas con la representación y el modelo.

Primer Experimento. Como se mencionó, los primeros experimentos buscaron comprobar la factibilidad de la representación propuesta, así que se realizó recuperación de información, empleando únicamente los términos de las colecciones.

La tabla 4.1 muestra los términos obtenidos para cada colección después de remover palabras vacías y aplicar truncamiento.

Implementación del modelo vectorial

El modelo vectorial clásico se empleó como referencia para comparar sus resultados con los obtenidos por el modelo propuesto. Se implementó usando el esquema de pesado tf.idf ,(1).

Como medida de similitud entre documentos y consultas, se utilizó el coseno, esto es si q es una consulta y d un documento, su similitud está dada por:

$$sim(q,d) = \frac{q \bullet d}{|q||d|} = \frac{\sum_i w_{iq} w_{id}}{\sqrt{\sum_i w_{iq}^2 \sum_i w_{id}^2}}$$

Implementación del modelo propuesto

Para el modelo propuesto, se definieron tantos patrones como términos del vocabulario de cada colección. Los patrones fueron de cinco dígitos binarios, esta longitud se determinó de manera empírica después de realizar experimentos con diferentes longitudes. Los documentos y las consultas se representaron empleando la suma vectorial para combinar los vectores de sus términos.

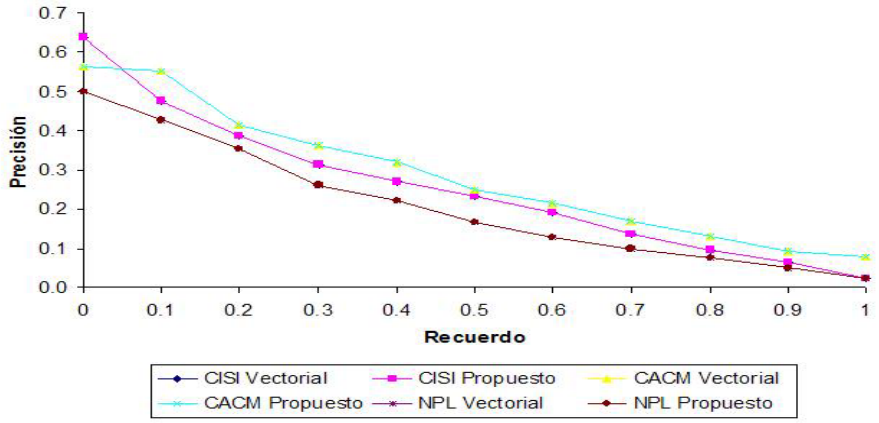

Figura 4.1 Efectividad de la recuperación en CISI, CACM y NPL empleando solo términos.

Los términos se ponderaron con el esquema tf.idf, es decir, se multiplicaron por el w_{ik} correspondiente antes de sumarlos para conformar los vectores de documentos y consultas. El producto punto se utilizó como medida de similitud entre documentos y consultas. La figura 4.1 muestra la gráfica recuerdo-precisión comparando el modelo vectorial clásico con el modelo propuesto. La precisión se calculó a valores estándar de recuerdo promediada entre el número de consultas. La efectividad de la recuperación del modelo vectorial clásico es equivalente a la obtenida con el modelo propuesto para las tres colecciones, por lo que las curvas aparecen traslapadas. Así que estos resultados fueron el punto de partida para los experimentos que se realizaron de manera posterior.

Tabla 4.2 Frase nominales obtenidos para las colecciones de prueba

Colección	Frases nominales
CISI	8940
CACM	9373
NPL	18643

Tabla 4.3 Frases nominales más comunes para las colecciones procesadas

CISI		CACM		NPL	
Frase	Frec.	Frase	Frec.	Frase	Frec.
retriev system	86	comput system	65	magnet field	321
Inform system	72	comput program	59	experiment result	112
Data base	64	program languag	59	electron density	110
Inform retriev	51	digit comput	49	electromagnet wave	94
Inform scienc	46	Oper system	42	electr field	93

Segundo Experimento. Para los siguientes experimentos se consideró la relación más simple que puede extraerse, la de colocación. Se extrajeron frases nominales después de analizar los documentos y consultas con Link Grammar [12], y se seleccionaron las frases nominales formadas por dos palabras contiguas.

La tabla 4.2 muestra el número de frases nominales obtenidas para cada colección y la tabla 4.3 las cinco frases más comunes en cada colección con su frecuencia (Frec.). Para estas frases se

calcularon sus vectores empleando el operador de convolución circular, como se explicó en la sección 4.5.2. Ya que se empleó truncamiento para obtener el vocabulario, también se usó para la obtención de las frases. Estas frases también se agregaron al modelo vectorial como nuevos términos Las frases se pesaron en ambos modelos empleando nuevamente tf.idf.

La similitud entre documentos y consultas para el modelo propuesto se calculó con la fórmula (3), donde factor se tomó igual a un sexto.

Tabla 4.4. Recuerdo-precisión para 76 consultas con frases nominales en CISI.

Recuerdo	Precisión		% Diferencia
	Modelo Vectorial	Modelo Propuesto	
0	0.5871	0.6423	9.40
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

Los resultados de recuerdo-precisión para 76 consultas de CISI (aquellas que tienen documentos relevantes) se muestran en la tabla 4.4. A valores de recuerdo estándar, la precisión mejoró en todos los niveles de recuerdo, alcanzando al inicio hasta un 9.4% de mejora y en promedio 3.06%.

Para comparar los resultados obtenidos, también se empleó la métrica de precisión promedio (*mean average precision MAP*). La media de la precisión promedio (AP) para una consulta está dada como:

$$AP = \left(\sum_{i=1}^{R} i / rango_i \right) \Big/ R$$

donde R es el número de documentos relevantes para la consulta y $rango_i$ es la posición del documento *i* en la lista de documentos ordenada de acuerdo a su relevancia . Entonces para un grupo de consultas, MAP es igual a la media de las APs.

Para las 76 consultas de CISI, la MAP fue de 0.2518 empleando el modelo vectorial clásico y de 0.2568 para el modelo propuesto, teniendo en promedio una mejora del 3.42%. El porcentaje de diferencia mayor para las APs, favoreció al modelo propuesto y se alcanzó en la consulta 43 con un AP de 0.0464 para el vectorial y de 0.0759 para el propuesto, esto es 63.64 % de mejora, tabla A1.1 del apéndice.

Otras métricas empleadas para comparar los resultados fueron: la precisión normalizada (NPREC) y el recuerdo normalizado (NREC) [13] con el fin de estimar los cambios en la posición alcanzada por los documentos relevantes, la precisión normalizada, que responde a cuán rápido se encuentra el primer documento relevante es definida como:

$$NPREC = 1 - \frac{\sum_{i=1}^{REL} \log RANK_i - \sum_{i=1}^{REL} \log i}{\log\left(\frac{N!}{(N-REL)!\,REL!} \right)}$$

donde: $RANK_i$ denota la posición del documento relevante i, REL el número total de documentos relevantes para la consulta y N el tamaño de la colección.
El recuerdo normalizado, que expresa cuánto toma alcanzar el último documento relevante, se define como:

$$NREC = 1 - \frac{\sum_{i=1}^{REL} RANK_i - \sum_{i=1}^{REL} i}{REL(N - REL)}$$

Tabla 4.5 Recuerdo-precisión para 51 consultas con frases nominales en CACM.

Recuerdo	Precisión		% Diferencia
	Modelo Vectorial	Modelo Propuesto	
0	0.6099	0.5842	-4.21
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

Los resultados para la precisión normalizada en CISI, se muestran en las tablas A.1.2, donde puede observarse que la mejora promedio fue de 0.39%, el porcentaje de diferencia mayor se tiene para la consulta 28 con 0.489 para el modelo vectorial y 0.532 para el propuesto, lo que representa un 8.79% de mejora. En la tabla A.1.3 están los resultado del recuerdo normalizado cuya mejora promedio fue del 0.02%. Cabe hacer notar que no se esperan grandes cambios en recuerdo normalizado dado que la posición de los últimos documentos relevantes difícilmente se puede mejorar. Los documentos localizados en las últimas posiciones generalmente comparten muy pocos términos con la consulta lo que implica que el número de relaciones que puedan compartir también será mínimo.

Los resultados para la colección CACM de recuerdo-precisión, para 51 consultas que son las que cuentan con documentos relevantes, se muestran en la tabla 4.5. Puede observarse que sólo en tres puntos de recuerdo los datos son favorables al modelo propuesto, sin embargo la diferencia promedio es de 0.87%, la cual no es muy desventajosa. El MAP para esta colección fue de 0.3155 para el modelo vectorial clásico y de 0.3144 para el propuesto, sin embargo el porcentaje promedio de mejora fue de 1.91%, favorable al modelo propuesto, como se muestra en la tabla A.2.1, donde también puede verse que el mayor porcentaje de diferencia es de 118.39% favorable al modelo propuesto para la consulta 48, que tiene un AP de 0.0564 obtenida con el modelo vectorial y de 0.1231 con el propuesto. Para la precisión normalizada y recuerdo normalizado, el porcentajes promedio de mejora favoreció ligeramente al modelo propuesto, teniendo un 0.18% y 0.01% respectivamente. Para la precisión normalizada la mayor diferencia se tuvo también para la consulta 48 y fue de 16.6 % favorable al modelo propuesto.

