BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

Facultad de Ciencias de la Computación

"Una Nueva Ponderación para el Modelo de Espacio Vectorial de Recuperación de Información"

Tesis Profesional:
Que para obtener el título de
Maestro en Ciencias de la Computación

Presenta
Rubí Jannet Cabrera Ramírez

Asesores:
Dr. Darnes Vilariño Ayala
M.C. David Eduardo Pinto Avendaño

Puebla, Pue. Otoño 2005

Gracias a mis padres Javier Cabrera Villa y Esla Ramírez Pérez, que con su dedicación en cada instante de mi vida como los grandes seres son, los amo porque con su invaluable amor, han logrado provocar en mi el impulso de ser cada día mejor profesionista y mejor persona, se los agradezco de corazón por hacer posible uno de mis grandes sueños. Los amo demasiado.

Agradezco a mis asesores, les doy gracias por haber confiado en mi y haber colaborado en mi tesis. A usted Dr. Darnes Vilariño Ayala, le doy las gracias por entenderme las diferentes circunstacias de la vida y por siempre tener una palabra de aliento en los momentos dificiles. A usted M.C David Eduardo Pinto Avendao le agradezco de manera especial la enorme paciencia al explicarme y lo principal, el darnos la confianza de que nosotros podemos realizar lo que nos propongamos.

Agradezco a mi hermano Fernando por siempre haber creído que esta etapa de mi vida valdria la pena.

Agradezco a mi hermana Elsa Cabrera por brindarme siempre una sonrisa cuyo signifaco es, sigue hermana sigue tu lo puedes lograr.

Agradezco a mi hermana Roxana Cabrera
por ayudarme en los momentos difíciles a
lo largo de este proyecto de vida.

Agradezco a Marcos de la Rosa Arana por estar siempre a mi lado no importando lo que pase, siendo una persona extrahordinaria. Te amo.

Agradezco a Jesús por demostrarme que yo puedo hacer todo lo que me propoga, además de ser un maravilloso amigo.

Agradezco a Gwendolyne, Rocío, Claudia
y Lorena por su amistad, ya que es un
gran aliciente en el camino de la vida.

Agradezco a la Dra. María Elena Franco Carcedo por haber brindado su enorme experiencia en éste proyecto de tesis.

Gracias a la Vicerrectoria de Estudios de Posgrado por la beca otorgada para la culminación de mis estudios de posgrado.

Índice general

Índice de Tablas

Índice de figuras

Introducción

Desde años atrás ha sido necesario trabajar con información, y con el paso de los años se ha vuelto de mayor importancia la elaboración de técnicas que ayuden a manejar los grandes volúmenes de información existentes, especialmente en la actualidad; Internet. La recuperación de información (RI) es un área dedicada al tratamiento de textos con la finalidad de recuperar aquellos que cumplan con ciertos criterios, de acuerdo a una determinada consulta. Esta área apoya en el proceso de tratamiento de textos; la representación de la información se ha convertido en una línea de trabajo dentro de la RI la cual requiere especial atención. A través de la historia, se han creado diversos mecanismos de representación para textos, que posibilitan la creación de consultas eficazmente. El modelo vectorial propuesto por Salton [1] es un ejemplo claro de la manera en que un texto puede ser representado; en este caso se crea un vector de valores, donde cada valor tiene un peso asociado. Así, el proceso de RI se reduce a operaciones matemáticas entre vectores, para decidir qué documentos son relevantes para una determinada consulta. En este trabajo se proponen dos técnicas matemáticas para representar documentos de texto. En el capítulo 1 se presenta el marco teórico que introduce al lector en el área de RI, en el capítulo 2 se muestra la idea sustancial del trabajo, que recae en el concepto denominado Punto de Transición, y en el planteamiento de las técnicas propuestas; en el capítulo 3 se realiza el análisis de las pruebas y resultados obtenidos sobre diferentes *corpora* y por último se muestran las conclusiones y perspectivas.

Capítulo 1

Marco Teórico

1.1. Antecedentes

Continua referencia se hace actualmente a la importancia que reviste en el mundo real la recuperación de información (RI), tal como en [10], donde se alude a ésta como una operación en la que se interpreta una necesidad de información de un usuario y se seleccionan los documentos más relevantes capaces de solucionarla, básicamente, este proceso consiste en buscar documentos que exhiban un mayor grado de similitud con una pregunta formulada.