Tabla 4.6 Recuerdo-precisión para 92 consultas con frases nominales en NPL.

Recuerdo	Precisión		% Diferencia
	Modelo Vectorial	Modelo Propuesto	
0	0.4430	0.5137	15.96
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	2.1242	2.3293	4.95

Para la colección NPL la tabla 4.6 muestra los resultados de recuerdo-precisión para 92 consultas, se observa que para siete puntos de recuerdo los datos favorecen al modelo propuesto y sólo en cuatro puntos el modelo vectorial clásico resulta ser mejor. La mayor precisión se alcanza en el primer punto de recuerdo obteniendo un 15.96% de mejora con el modelo propuesto, mismo que supera al vectorial clásico en un 4.95% en promedio. El MAP obtenido fue de 0.1891 para el modelo vectorial y de 0.2048 para el modelo propuesto con un promedio de mejora de 11.13%, mostrado en la tabla A.3.1. El mayor porcentaje de diferencia se tiene en la consulta 18, donde la AP para el modelo vectorial es de 0.0551 y de 0.1698 para el propuesto representando un 208.47 % de mejora. El promedio de mejora en NPREC es del 0.21 % teniendo el mayor porcentaje de diferencia en la consulta 52 con 0.6758 para el vectorial y 0.735 para el propuesto, esto es un 8.76% de mejora, como puede verse en la tabla A.3.2. El recuerdo normalizado, en este caso favoreció al modelo vectorial, con una diferencia promedio de -0.12%, tabla A.3.2.

Se realizaron pruebas estadísticas (prueba del signo [22]) para evaluar cuán significativos son los cambios entre el modelo vectorial y el propuesto en términos de precisión a niveles estándar y MAP, empleando el total de consultas con documentos relevantes en las tres colecciones. La hipótesis nula a probar fue que el modelo vectorial en términos de precisión se comporta al menos igual que el modelo con la representación propuesta. Los resultados se muestran en la tabla 4.7. Donde puede verse que para niveles estándar de precisión la hipótesis nula se rechaza para un nivel de significancia $\alpha = 0.01$ para CISI, para NPL $\alpha = 0.2$ y para CACM los datos favorecieron al modelo vectorial. En cuanto a la MAP en NPL se rechaza con $\alpha = 0.05$ y CACM con $\alpha = 0.06$. Dado que no se pudo rechazar la hipótesis nula para CISI en términos de MAP, se realizó la misma prueba para esta colección con la NPREC, rechazándose la hipótesis con $\alpha = 0.06$, teniendo 28 valores a favor del modelo vectorial, 41 del propuesto y 7 iguales (datos extraídos de las tablas 4.4, 4.5, 4.6 y del apéndice).

También se efectuó un análisis cualitativo que permite ver cómo se comporta la representación plasmada en el modelo. A continuación se presentan cuatro consultas por cada colección, primero las dos en las que se obtuvo mejor resultado (MAP) con el modelo propuesto y después las dos en los que los resultados fueron más favorables para el modelo vectorial. Para cada consulta, se presentan los primeros cinco documentos relevantes (o menos si no existe ese número de relevantes), su posición dentro de la lista de documentos recuperados, el recuerdo y la precisión asociados para ambos modelos a fin de observar su comportamiento.

Tabla 4.7 P-valores de diferencias de precisión para las tres colecciones.

Niveles estándar de recuerdo				
Colección	Vectorial	Propuesto	Iguales	Probabilidad de un sólo lado
CISI	0	11	0	0.0017**
CACM	7	4	0	-
NPL	4	7	0	0.18
Mean Average Precision				
Colección	Vectorial	Propuesto	Iguales	Probabilidad de un sólo lado
CISI	38	38	0	-
CACM	18	29	4	0.055
NPL	35	57	0	0.011*
Precisión Normalizada				
Colección	Vectorial	Propuesto	Iguales	Probabilidad de un sólo lado
CISI	28	41	7	0.0594

* significativo con p-valor < α =0.05

** significativo con p-valor < α =0.01

Colección CISI:

Consulta 28: Computerized information systems in fields related to chemistry.

Frases: inform system

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
1460	1	0.0167	1.0000	1460	3	0.0167	0.3333
696	2	0.0333	1.0000	696	8	0.0333	0.2500
116	3	0.0500	1.0000	116	9	0.0500	0.3333
1370*	8	0.0667	0.5000	375	11	0.0667	0.3636
1164*	9	0.0833	0.5556	1092	13	0.0833	0.3846

* Documentos promovidos por sus frases

Consulta 43: The difficulties encountered in information retrieval systems are often less related to the equipment used than to the failure to plan adequately for document analysis, indexing, and machine coding. The position of the programmer is to take a problem and write it in a way in which the equipment will understand. What articles have been written describing research in maximizing the effectiveness of programming.

Frases: analysi index, describ research, machin code, retriev system

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
114	2	0.0769	0.5000	114	5	0.0769	0.2000
135	17	0.1538	0.1176	135	30	0.1538	0.0667
835	25	0.2308	0.1200	835	41	0.2308	0.0732
157	117	0.3077	0.0342	157	109	0.3077	0.0367
350	148	0.3846	0.0338	350	137	0.3846	0.0365

Consulta 57: In catalogs which are either arranged alphabetically or arranged by classification number, the LC entry, printed in readable language, is ultimately important because the individual looking for information has a definite author, title, or subject phrase in his language (probably English in our case) in mind. Will LC entries and subject headings be used in the same manner in automated systems.

Frases: autom system, case mind, lc entri, subject head

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
480	3	0.0556	0.3333	1230	2	0.0556	0.5000
1230	5	0.1111	0.4000	480	3	0.1111	0.6667
1216	19	0.1667	0.1579	1216	7	0.1667	0.4286
388	28	0.2222	0.1429	825	26	0.2222	0.1538
530*	36	0.2778	0.1389	388	40	0.2778	0.1250

* Documento promovido por sus frases

Consulta 7: Describe presently working and planned systems for publishing and printing original papers by computer, and then saving the byproduct, articles coded in data-processing form, for further use in retrieval.

Frases: origin paper

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
376	19	0.1250	0.0526	376	9	0.1250	0.1111
725	125	0.2500	0.0160	725	141	0.2500	0.0142
375	182	0.3750	0.0165	375	172	0.3750	0.0174
310	260	0.5000	0.0154	310	285	0.5000	0.0140
724*	756	0.6250	0.0066	332	742	0.6250	0.0067

* Documento promovido por sus frases

Colección CACM.

Consulta 48: The use of computer science principles (e.g. data structures, numerical methods) in generating optimization (e.g. linear programming) algorithms. This includes issues of the Khachian (Russian, ellipsoidal) algorithm and complexity of such algorithms.

Frases: scienc principl, data structur , program algorithm, numer method

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
1797	6	0.0833	0.1667	2325	27	0.0833	0.0370
2325	8	0.1667	0.2500	1797	35	0.1667	0.0571
2223	28	0.2500	0.1071	1863	59	0.2500	0.0508
1863	29	0.3333	0.1379	1729	60	0.3333	0.0667
1729	31	0.4167	0.1613	2223	61	0.4167	0.0820
2285*	111	0.6667	0.0721	2589	135	0.6667	0.0593

* Documento promovido por sus frases

Consulta 45: The use of operations research models to optimize information system performance. This includes fine tuning decisions such as secondary index selection, file reorganization, and distributed databases.

Frases: distribut database, tune decision, oper research, optim inform, index select, system perform

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
2816	1	0.0385	1.0000	2816	1	0.0385	1.0000
2493	2	0.0769	1.0000	2493	3	0.0769	0.6667
2964	4	0.1154	0.7500	3152	10	0.1154	0.3000
3152	6	0.1538	0.6667	2882	13	0.1538	0.3077
2882	20	0.1923	0.2500	2964	19	0.1923	0.2632

Consulta 20: Graph theoretic algorithms applicable to sparse matrices
Frases: spars matriz

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
2986	25	0.3333	0.0400	2695	9	0.3333,	0.1111
2695	36	0.6667	0.0556	2986	22	0.6667,	0.0909
1563	58	1.0000	0.0517	1563	48	1.0000,	0.0625

Consulta 23: Distributed computing structures and algorithms
Frases: distribut comput

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
3148	14	0.2500	0.0714	3148	1	0.2500	1.0000
2578	33	0.5000	0.0606	2578	35	0.5000	0.0571
3137	47	0.7500	0.0638	3137	57	0.7500	0.0526
2849	84	1.0000	0.0476	2849	72	1.0000	0.0556

Colección NPL

Consulta 18: Diurnal variations of fluctuations in the earths magnetic field
Frases: diurnal variat

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
739	1	0.0769	1.0000	739	17	0.0769	0.0588
345	6	0.1538	0.3333	345	23	0.1538	0.0870
4685	7	0.2308	0.4286	4685	27	0.2308	0.1111
93	27	0.3077	0.1481	1290	41	0.3077	0.0976
1290	72	0.3846	0.0694	93	52	0.3846	0.0962

Consulta 28: The effect of small distortions in the surface of a cavity resonator
Frases: caviti reson, small distort

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
8230	1	0.1250	1.0000	8230	6	0.1250	0.1667
5472	20	0.2500	0.1000	5472	17	0.2500	0.1176
4720	37	0.3750	0.0811	4720	29	0.3750	0.1034
7482	54	0.5000	0.0741	7482	44	0.5000	0.0909
8441	55	0.6250	0.0909	8441	45	0.6250	0.1111

Consulta 92: The phenomenon of radiation caused by charged particles moving in varying electric and magnetic fields
Frases: Ninguna

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
8381	2	0.0714	0.5000	8381	1	0.0714	1.0000
4978	13	0.1429	0.1538	4978	4	0.1429	0.5000
10234	28	0.2143	0.1071	10234	19	0.2143	0.1579
4851	116	0.2857	0.0345	4851	107	0.2857	0.0374
10869*	123	0.3571	0.0407	5516	124	0.3571	0.0403

* Documento promovido por sus frases

Consulta 25: Equations governing the propagation of electromagnetic and hydromagnetic waves in the solar corona
Frases: solar corona

Modelo Propuesto				Modelo vectorial clásico			
Documento	Posición	Recuerdo	Precisión	Documento	Posición	Recuerdo	Precisión
2632	3	0.0137	0.3333	2632	1	0.0137	1.0000
3351	5	0.0274	0.4000	3351	4	0.0274	0.5000
2016	16	0.0411	0.1875	1663	5	0.0411	0.6000
1663	19	0.0548	0.2105	2016	6	0.0548	0.6667
7613*	21	0.0685	0.2381	8741	8	0.0685	0.6250

* Documento promovido por sus frases

Las tablas mostradas nos permiten observar que de las seis consultas que resultan favorables al modelo propuesto, en cuatro de ellas la precisión para el primer documento recuperado es de 1.00, en la consulta 28 de CISI se tiene esta precisión para los tres primeros documentos y en la 45 de CACM para los dos primeros. El modelo vectorial presenta un comportamiento similar, puede observarse que de las seis consultas que le son favorables, en tres de ellas también alcanza una precisión de 1.00, únicamente para el primer documento relevante que se recupera.

Cabe hacer notar que a partir del análisis cualitativo se observan los cambios en el orden en el que aparecen los documentos, indicando que documentos con detalles de los solicitados son promovidos.

Para ejemplificar la utilidad de la representación propuesta en otras tareas de tratamiento de información consideremos la búsqueda de respuesta.

Supongamos que nos interesa contestar la pregunta: *Who was Pilates?*, después procesarla con un etiquetador de entidades nombradas obtenemos:

```
<?xml version="1.0" encoding="ISO-8859-15" ?>
    -<output>
      - <s i="0">
            Who was
                <ENAMEX TYPE="PERSON"> Pilates</ENAMEX>
             ?
        </s>
    </output>
```

Ahora sabremos que Pilates es una entidad nombrada de tipo persona. Utilizamos esta información para trasladar la pregunta a la representación propuesta y obtenemos el vector C definido como:

$$\vec{C} = \langle per \otimes Pil\bar{a}tes \rangle$$

Si se tiene dos textos D1 y D2, y queremos saber cuál contiene la respuesta con mayor probabilidad, procederíamos a etiquetarlos y trasladarlos a la representación propuesta:
D1: Pilates, (1880-1967), born in Germany was a sickly child who developed an avid interest in athletics, physical fitness and yoga. He developed a method called "Contrology", related to encouraging the use of the mind to control muscles

Resultado del etiquetador:

```
<?xml version="1.0" encoding="ISO-8859-15" ?>
- <output>
- <s i="0">
  <ENAMEX TYPE="PERSON">Pilates</ENAMEX>
  , (1880-1967), born in
  <ENAMEX TYPE="LOCATION">Germany</ENAMEX>
   was a sickly child who developed an avid interest in athletics, physical
fitness and yoga.
  </s>
  He developed a method called "Contrology", related to encouraging the use of
the mind to control muscles
  </output>
```

Del este resultado, obtenemos que Pilates es una entidad nombrada de tipo persona y Germany es una de tipo localidad, así que el vector de D1 será la suma de los vectores de sus términos, más la codificación de las entidades identificadas:

$$\vec{D}1 = \langle \vec{t}_1 + \vec{t}_2 + ... + \vec{t}_n + (per \otimes Pilates) + (loc \otimes Germany) \rangle$$

D2: Imagine an exercise program that you look forward to, that engages you, and that leaves you refreshed and alert with a feeling of physical and mental well-being. The pilates method will do all this and more.

Resultado del etiquetador

```
<?xml version="1.0" encoding="ISO-8859-15" ?>
- <output>
   <s i="0">Imagine an exercise program that you look forward to, that engages
you, and that leaves you refreshed and alert with a feeling of physical and
mental well-being.</s>
  <s i="1">The pilates method will do all this and more.</s>
  </output>
```

Por el resultado del etiquetador, el vector de D2 será sólo la suma de los vectores de sus términos:

$$\vec{D}2 = \langle \vec{t}_1 + \vec{t}_2 + ... + \vec{t}_n \rangle$$

y en consecuencia el documento D1, de manera correcta, se elegirá como el que contiene la respuesta con mayor probabilidad, dado que comparte con la consulta la codificación $per \otimes Pilates$

Para probar la utilidad de la representación propuesta en la tarea de agrupamiento, se realizó un experimento sencillo: un documento de cada colección fue considerado como consulta, tratando de agrupar de abajo hacia arriba (bottom-up). A continuación se muestran para cada colección el identificador de los documentos, el texto empleado como consulta, el texto de los primeros dos documentos relevantes y su similitud con la consulta.

Colección CISI:

ID	Texto (consulta)	
33	The "**half life**" of some scientific and technical literatures. A consideration of the analogy between the **half-life** of radioactive substances and the rate of obsolescence of scientific literature. The validity of this analogy suggest the possibility of more accurate prognostications concerning the period of time during which scientific literature may by used and hence might help to guide the planning of library collections and technical information services..	
ID	Texto	Sim.
793	The '**half life**' of periodical literature: apparent and real obsolescence. The expression '**half-life**', borrowed from physics, has appeared quite frequently in the literature on documentation since 1960, when an article by Burton and Kebler on The '**half-life**' of some scientific and technical literatures was published, although it had certainly been used previously. Burton and Kebler point out that literature becomes obsolescent rather than disintegrating (as in its original meaning), so that '**half-life**' means 'half the active life', and this is commonly understood as meaning the time during which one-half of the currently active literature was published. Numerous studies have been carried out, mainly by the analysis of citations, to establish obsolescence rates of the literature of different subjects. Bourne points out that different studies have given widely different results, so that many of the '**half-life**' figures reported are not valid beyond the particular sample of literature or users surveyed; certainly they cannot be used as accurate measures for discriminating between different subject-fields.	0.44
764	Progress in documentation. The term 'obsolescence' occurs frequently in the literature of librarianship and information science. In numerous papers we are told how most published literature becomes obsolete within a measurable time, and that an item receives half the uses it will ever receive ('**half-life**') in a few years. 'Obsolescence' is however very rarely defined, and its validity, interest, and practical value are often assumed rather than explained. Before reviewing studies on 'obsolescence', therefore, it is necessary to look at the concept and to identify the reasons why it should be of interest.	0.3

El primer documento, que habla de un tema diferente, tiene similitud de 0.2 con la consulta y se ubica en la posición 4, debe notarse la presencia de la frase: *half life* en los documentos con mayor similitud a la consulta.