La recuperación de información ha generado incluso mayor expectativa, ya que en la actualidad se ha observado que los grandes volúmenes de información en Internet son prácticamente inmanejables bajo los paradigmas clásicos de tratamiento de documentos. Así, surge la necesidad de crear métodos adecuados para la representación y recuperación de la información.

En [5] se refiere a la progresiva y notoria proliferación de herramientas para buscar información en la Web; se estima que en la actualidad existen más de 2000 motores de búsqueda diferentes en la Web, mientras que en 1995 había tan sólo una docena. Cada uno de ellos tiene sus propias características, utilidades e interfaces de usuario.

La creación de los buscadores de información en Internet impulsaron la generación de nuevas técnicas para la representación de información y por consiguiente la recuperación de la misma. La forma más común de encontrar información en Internet es, justamente, utilizando los llamados motores de búsqueda o buscadores. Algunos de los más populares son: Google, Yahoo, Altavista, Excite, InfoSeek, Web Crawler, entre otros. Existen buscadores en casi todos los idiomas del mundo y algunos de ellos poseen ciertas especialidades. Otros se limitan a páginas en algún idioma o referentes a un cierto país o región. También han sufrido grandes mejoras gracias a las diversas investigaciones acerca de la evolución de la información como lo es Google, que optimizó sus técnicas de indización automática para lograr y brindar un excelente funcionamiento [3] [4]. Ciertamente, resulta necesaria la creación de nuevas técnicas y mejoras a las ya existentes para proporcionar información de una forma eficaz ante una búsqueda dada, sobre todo en Español. Así lo que se propone en este trabajo es diseñar una técnica novedosa para la indización automática de documentos no estructurados, basada en el uso de una técnica llamada punto de transición.

En la generación y mejora de las técnicas de indización, existen artículos de gran interés como en [13], donde se plantea una aproximación teórica a la indización automática, poniendo de relieve el papel que ha desempeñado el lenguaje natural, no controlado, en su evolución, y señalando sobre todo las últimas tendencias en indización automatizada fundamentadas en bases de conocimiento. Aunque, sin duda, el marco más importante en el aspecto de representación de información, ha sido el de Gerard Salton [1], que introdujo el modelo de espacio vectorial.

Actualmente, se sigue realizando investigación en esta área y se espera que los resultados arrojados mejoren los valores de precisión y evocación existentes en los sistemas de recuperación de información. En la Facultad de Ciencias de la Computación, BUAP, se está desarrollando un conjunto de herramientas destinadas al proceso de RI. Una

técnica particular que se encuentra en investigación en este grupo es precisamente la del punto de transición. Un trabajo derivado del estudio de esta técnica se puede ver en [12], donde se presenta un mecanismo para reducir los términos de representación de un documento mediante el PT.

Seguramente, esta técnica podrá ser utilizada en mayor medida, ya que sus cualidades proporcionan una plataforma de experimentación, no solamente para la representación de la información, sino incluso para otras áreas ligadas al procesamiento del lenguaje natural (PLN), como es el caso de categorización de textos, generación de extractos, etc.

1.2. Recuperación de Información

Puede decirse, que a la RI la compone un método de acceso a la información que consiste en revisar con la vista un espacio con el propósito de reconocer objetos en él. Puede realizarse en espacios de una dimensión de forma secuencial (por ejemplo en un vector), o puede tener lugar en un contexto estructurado que contiene relaciones jerárquicas (por ejemplo en forma de árbol) o bien semánticas o asociativas (por ejemplo en forma de mapa). Además, puede realizarse tanto en contextos analógicos (por ejemplo sería el caso de revisar las estanterías de una biblioteca) como en contextos digitales (podría ser en el seno de documentos de un procesador de texto, en presentaciones multimedia, en la web, etc).

En la actualidad, la recuperación de la información a partir de una colección de documentos cobra día con día gran importancia en el mundo real, así los documentos recuperados pueden satisfacer una necesidad de información de un usuario expresada normalmente en lenguaje natural. Desde un punto de vista diferente, se considera como un proceso automatizable compuesto por representación, almacenamiento, organización

y acceso a la información. En los últimos 20 años, la RI ha tomado gran auge en la indexación de textos y búsqueda efectiva de los documentos, y con el paso de los años se estudia la manera de modelar, clasificar y categorizar documentos, arquitectura de sistemas, interfaz de usuario, visualización de datos, filtrado, lenguajes etc.