Colección CACM

ID	Texto (consulta)	
1237	Conversion of **Decision Tables** To Computer Programs Several translation procedures for the conversion of **decision tables** to programs are presented and then evaluated in terms of storage requirements, execution time and compile time. The procedures are valuable as hand-coding guides or as algorithms for a compiler. Both limited-entry and extended-entry tables are analyzed. In addition to table analysis, the nature of table-oriented programming languages and features is discussed. It is presumed that the reader is familiar with the nature of **decision tables** and conventional definitions.	
ID	Texto	Sim.
2053	On the Conversion of **Decision Tables** to Computer Programs The use of execution time diagnostics in pinpointing ambiguities in decision tables is discussed. It is pointed out that any attempt at resolving ambiguities at compile time will, in general, be impossible. It is shown that, as a consequence, tree methods of converting decision tables to programs are inadequate in regard to ambiguity detection. Two algorithms for programming decision tables whose merits are simplicity of implementation and detection of ambiguities at execution time are presented. The first algorithm is for limited entry **decision tables** and clarifies the importance of proper coding of the information in the decision table. The second algorithm programs a mixed entry **decision table** directly without going through the intermediate step of conversion to a limited entry form, thereby resulting in storage economy. A comparison of the algorithms and others proposed in the literature is made. Some features of a **decision table** to Fortran IV translator for the IBM 7044 developed by the authors are given.	0.63
2221	Comment on the Conversion of **Decision Tables** to Computer Programs	0.59

En este grupo de documentos puede observarse la presencia de la frase *decision table*, la similitud con el primer documento, que ya no toca el tema, es de 0.49, en la posición 10.

Colección NPL

ID	Texto (consulta)	
6368	dissected amplifiers using **negative resistance** possible circuit arrangements of semiconductor devices to give hf amplification are discussed based on segregation of the amplifier properties of **negative resistance** and directionality an arrangement of a hexagonal ge plate with magnetic field perpendicular to its plane and with an appropriate network of resistances corresponds to a pentode amplifier	
ID	Texto	Sim.
8853	**negative resistance** amplifier design	0.43
2808	a **negative resistance** for dc computers a nomogram is presented for the design of a circuit using transistors and resistors to give any required value of **negative resistance**	0.29

Este grupo de documentos comparte la frase: *negative resistance*, la similitud con el primer documento de un tema diferente es de 0.19, posición 21.

4.6 Conclusiones preliminares

Los experimentos realizados dan un indicio de que la representación propuesta para la codificación de relaciones tiene la posibilidad de contribuir a mejorar la precisión, permitiendo representar documentos y consultas de manera más detallada.

La representación propuesta da auspicios de ser benéfica en otras tareas de tratamiento de información. Por ejemplo en búsqueda de respuesta, puede contribuir a una adecuada selección de la respuesta y en el agrupamiento, a asociar documentos con atributos más específicos

Concluimos que la precisión no empeora con el modelo propuesto y que si se enriquece dicho modelo agregando nuevas relaciones como entidades nombradas, sujeto-verbo, verbo-objeto, sustantivo-adjetivo, verbo-adverbio, o relaciones que reflejen conocimiento de un dominio particular, podrá alcanzarse una mejora mayor.

4.7 Plan de trabajo

En la siguiente tabla se presenta el cronograma de actividades para los 36 meses (9 cuatrimestres) de duración de la investigación doctoral.

	2007			2008			2009		
Actividad	1	2	3	4	5	6	7	8	9
Definición del tema									
Estudio del estado del arte									
Determinación de la representación									
Especificación de las relaciones a extraer									
Elección de herramientas									
Obtención de colecciones de pruebas									
Formulación de la representación									
Instrumentación del sistema de RI									
Validación con diferentes colecciones									
Redacción de propuesta de investigación									
Redacción de reporte de avance									
Redacción de artículos para congresos									
Redacción del documento de tesis									
Defensa de tesis									

Concluida
En proceso
No iniciada

Durante el primer año se definió de manera precisa el tema de investigación, se inició el estudio del estado del arte, se determinó la representación para los documentos, se inició la

especificación de las relaciones, se buscaron herramientas para iniciar la extracción de relaciones, se obtuvieron colecciones para realizar pruebas con la representación propuesta.

En el segundo año se continuará con las actividades iniciadas durante el primer año, se preparará el reporte de avance, se iniciará la redacción del documento de tesis y de artículos para presentarlos en los foros que se identifiquen.

Durante el último año se concluirá la instrumentación y validación de la representación, la redacción del documento de tesis iniciada desde el segundo cuarto del segundo año se concluirá en el tercer cuarto del tercer año, para reservar el último cuarto del tercer año, a preparar la defensa de tesis y realizar los ajustes necesarios.

Referencias

1. Shuming Shi, Ji-Rong Wen, Qing Yu, Ruihua Song, Wei-Ying Ma, Gravitation-Based Model for Information Retrieval, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval 2005, pp. 488-495, Salvador, Brazil August 15 - 19, 2005.
2. Alexandre Gonçalves, Jianhan Zhu, Dawei Song, Victoria Uren, Roberto Pacheco, LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval, In Proc. of The Seventh International Conference on Web- Age Information Management (WAIM 2006), pages 122-133, June, Hong Kong, China.
3. Jörg Becker, Dominik Kuropa, Topic-based Vector Space Model, Proceedings of the 6th International Conference on Business Information Systems, July 2003 Colorado,USA, pp. 7-13.
4. David D. Lewis, Karen Sparck Jones, Natural language processing for information retrieval. Communications ACM 39, Jan. 1996, pp. 92-101.
5. Mandar Mitra, Chris Buckley, Amit Singhal, Claire Cardie, An analysis of statistical and syntactic phrases in Proceedings of RIAO-97, 5th International Conference, pp. 200-214.
6. David A.Evans, Chengxiang Zhai, Noun-phrase analysis in unrestricted text for information retrieval, In Proceedings of the 34th Annual Meeting on Association For Computational Linguistics, June 1996, pp. 17-24.
7. Miguel A. Alonso, Jesús Vilares, and Víctor M. Darriba, On the Usefulness of Extracting Syntactic Dependencias for Text Indexing, in Michael O'Neill, Richard F. E. Sutcliffe, Conor Ryan, Malachy Eaton and Niall J. L. Griffith (eds.), Artificial Intelligence and Cognitive Science, volume 2464 of Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin-Heidelberg-New York, 2002, pp. 3-11.
8. Jesus Vilares, Miguel A. Alonso and Manuel Vilares, Morphological and syntactic processing for Text Retrieval, in Fernando Galindo, Makoto Takizawa and Roland Traunmüller (eds.), Database and Expert Systems Applications, volume 3180 of Lecture Notes in Computer Science, Springer-Verlag, Berlin-Heidelberg-New York, 2004, pp. 371-380.
9. Jesús Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso, Managing Syntactic Variation in Text Retrieval, in Peter R. King, Proceedings of the 2005 ACM Symposium on Document Engineering. November 2-4, 2005. Bristol, United Kingdom, ACM Press, New York, USA, 2005, pp. 162-164.
10. Aurelio López-López, Sung H. Myaeng, Extending the Capabilities of Retrieval Systems by a Two-Level Representation of Content, Aurelio López-López and Sung H. Myaeng, *Proceedings of the Australian Document Computing Symposium*, Justin Zobel (Ed.), Part I, 1996, pp. 15-20
11. Tony A. Plate, Analogy retrieval and processing with distributed vector representation, Technical report CS-TR-98-4 Victoria University of Wellington, Computer Science. 16 p. Longer version of an invited submission to the Workshop on Advances in Analogy Research held at New Bulgarian University, Sofia, Bulgaria, July 1998.
12. Dennis Grinberg, John Lafferty and Daniel Sleator. A robust parsing algorithm for link grammars. Carnegie Mellon University, Computer Science technical report CMU-CS-95-125, and Proceedings of the Fourth International Workshop on Parsing Technologies, Prague,September, (1995). 17p.
13. C.J. van Rijsbergen, Information Retrieval, online book (http://www.dcs.gla.ac.uk/~iain/keith/).
14. Tony A. Plate. Holographic Reduced Representation, Distributed Representation for Cognitive Structures, CSLI Publications, 2003.

15. Gerald Kowalski, Information Retrieval Systems Theory and Implementation, Kluwer Academic Publishers, 1997.
16. Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, ACM press, 1999.
17. David A. Grossman, Ophir Frieder, Infromation Retrieval Algorithms and Heuristics, Springer, 2004.
18. A.T. Arampatzis, Th.P.van der Weide, P. van Bommel, C.H.A. Koster, Linguistically-motivated Information Retrieval. Technical Report CSI-R9918, Dept. of Information system, Faculty of Mathematics and Computer Science, University of Nijmegen, Netherlands, September 1999.
19. Amit Singhal, Modern Information Retrieval: A Brief Overview, IEEE Data Engineering Bulletin,Volume: 24, Issue: 4, p. 35 – 43, 2001.
20. Keith van Rijsbergen, The Geometry of Information Retrieval, Cambridge University Press, 2004.
21. Jhon I. Teit (editor), Chartering a New Course: Natural Language Processing and Information Retrieval, Essays in Honor of Karen Spärck Jones, Springer 2005.
22. John E. Freund, Ronald E. Walpole, Estadística Matemática con Aplicaciones, 4ª. Edición, Prentice-Hall.
23. Guy Lebanon, Yi Mao, Joshua Dillon, The Locally Weighted Bag of Words Framework for Document Representation, Journal of Machine Learning Research 8, 2007, pp. 2405-2441.
24. Renxu Sun, Hang Cui, Keya Li, Min- Yen Kan, Tat-Seng Chua, Dependency Relation Matching for Answer Selection, Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Poster Sesion, Salvador, Brazil August 15 - 19, 2005 pp.651-652
25. Hang Cui, Renxu Sun, Keya Li, Min-Yen Kan Tat-Seng Chua, Question Answering Passage Retrieval Using Dependency Relations, Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil August 15 - 19, 2005, pp. 400-407.
26. Azam Jalali, Farhad Oroumchian, Rich Document Representation for Document Clustering, Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, Avignon (Vaucluse), France, vol.1, April 2004, pp. 802-808.
27. Matthew Lease. Natural Language Processing for Information Retrieval: the time is ripe (again), In Proceedings of the 1st Ph.D. Workshop at the ACM Conference on Information and Knowledge Management (PIKM), 2007, pp. 1-8.
28. Cheng Xiang Zhai, A brief review of information retrieval models, Technical report, Department of Computer Science, University of Illinois at Urbana-Champaign, 2007.
29. James Allan, Jaime Carbonell, George Doddingtonz, Jonathan Yamronx, and Yiming Yang, Topic Detection and Tracking Pilot Study Final Report, in Proceedings of the DARPA Broadcast News Transcriptions and Understanding Workshop, 1998, pp. 194-218.

Anexo A: Tablas MAP, NPREC y NREC

A.1 Colección CISI

Tabla A.1.1 Mean Average Precision para 76 consultas en CISI

Query ID	AP vectorial	AP propuesta	% Diferencia
.Q001	0.5353	0.5655	5.65
.Q002	0.0351	0.0343	-2.08
.Q003	0.4705	0.4342	-7.71
.Q004	0.0648	0.0853	31.65
.Q005	0.0483	0.0499	3.33
.Q006	0.0278	0.0204	-26.62
.Q007	0.0237	0.0166	-29.92
.Q008	0.0309	0.0358	16.04
.Q009	0.1995	0.1903	-4.65
.Q010	0.3114	0.2761	-11.32
.Q011	0.2615	0.2458	-6.00
.Q012	0.0567	0.0627	10.52
.Q013	0.2820	0.3008	6.67
.Q014	0.0131	0.0146	11.45
.Q015	0.2177	0.2290	5.20
.Q016	0.0589	0.0818	38.95
.Q017	0.0664	0.0660	-0.65

Query ID	AP vectorial	AP propuesta	% Diferencia
.Q018	0.2895	0.2880	-0.51
.Q019	0.3254	0.3188	-2.02
.Q020	0.2033	0.2024	-0.47
.Q021	0.0891	0.0832	-6.63
.Q022	0.1004	0.1013	0.90
.Q023	0.1077	0.1130	4.89
.Q024	0.3566	0.3448	-3.29
.Q025	0.2391	0.2158	-9.72
.Q026	0.4422	0.4463	0.93
.Q027	0.2937	0.2903	-1.15
.Q028	0.1399	0.2194	56.87
.Q029	0.1824	0.2084	14.24
.Q030	0.5042	0.5037	-0.10
.Q031	0.1850	0.1961	5.98
.Q032	0.2344	0.2466	5.19
.Q033	0.0532	0.0700	31.76
.Q034	0.1842	0.1787	-2.98
.Q035	0.2009	0.2029	1.01
.Q037	0.2322	0.2309	-0.57
.Q039	0.0710	0.0811	14.28
.Q041	0.3800	0.3635	-4.32
.Q042	0.1379	0.1251	-9.28
.Q043	0.0464	0.0759	63.64
.Q044	0.3331	0.3278	-1.58
.Q045	0.1515	0.1488	-1.81
.Q046	0.3512	0.3601	2.55
.Q049	0.1862	0.1681	-9.70
.Q050	0.4791	0.4801	0.22
.Q052	0.6117	0.6327	3.43
.Q054	0.1692	0.1654	-2.27
.Q055	0.7638	0.7965	4.27
.Q056	0.1185	0.1128	-4.79
.Q057	0.1792	0.1316	-26.57
.Q058	0.4878	0.5216	6.94
.Q061	0.0388	0.0464	19.60
.Q062	0.6041	0.5783	-4.27
.Q065	0.5682	0.5720	0.66
.Q066	0.6103	0.5968	-2.20
.Q067	0.1298	0.1423	9.63
.Q069	0.1923	0.1812	-5.78
.Q071	0.3560	0.3528	-0.88
.Q076	0.5201	0.5384	3.52
.Q079	0.2827	0.2664	-5.76
.Q081	0.1181	0.1565	32.54
.Q082	0.0890	0.0774	-13.09
.Q084	0.1315	0.1255	-4.59
.Q090	0.1211	0.1257	3.83
.Q092	0.1421	0.1429	0.58
.Q095	0.2763	0.2514	-8.98
.Q096	0.2410	0.1991	-17.39
.Q097	0.4385	0.5564	26.89
.Q098	0.4100	0.4129	0.71
.Q099	0.3734	0.3677	-1.51
.Q100	0.0953	0.0851	-10.67
.Q101	0.3333	0.5000	50.02
.Q102	0.5751	0.5769	0.30
.Q104	0.0794	0.0908	14.24
.Q109	0.1929	0.1870	-3.08
.Q111	0.6834	0.7230	5.79
MAP	0.2518	0.2568	3.42

Tabla A.1.2 Precisión normalizada para 76 consultas en CISI

Query ID	NPREC Vectorial	NPREC Propuesta	% Diferencia
.Q001	0.8293	0.8415	1.47
.Q002	0.3492	0.3449	-1.23
.Q003	0.7444	0.7292	-2.04
.Q004	0.5948	0.6198	4.2
.Q005	0.4039	0.4129	2.23
.Q006	0.7541	0.7329	-2.81
.Q007	0.3812	0.3693	-3.12
.Q008	0.4072	0.4314	5.94
.Q009	0.5532	0.5611	1.43
.Q010	0.6968	0.7032	0.92
.Q011	1.0000	1.0000	0
.Q012	0.5230	0.5310	1.53
.Q013	0.6540	0.6663	1.88
.Q014	0.5899	0.5994	1.61
.Q015	0.5945	0.6043	1.65
.Q016	0.4377	0.4720	7.84
.Q017	0.3105	0.3070	-1.13
.Q018	0.7152	0.7141	-0.15
.Q019	0.6779	0.6749	-0.44
.Q020	1.0000	1.0000	0
.Q021	0.5083	0.5012	-1.4
.Q022	0.4892	0.4908	0.33
.Q023	0.4709	0.4825	2.46
.Q024	0.6869	0.6796	-1.06
.Q025	0.684	0.6676	-2.4
.Q026	0.7768	0.7787	0.24
.Q027	1.0000	1.0000	0
.Q028	0.4890	0.5320	8.79
.Q029	0.6096	0.6213	1.92
.Q030	1.0000	1.0000	0
.Q031	0.5393	0.5442	0.91
.Q032	1.0000	1.0000	0
.Q033	0.4324	0.4513	4.37
.Q034	0.5835	0.5770	-1.11
.Q035	0.6113	0.6100	-0.21
.Q037	0.6144	0.6165	0.34
.Q039	0.4756	0.4817	1.28
.Q041	0.7609	0.7435	-2.29
.Q042	0.5148	0.5023	-2.43
.Q043	0.4486	0.4659	3.86
.Q044	1.0000	1.0000	0
.Q045	0.5052	0.5041	-0.22
.Q046	1.0000	1.0000	0
.Q049	0.5276	0.5162	-2.16
.Q050	0.7541	0.7550	0.12
.Q052	0.8574	0.8587	0.15
.Q054	0.5773	0.5722	-0.88
.Q055	0.9149	0.9230	0.89
.Q056	0.5179	0.5031	-2.86
.Q057	0.6294	0.6034	-4.13
.Q058	0.7781	0.7862	1.04
.Q061	0.4619	0.4662	0.93
.Q062	0.8451	0.8392	-0.7
.Q065	0.8237	0.8254	0.21
.Q066	0.8413	0.8367	-0.55
.Q067	0.5531	0.5594	1.14
.Q069	0.6283	0.6266	-0.27
.Q071	0.7034	0.7048	0.2
.Q076	0.7905	0.7958	0.67
.Q079	0.695	0.6796	-2.22
.Q081	0.5146	0.5180	0.66

Query ID	NPREC Vectorial	NPREC Propuesta	% Diferencia
.Q082	0.5675	0.5705	0.53
.Q084	0.6600	0.6572	-0.42
.Q090	0.4546	0.4622	1.67
.Q092	0.4696	0.4729	0.7
.Q095	0.7214	0.7047	-2.31
.Q096	0.7728	0.7474	-3.29
.Q097	0.8676	0.8927	2.89
.Q098	0.7411	0.7474	0.85
.Q099	0.6622	0.6666	0.66
.Q100	0.3468	0.3381	-2.51
.Q101	0.9246	0.9524	3.01
.Q102	0.8007	0.7998	-0.11
.Q104	0.5102	0.5162	1.18
.Q109	0.5368	0.5396	0.52
.Q111	0.9027	0.9078	0.56
% Total	0.6549	0.6567	0.39

Tabla A.1.3 Recuerdo normalizado para 76 consultas en CISI.