Para la indexación de los documentos y facilitar el acceso a ellos, se suelen usar términos o palabras clave, los cuales se consideran en una extracción automática como las palabras más representativas a un documento dado, sin olvidar factores importantes como lo son la estructura del documento, es decir, si son documentos sin estructura (texto libre), estructura fija (estructura externa al documento), metadatos (estructura interna al documento, por ejemplo, XML). Otro factor importante es la caracterización de la pregunta, la cual consiste en cómo el usuario realiza una pregunta que le devuelve un conjunto de documentos o el usuario selecciona los documentos más relevantes de los recuperados en función de esos documentos.

Por lo anterior, un Sistema de Recuperación de Información (SRI) debe soportar una serie de operaciones básicas sobre los documentos como los son la eliminación, modificación, e inserción de documentos. Se debe contar también con un método de localización de documentos para que sean presentados al usuario. Los SRI manejan dichas operaciones de forma diversa, por lo que regularmente es posible encontrar variaciones con respecto a los métodos de búsqueda y las técnicas de representación de documentos. Sin embargo, en los SRI se suele utilizar el enfoque de frecuencias de ocurrencias de los términos, se utiliza también la morfología de los términos, e incluso un enfoque no semántico. Después de haber representado los documentos es posible, aplicar la función de similitud, cuya finalidad es saber qué tan similar es la representación de los documentos con la representación de la consulta. En un SRI compiten la precisión y evocación de tal manera que a mayor precisión menor evocación y viceversa.

Se distinguen principalmente dos divisiones importantes dentro de dichos modelos

para realizar la tarea de recuperación, que son: los modelos clásicos y los modelos estructurados. A los modelos clásicos pertenecen tres modelos importantes que son la base de muchos otros: modelo booleano, vectorial y probabilístico. Estos modelos han sido el fundamento para el desarrollo de otros nuevos que han permitido mejoras en su eficiencia y precisión. H. Lee, en [7], enuncia, analiza y compara las diferentes extensiones que ha tenido el modelo booleano como son: el modelo basado en la teoría de conjuntos difusos, el modelo Waller-Kraft, el modelo de Paice, el booleano extendido y el Infinite-One.

Como extensiones del modelo vectorial es posible ver los modelos de vector generalizado, indexación semántica latente (LSI), redes neuronales y algoritmos genéticos. Y por último, en la generalización del modelo probabilístico se tienen las redes de inferencia y las redes de creencia. Tanto los modelos clásicos como sus extensiones permiten realizar búsquedas en documentos cuyas vistas lógicas van de términos indexados a texto completo; sin embargo, los modelos estructurados, como su nombre lo dice, permiten la recuperación únicamente en documentos que cuentan con una estructura. La figura 1.1 muestra esquemáticamente estas divisiones.

Figura 1.1: Modelos clásicos.

1.3. Modelos de Representación de Documentos

1.3.1. Modelo Booleano

El modelo booleano permite representar y recuperar documentos mediante funciones de similaridad booleanas. Así, dada una consulta, el modelo regresa solamente aquellos documentos que empatan totalmente con la consulta. En este caso, cada documento es evaluado mediante una función booleana, y de ahí el nombre del modelo. Éste no permite ordenar los resultados por relevancia; es decir:

- Los documentos se representan mediante vectores binarios
- Las consultas son combinaciones de términos índice conectados por los operadores lógicos AND, OR y NOT
- Los documentos relevantes son los que satisfacen totalmente la consulta, el resto son irrelevantes (no hay grados de relevancia como en el modelo de espacio vectorial, el cual discutremos mas adelante).

La descripción formal del modelo booleano, incluyendo preproceso, representación y consulta, es dada a continuación:

Preproceso Dado un texto D, denótese con D' el que se obtiene por eliminar las palabras cerradas (preposiciones, artículos, etc.) y lematizando cada una de las restantes.

Definición de Índices Sea $D = D_1, ..., D_k$ una colección de documentos y D' la colección preprocesada. Sea $V = \bigcup_i D'$ el vocabulario de la colección, y $V_0 = [v_i]_i$ el vocabulario ordenado lexicográficamente. La representación de un texto D es el vector $\vec{D} = [d_i]_{i \leq n}$ donde:

$$\mathrm{d}_i = \begin{cases} 1, & \text{si } V_i \in \text{D'}; \\ 0, & \text{si } V_i \notin \text{D'}, \text{ con n} = \#V_0. \end{cases}$$

Búsqueda Dada una consulta q formada por términos preprocesados, $q_1, ..., q_k$, se forma el vector $\vec{q} = [q_i]_{i \leq k}$. Los documentos recuperados bajo el modelo booleano son D_j tales que $\vec{D}_j \cdot \vec{q} = 0$.