Query ID	NREC Vectorial	NREC Propuesta	% Diferencia
.Q001	0.9542	0.9566	0.25
.Q002	0.6907	0.6887	-0.29
.Q003	0.8986	0.8952	-0.38
.Q004	0.9185	0.9231	0.5
.Q005	0.7211	0.7296	1.18
.Q006	0.9760	0.9671	-0.91
.Q007	0.6801	0.6774	-0.4
.Q008	0.8209	0.8335	1.53
.Q009	0.8197	0.8223	0.32
.Q010	0.9227	0.9316	0.96
.Q011	0.8030	0.8031	0.01
.Q012	0.8850	0.8839	-0.12
.Q013	0.8710	0.8738	0.32
.Q014	0.9194	0.9242	0.52
.Q015	0.8419	0.8445	0.31
.Q016	0.7732	0.7807	0.97
.Q017	0.5917	0.5888	-0.49
.Q018	0.9491	0.9486	-0.05
.Q019	0.8754	0.875	-0.05
.Q020	0.7839	0.7876	0.47
.Q021	0.8344	0.8303	-0.49
.Q022	0.8150	0.8149	-0.01
.Q023	0.7537	0.7624	1.15
.Q024	0.8708	0.8679	-0.33
.Q025	0.9159	0.9133	-0.28
.Q026	0.9358	0.9361	0.03
.Q027	0.8814	0.8804	-0.11
.Q028	0.7407	0.7457	0.68
.Q029	0.8893	0.8888	-0.06
.Q030	0.9262	0.9254	-0.09
.Q031	0.7941	0.7946	0.06
.Q032	0.7246	0.7255	0.12
.Q033	0.7661	0.7731	0.91
.Q034	0.8386	0.8353	-0.39
.Q035	0.8613	0.8591	-0.26
.Q037	0.8556	0.8576	0.23
.Q039	0.7906	0.7914	0.1
.Q041	0.9654	0.9501	-1.58
.Q042	0.7712	0.7693	-0.25
.Q043	0.7939	0.7797	-1.79
.Q044	0.7867	0.7883	0.2
.Q045	0.7690	0.7711	0.27
.Q046	0.8489	0.847	-0.22

Query ID	NREC Vectorial	NREC Propuesta	% Diferencia
.Q049	0.7639	0.7658	0.25
.Q050	0.8653	0.8659	0.07
.Q052	0.9577	0.956	-0.18
.Q054	0.8458	0.8453	-0.06
.Q055	0.9681	0.9683	0.02
.Q056	0.8169	0.8053	-1.42
.Q057	0.8934	0.8897	-0.41
.Q058	0.9082	0.9065	-0.19
.Q061	0.8090	0.807	-0.25
.Q062	0.9505	0.9479	-0.27
.Q065	0.9595	0.9573	-0.23
.Q066	0.9341	0.9334	-0.07
.Q067	0.8460	0.8459	-0.01
.Q069	0.8474	0.8473	-0.01
.Q071	0.8886	0.8888	0.02
.Q076	0.9136	0.9119	-0.19
.Q079	0.9211	0.9216	0.05
.Q081	0.7308	0.728	-0.38
.Q082	0.9048	0.9124	0.84
.Q084	0.9533	0.9544	0.12
.Q090	0.6933	0.6975	0.61
.Q092	0.7756	0.7817	0.79
.Q095	0.9243	0.9164	-0.85
.Q096	0.9824	0.9799	-0.25
.Q097	0.9936	0.9938	0.02
.Q098	0.9110	0.9148	0.42
.Q099	0.8267	0.8320	0.64
.Q100	0.5213	0.5226	0.25
.Q101	0.9986	0.9993	0.07
.Q102	0.8998	0.8993	-0.06
.Q104	0.8199	0.8171	-0.34
.Q109	0.7576	0.7600	0.32
.Q111	0.9892	0.9877	-0.15
% Total	0.8500	0.8500	0.02

A.2 Colección CACM

Tabla A.2.1 Mean Average Precision para 51 consultas en CACM

Query ID	AP vectorial	AP propuesta	% Diferencia
Q01	0.1835	0.1775	-3.30
.Q03	0.2216	0.2221	0.20
.Q04	0.0720	0.0614	-14.79
.Q05	0.0879	0.0897	1.98
.Q06	0.1422	0.1517	6.71
.Q07	0.1658	0.1762	6.28
.Q08	0.1366	0.1356	-0.73
.Q09	0.1745	0.1817	4.09
.Q10	0.7756	0.7959	2.61
.Q11	0.4261	0.4051	-4.94
.Q12	0.3051	0.2936	-3.78
.Q13	0.3114	0.3477	11.66
.Q14	0.3875	0.3947	1.86
.Q15	0.1301	0.1424	9.43
.Q16	0.0417	0.0454	8.85
.Q17	0.1633	0.1666	1.99
.Q18	0.2512	0.2587	2.99
.Q19	0.4940	0.4343	-12.09
.Q20	0.0882	0.0491	-44.31
.Q21	0.0488	0.0443	-9.34
.Q22	0.5852	0.5838	-0.23
.Q23	0.2913	0.0609	-79.11
.Q24	0.4405	0.4443	0.85

Query ID	AP vectorial	AP propuesta	% Diferencia
.Q25	0.2136	0.2434	13.97
.Q26	0.4031	0.3973	-1.44
.Q27	0.2955	0.3077	4.13
.Q28	0.6572	0.6572	0.00
.Q29	0.7578	0.7952	4.94
.Q30	0.1403	0.1673	19.27
.Q31	1.0000	1.0000	0.00
.Q32	0.4333	0.4294	-0.91
.Q33	0.1250	0.1000	-20.00
.Q36	0.1994	0.2039	2.25
.Q37	0.1488	0.1904	27.94
.Q38	0.5808	0.5057	-12.94
.Q39	0.2989	0.3019	1.00
.Q40	0.3568	0.3210	-10.03
.Q42	0.0546	0.0579	6.12
.Q43	0.1802	0.1822	1.11
.Q44	0.2019	0.2003	-0.83
.Q45	0.1832	0.2345	27.98
.Q48	0.0564	0.1231	118.39
.Q49	0.1421	0.1277	-10.13
.Q57	1.0000	1.0000	0.00
.Q58	0.3623	0.3747	3.42
.Q59	0.3778	0.3998	5.82
.Q60	0.3491	0.3853	10.38
.Q61	0.2045	0.2443	19.45
.Q62	0.0644	0.0698	8.37
.Q63	0.3801	0.3512	-7.61
.Q64	1.0000	1.0000	0.00
MAP	0.3155	0.3144	1.91

Tabla A.2.2 Precisión normalizada para 51 consultas en CACM

Query ID	NPREC vectorial	NPREC propuesta	% Diferencia
Q01	0.7495	0.7404	-1.21
.Q03	0.6654	0.6678	0.36
.Q04	0.5652	0.5567	-1.5
.Q05	0.663	0.6641	0.17
.Q06	0.8044	0.8135	1.13
.Q07	0.6153	0.623	1.25
.Q08	0.7484	0.7465	-0.25
.Q09	0.7095	0.7074	-0.3
.Q10	0.932	0.9362	0.45
.Q11	0.8135	0.8052	-1.02
.Q12	0.8094	0.8076	-0.22
.Q13	0.7981	0.8171	2.38
.Q14	0.7872	0.7881	0.11
.Q15	0.7265	0.7418	2.11
.Q16	0.4815	0.4889	1.54
.Q17	0.688	0.6957	1.12
.Q18	0.6973	0.7089	1.66
.Q19	0.8879	0.8765	-1.28
.Q20	0.7584	0.7026	-7.36
.Q21	0.5636	0.5535	-1.79
.Q22	0.8847	0.8829	-0.2
.Q23	0.7661	0.6977	-8.93
.Q24	0.7837	0.7863	0.33
.Q25	0.6627	0.6753	1.9
.Q26	0.7589	0.7584	-0.07
.Q27	0.7171	0.72	0.4
.Q28	0.9339	0.9339	0
.Q29	0.9169	0.925	0.88
.Q30	0.7694	0.7717	0.3
.Q31	1	1	0

Query ID	NPREC vectorial	NPREC propuesta	% Diferencia
.Q32	0.8717	0.8676	-0.47
.Q33	0.8712	0.8574	-1.58
.Q36	0.672	0.6739	0.28
.Q37	0.5893	0.596	1.14
.Q38	0.8648	0.8447	-2.32
.Q39	0.7708	0.7566	-1.84
.Q40	0.7659	0.7499	-2.09
.Q42	0.5206	0.5301	1.82
.Q43	0.6362	0.6353	-0.14
.Q44	0.6151	0.611	-0.67
.Q45	0.6557	0.6752	2.97
.Q48	0.589	0.6868	16.6
.Q49	0.7291	0.7243	-0.66
.Q57	1	1	0
.Q58	0.707	0.7076	0.08
.Q59	0.7899	0.7946	0.6
.Q60	0.7844	0.7834	-0.13
.Q61	0.7176	0.7423	3.44
.Q62	0.6448	0.6521	1.13
.Q63	0.8571	0.8482	-1.04
.Q64	1	1	0
% Total	0.7512	0.7516	0.18

Tabla A.2.3 Recuerdo normalizado para 51 consultas en CACM

Query ID	NREC vectorial	NREC propuesta	% Diferencia
Q01	0.9624	0.9584	-0.42
.Q03	0.9627	0.964	0.14
.Q04	0.9187	0.9166	-0.23
.Q05	0.9685	0.967	-0.15
.Q06	0.9959	0.9964	0.05
.Q07	0.8585	0.8581	-0.05
.Q08	0.9854	0.9853	-0.01
.Q09	0.9559	0.9537	-0.23
.Q10	0.9862	0.9862	0
.Q11	0.9773	0.9772	-0.01
.Q12	0.9893	0.9903	0.1
.Q13	0.9851	0.9862	0.11
.Q14	0.9538	0.9519	-0.2
.Q15	0.9881	0.9898	0.17
.Q16	0.8742	0.8796	0.62
.Q17	0.9634	0.9644	0.1
.Q18	0.9713	0.9724	0.11
.Q19	0.996	0.996	0
.Q20	0.9924	0.9882	-0.42
.Q21	0.9362	0.9355	-0.07
.Q22	0.9922	0.9918	-0.04
.Q23	0.9879	0.9869	-0.1
.Q24	0.9745	0.9746	0.01
.Q25	0.9257	0.9265	0.09
.Q26	0.9468	0.9474	0.06
.Q27	0.9429	0.9423	-0.06
.Q28	0.9984	0.9984	0
.Q29	0.9736	0.9736	0
.Q30	0.9893	0.9867	-0.26
.Q31	1	1	0
.Q32	0.9964	0.9959	-0.05
.Q33	0.9978	0.9972	-0.06
.Q36	0.9223	0.9238	0.16
.Q37	0.8798	0.8772	-0.3
.Q38	0.982	0.9812	-0.08
.Q39	0.966	0.9639	-0.22
.Q40	0.9692	0.9672	-0.21

Query ID	NREC vectorial	NREC propuesta	% Diferencia
.Q42	0.8923	0.8972	0.55
.Q43	0.9062	0.9041	-0.23
.Q44	0.9329	0.9322	-0.08
.Q45	0.9445	0.946	0.16
.Q48	0.9621	0.9749	1.33
.Q49	0.9806	0.9823	0.17
.Q57	1	1	0
.Q58	0.8999	0.9025	0.29
.Q59	0.967	0.9667	-0.03
.Q60	0.9779	0.9739	-0.41
.Q61	0.9685	0.9703	0.19
.Q62	0.9778	0.9776	-0.02
.Q63	0.9954	0.9952	-0.02
.Q64	1	1	0
% Total	0.9622	0.9622	0.01

A.3 Colección NPL

Tabla A.3.1 Mean Average Precision para 92 consultas en NPL

Query ID	AP vectorial	AP propuesto	% Diferencia
.Q01	0.1271	0.1446	13.76
.Q02	0.0328	0.0340	3.77
.Q03	0.0993	0.1464	47.46
.Q04	0.3915	0.4662	19.09
.Q06	0.1244	0.1037	-16.67
.Q07	0.5567	0.5584	0.30
.Q08	0.3333	0.5000	50.02
.Q09	0.2645	0.5152	94.76
.Q10	0.0596	0.0674	13.08
.Q11	0.2911	0.2203	-24.32
.Q12	0.1536	0.1098	-28.49
.Q13	0.1906	0.1826	-4.21
.Q14	0.1837	0.1333	-27.42
.Q15	0.0808	0.0818	1.21
.Q16	0.0497	0.0403	-19.03
.Q17	0.2956	0.3138	6.15
.Q18	0.0551	0.1698	208.47
.Q19	0.1287	0.1317	2.34
.Q20	0.1294	0.1417	9.51
.Q21	0.4017	0.4083	1.66
.Q22	0.1037	0.1181	13.89
.Q23	0.1831	0.2196	19.92
.Q24	0.1693	0.1365	-19.37
.Q25	0.2793	0.1944	-30.40
.Q26	0.4317	0.4506	4.39
.Q27	0.2629	0.2884	9.68
.Q28	0.0894	0.1834	105.29
.Q29	0.1193	0.1469	23.13
.Q30	0.2717	0.3147	15.86
.Q31	0.2612	0.3172	21.42
.Q32	0.5098	0.4944	-3.03
.Q33	0.1430	0.1122	-21.58
.Q34	0.1580	0.1769	11.95
.Q35	0.2476	0.2851	15.16
.Q36	0.0384	0.0368	-4.17
.Q37	0.3905	0.5598	43.37
.Q38	0.4259	0.3777	-11.31
.Q39	0.1279	0.1600	25.11
.Q40	0.3962	0.4013	1.29
.Q41	0.1252	0.1357	8.41
.Q42	0.4099	0.4013	-2.10
.Q43	0.1488	0.1991	33.84

Query ID	AP vectorial	AP propuesto	% Diferencia
.Q44	0.3560	0.3463	-2.73
.Q45	0.3167	0.3633	14.73
.Q46	0.5373	0.5211	-3.02
.Q47	0.1630	0.2735	67.76
.Q48	0.1270	0.1343	5.75
.Q49	0.3016	0.2189	-27.41
.Q50	0.0476	0.0333	-30.04
.Q51	0.2269	0.2628	15.83
.Q52	0.1004	0.1963	95.45
.Q53	0.0479	0.0520	8.41
.Q54	0.3964	0.4163	5.02
.Q55	0.1754	0.1610	-8.22
.Q56	0.1700	0.1882	10.73
.Q57	0.2579	0.1960	-23.98
.Q58	0.0272	0.0207	-23.94
.Q59	0.0062	0.0084	35.48
.Q60	0.1314	0.1537	16.98
.Q61	0.1302	0.1306	0.34
.Q62	0.2369	0.3146	32.80
.Q63	0.4850	0.4573	-5.72
.Q64	0.0444	0.0558	25.63
.Q65	0.1607	0.1231	-23.38
.Q66	0.0214	0.0171	-20.14
.Q67	0.0488	0.0463	-5.12
.Q68	0.0749	0.0612	-18.30
.Q69	0.1986	0.3829	92.79
.Q70	0.0811	0.0940	15.85
.Q71	0.0576	0.0614	6.73
.Q72	0.2412	0.2358	-2.24
.Q73	0.2200	0.2699	22.68
.Q74	0.1313	0.1376	4.86
.Q75	0.3129	0.3199	2.24
.Q76	0.0763	0.1265	65.84
.Q77	0.4202	0.3872	-7.84
.Q78	0.0778	0.0859	10.41
.Q79	0.0472	0.0453	-4.00
.Q80	0.0098	0.0106	8.37
.Q81	0.0789	0.0695	-11.94
.Q82	0.2400	0.2319	-3.36
.Q83	0.1200	0.1298	8.18
.Q84	0.2780	0.3133	12.70
.Q85	0.0061	0.0056	-9.07
.Q86	0.0910	0.0704	-22.67
.Q87	0.1653	0.2489	50.61
.Q88	0.0336	0.0309	-8.05
.Q89	0.0876	0.0953	8.80
.Q90	0.0584	0.1056	80.77
.Q91	0.2123	0.2207	3.91
.Q92	0.1401	0.0739	-47.26
.Q93	0.1785	0.1609	-9.81
MAP	0.1891	0.2048	11.13

Tabla A.3.2 Precisión normalizada para 92 consultas en NLP

Query ID	NPREC vectorial	NPREC propuesto	% Diferencia
.Q01	0.6694	0.6561	-1.99
.Q02	0.5724	0.5574	-2.62
.Q03	0.6558	0.6877	4.86
.Q04	0.8821	0.8795	-0.29
.Q06	0.7291	0.7205	-1.18
.Q07	1	1	0
.Q08	0.9412	0.9629	2.31