1.3.2. Modelo de Espacio Vectorial

El modelo de espacio vectorial fue propuesto por Salton en la década de los 70's [1]. La idea básica de este modelo reside en la construcción de una matriz llamada también vector de términos y documentos, donde las filas fueran documentos y las columnas correspondieran a los términos incluidos en ellos. Así, las filas de esta matriz serían equivalentes a los documentos que se expresarían en función de las apariciones (frecuencia) de cada término. De esta manera, un documento podría expresarse de la manera $d_1 = (1, 2, 0, 0, 0, ..., 1, 3)$ siendo cada uno de estos valores el número de veces que aparece cada término en el documento. La longitud del vector de documentos sería igual al total de términos de la matriz.

De esta manera, un conjunto de m documentos se almacenaría en una matriz de m filas por n columnas, siendo n el total de términos almacenamos en ese conjunto de documentos. La segunda idea asociada a este modelo es calcular la similitud entre la pregunta (que se convertiría en el vector pregunta, expresado en función de la aparición de los n términos en la expresión de búsqueda) y los m vectores de documentos almacenados. Los más similares serían aquellos que deberían colocarse en los primeros lugares de la respuesta.

La frecuencia de los términos en un documento parece ser un buen indicador de la importancia del mismo, por lo que muchos autores lo han usado, por otro lado,

algunos autores, entre los que destaca Sparck-Jones [14], han apreciado la capacidad de discriminación de un término frente a otro, dando importancia o generalidad a un término dentro de la colección, en función del conjunto,y no de un único documento; así, se ha pensado en incentivar la presencia de aquellos términos que aparecen en menos documentos frente a los que aparecen en todos o casi todos, ya que realmente los muy frecuentes discriminan poco o nada en el momento de la representación del contenido de un documento. Para medir este valor de discriminación se propone la medida tf (frecuencia de aparición de un término en el documento) e idf (frecuencia inversa del documento), respectivamente, los cuales presentan un buen comportamiento, aunque complican el tiempo de cálculo de la representación.

El vector para cada documento tiene k elementos, siendo k igual al número de términos indizables que existen en la colección documental *(corpus)*. Los componentes en el vector se fijan con los pesos calculados para cada término en la colección de documentos. A los términos en cada documento se les asignan pesos automáticamente, basándose en la frecuencia con que ocurren en la colección entera de documentos y en la aparición de un término en un documento particular. Así, el peso de un término en un documento aumenta si éste aparece más a menudo en un documento y disminuye si aparece más a menudo en todos los demás documentos. El peso para un término en un vector de documento es distinto de cero sólo si el término aparece en el documento. Para una colección de documentos grande que consiste en numerosos documentos pequeños, es probable que los vectores de los documentos contengan ceros principalmente.

La descripción formal del modelo de espacio vectorial, incluyendo preprocesamiento, definición de índices, ambos previamente definidos en la sección 1.3.1, por último representación y consulta, es dada a continuación:

Figura 1.2: formación de vectores.

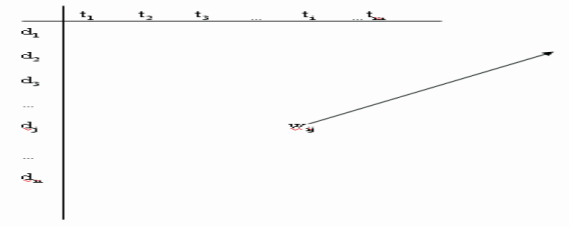

Figura 1.3: Representación del vector de pesos.

Asignación de pesos Así, para la construcción de la matriz de términos y documentos, se consideran las siguientes definiciones:

Los componentes de cada vector $\vec{D} = [d_{i1}, ..., d_{in}]$ son ponderados de la siguiente forma: $d_{ik} = tf_{ik} \cdot idf_k$ donde tf_{ik} [2] es la frecuencia del término k en D_i, e idf_k está definido como $idf_k = log_2(M) - log_2(df_k) + 1$, siendo df_k el número de documentos que usan el término k, y M el número de documentos en la colección.