Query ID	NPREC vectorial	NPREC propuesto	% Diferencia
.Q09	0.8451	0.8721	3.19
.Q10	0.6006	0.6046	0.67
.Q11	0.8748	0.8374	-4.28
.Q12	0.6965	0.655	-5.96
.Q13	0.7098	0.7044	-0.76
.Q14	0.7104	0.6759	-4.86
.Q15	0.6237	0.6199	-0.61
.Q16	0.5636	0.5443	-3.42
.Q17	0.7803	0.7776	-0.35
.Q18	0.6288	0.6708	6.68
.Q19	0.6744	0.6769	0.37
.Q20	0.704	0.68	-3.41
.Q21	0.829	0.8244	-0.55
.Q22	0.6061	0.6186	2.06
.Q23	0.7286	0.7482	2.69
.Q24	0.719	0.6916	-3.81
.Q25	0.7507	0.7072	-5.79
.Q26	0.8183	0.8173	-0.12
.Q27	0.764	0.7618	-0.29
.Q28	0.7182	0.7307	1.74
.Q29	0.7081	0.7332	3.54
.Q30	0.7996	0.818	2.3
.Q31	0.8101	0.8349	3.06
.Q32	0.8792	0.865	-1.62
.Q33	0.6931	0.6523	-5.89
.Q34	0.6775	0.7128	5.21
.Q35	0.7563	0.7623	0.79
.Q36	0.6383	0.6353	-0.47
.Q37	0.8505	0.8809	3.57
.Q38	0.8632	0.8501	-1.52
.Q39	0.7827	0.7977	1.92
.Q40	0.8487	0.8494	0.08
.Q41	0.6602	0.6695	1.41
.Q42	0.831	0.8287	-0.28
.Q43	0.7491	0.7709	2.91
.Q44	0.8255	0.8266	0.13
.Q45	0.815	0.8164	0.17
.Q46	0.8468	0.8439	-0.34
.Q47	0.7273	0.7705	5.94
.Q48	0.7578	0.773	2.01
.Q49	0.7571	0.7449	-1.61
.Q50	0.8371	0.818	-2.28
.Q51	0.7507	0.7479	-0.37
.Q52	0.6758	0.735	8.76
.Q53	0.5599	0.5596	-0.05
.Q54	0.7725	0.775	0.32
.Q55	0.7332	0.7149	-2.5
.Q56	0.6874	0.6927	0.77
.Q57	0.7483	0.7531	0.64
.Q58	0.5724	0.5503	-3.86
.Q59	0.7281	0.7443	2.22
.Q60	0.7728	0.8133	5.24
.Q61	0.6222	0.6154	-1.09
.Q62	0.7741	0.7978	3.06
.Q63	0.8677	0.8587	-1.04
.Q64	0.5854	0.608	3.86
.Q65	0.8021	0.7747	-3.42
.Q66	0.5767	0.555	-3.76
.Q67	0.5546	0.549	-1.01
.Q68	0.585	0.5615	-4.02
.Q69	0.7398	0.7899	6.77
.Q70	0.738	0.7451	0.96
.Q71	0.6143	0.6215	1.17

Query ID	NPREC vectorial	NPREC propuesto	% Diferencia
.Q72	0.7591	0.7599	0.11
.Q73	0.7182	0.75	4.43
.Q74	0.687	0.6906	0.52
.Q75	0.7979	0.7988	0.11
.Q76	0.6741	0.7088	5.15
.Q77	0.854	0.8398	-1.66
.Q78	0.6499	0.6625	1.94
.Q79	0.5525	0.5416	-1.97
.Q80	0.4809	0.4912	2.14
.Q81	0.7332	0.7303	-0.4
.Q82	0.725	0.7082	-2.32
.Q83	0.6958	0.707	1.61
.Q84	0.7659	0.769	0.4
.Q85	0.4571	0.4453	-2.58
.Q86	0.662	0.6363	-3.88
.Q87	0.7637	0.7735	1.28
.Q88	0.5803	0.5729	-1.28
.Q89	0.7108	0.7159	0.72
.Q90	0.5995	0.6353	5.97
.Q91	0.7647	0.7673	0.34
.Q92	0.6203	0.5884	-5.14
.Q93	0.7137	0.6975	-2.27
% Total	0.7211	0.7227	0.21

Tabla A.3.2 Recuerdo normalizado para 92 consultas en NLP

Query ID	NREC vectorial	NREC propuesto	% Diferencia
.Q01	0.9752	0.97	-0.53
.Q02	0.9744	0.9689	-0.56
.Q03	0.9758	0.9765	0.07
.Q04	0.9982	0.997	-0.12
.Q06	0.9924	0.9925	0.01
.Q07	0.9914	0.9913	-0.01
.Q08	0.9998	0.9999	0.01
.Q09	0.997	0.9972	0.02
.Q10	0.9645	0.9615	-0.31
.Q11	0.999	0.9982	-0.08
.Q12	0.9746	0.971	-0.37
.Q13	0.9726	0.9712	-0.14
.Q14	0.9779	0.9754	-0.26
.Q15	0.9662	0.9635	-0.28
.Q16	0.9598	0.9584	-0.15
.Q17	0.9857	0.9828	-0.29
.Q18	0.98	0.9778	-0.22
.Q19	0.9664	0.9645	-0.2
.Q20	0.9848	0.9793	-0.56
.Q21	0.9902	0.9882	-0.2
.Q22	0.9426	0.9434	0.08
.Q23	0.9827	0.9849	0.22
.Q24	0.9811	0.9791	-0.2
.Q25	0.9729	0.9696	-0.34
.Q26	0.9811	0.98	-0.11
.Q27	0.9868	0.9857	-0.11
.Q28	0.991	0.991	0
.Q29	0.9898	0.9913	0.15
.Q30	0.9951	0.9957	0.06
.Q31	0.9818	0.9861	0.44
.Q32	0.9965	0.9955	-0.1
.Q33	0.9828	0.9737	-0.93
.Q34	0.9854	0.9876	0.22
.Q35	0.9773	0.9763	-0.1
.Q36	0.9835	0.9838	0.03
.Q37	0.993	0.992	-0.1

Query ID	NREC vectorial	NREC propuesto	% Diferencia
.Q38	0.9956	0.9953	-0.03
.Q39	0.9958	0.9957	-0.01
.Q40	0.9945	0.994	-0.05
.Q41	0.9722	0.9725	0.03
.Q42	0.989	0.9883	-0.07
.Q43	0.9931	0.9936	0.05
.Q44	0.9943	0.995	0.07
.Q45	0.985	0.9867	0.17
.Q46	0.9833	0.9839	0.06
.Q47	0.9856	0.9842	-0.14
.Q48	0.9934	0.9941	0.07
.Q49	0.968	0.9701	0.22
.Q50	0.9982	0.9975	-0.07
.Q51	0.9856	0.9824	-0.32
.Q52	0.9807	0.9822	0.15
.Q53	0.9404	0.9392	-0.13
.Q54	0.9813	0.9815	0.02
.Q55	0.989	0.9885	-0.05
.Q56	0.9621	0.9625	0.04
.Q57	0.9908	0.9909	0.01
.Q58	0.9718	0.9695	-0.24
.Q59	0.986	0.9897	0.38
.Q60	0.9952	0.9976	0.24
.Q61	0.9478	0.9449	-0.31
.Q62	0.9896	0.9898	0.02
.Q63	0.995	0.9949	-0.01
.Q64	0.9761	0.9766	0.05
.Q65	0.9967	0.9955	-0.12
.Q66	0.9802	0.9747	-0.56
.Q67	0.9495	0.9484	-0.12
.Q68	0.952	0.9448	-0.76
.Q69	0.9758	0.9743	-0.15
.Q70	0.994	0.9932	-0.08
.Q71	0.9761	0.976	-0.01
.Q72	0.9843	0.9849	0.06
.Q73	0.9824	0.9834	0.1
.Q74	0.9787	0.9786	-0.01
.Q75	0.9881	0.9876	-0.05
.Q76	0.9824	0.9832	0.08
.Q77	0.9946	0.9935	-0.11
.Q78	0.9832	0.9845	0.13
.Q79	0.9505	0.9423	-0.86
.Q80	0.9514	0.9552	0.4
.Q81	0.9936	0.9944	0.08
.Q82	0.9799	0.9755	-0.45
.Q83	0.9851	0.985	-0.01
.Q84	0.9861	0.9846	-0.15
.Q85	0.9328	0.9189	-1.49
.Q86	0.9825	0.9796	-0.3
.Q87	0.9936	0.9926	-0.1
.Q88	0.9679	0.9681	0.02
.Q89	0.9927	0.991	-0.17
.Q90	0.9704	0.9689	-0.15
.Q91	0.9905	0.99	-0.05
.Q92	0.9641	0.9573	-0.71
.Q93	0.9783	0.9751	-0.33
% Total	0.9807	0.9796	-0.12