Similitud En el caso de la representación vectorial se emplea el coseno del ángulo entre los vectores que representan a los documentos:

$$sim(D_i, q) = \frac{\sum_{k=1}^{m} d_{ik} q_k}{\sqrt{\sum_{k=1}^{m} d_{ik}^2} \cdot \sum_{k=1}^{m} q_k^2} \tag{1.1}$$

El resultado del cálculo anterior mide la semejanza entre la consulta y cada uno de los documentos, de manera que, aquéllos que, en teoría, se ajustan más a la consulta formulada, producen un índice más alto de similitud.

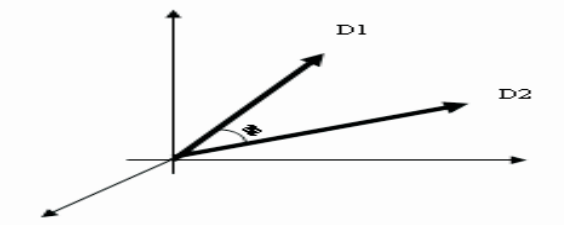

Figura 1.4: Similitud de vectores.

Formalmente es común considerar la función de similitud normalizada ($\in [0, 1]$). El índice de Jaccard $sim(D, q) = \#(D \cap q)/\#(D \cup q)$ se emplea en la representacin booleana. Para el modelo booleano también suele emplearse otra medida $sim(D, q) = 2 \times \#(D_i \cap q)/\#D_i + \#q$, que es llamada el coeficiente de Dice.

1.3.3. Modelo Probabilístico

El modelo probabilístico [15] trabaja con los conceptos que proceden del área de la probabilidad y de la estadística. En este modelo, los términos de la indización de documentos y de las consultas no poseen el cálculo de pesos. El orden de los documentos se calcula con el peso dinámico de los términos de la consulta relativamente a los documentos. Se basa en el principio de la ordenación probabilística (Probability Ranking Principle). Este modelo busca saber la probabilidad de un documento D relevante para una consulta Q_a. Tal información puede ser obtenida asumiendo que la distribución de términos no sea capaz de informar la relevancia para un documento cualquiera de la colección de documentos.

Principio de Ordenación Probabilística:

- Sea $+R_a$ un documento relevante para la consulta Q_a
- Sea R_a un documento no relevante para la consulta Q_a

- Sea $P(+R_a/D)$ la probabilidad de que un documento D sea relevante para la consulta Q_a
- Sea $P(-R_a/D)$ la probabilidad de que un documento D no sea relevante para la consulta Q_a

Si se asume que la relevancia de un documento es independiente de la relevancia de todos los demás (esto no es verdad), un documento D es relevante a la consulta Q_a cuando: $P(+R_a/D)\ P(-R_a/D)$.

Así mismo, dada una consulta Q_a, o modelo probabilístico con atributos de cada documento D (como medida de simililitud) un peso WD/Q_a, como:

$$W_{D/Q_a} = \frac{P(+R_a/D)}{P(-R_a/D)} \tag{1.2}$$

La fórmula anterior calcula a probabilidad de que D puede ser tanto relevante como irrelevante. La teoría de Bayes ayuda a identificar para cada término de la consulta, el grado de relevancia o de irrelevancia de un documento, seleccionando el más adecuado (el que produzca menor error) para la sumatoria final de probabilidades de relevancia, lo cual esta dado por la sumatoria de grados de relevancia de cada término. Así, aplicando la regla de Bayes.

$$W_{D/Q_a} = \frac{P(D/+R_a) \times P(+R_a)}{P(D/-R_a) \times P(-R_a)} \tag{1.3}$$

Donde:

- $P(D/+R_a)$ es la probabilidad de que el documento relevante para la consulta Q_a, sea D
- $P(D/-R_a)$ es la probabilidad de que el documento no relevante para Q_a, sea D

- $P(+R_a)$ es la probabilidad de que el documento sea relevante.
- $P(D/-R_a)$ es la probabilidad que un documento no sea relevante.

Para calcular $P(D/+R_a)$ y $P(D/-R_a)$, como los términos indexados de los documentos, el documento se puede representar por el vector: $D = \{x_1, x_2, ..., x_n\}$, x_k $\{0, 1\}$. Colocando esto en en la siguiente fórmula:

$$P(D/+R_a) = \prod_{k=1}^{n} P(X_k/+R_a) \tag{1.4}$$

Donde:

- $P(X_k/+R_a)$ es la probabilidad de que un documento D sea relevante para la consulta Q_a; si el evento decrito X_k presencia o ausencia del término$k$ en el documento D ocurre.
- $r_{ak} = P(X_k = 1/+R_a)$ es la probabilidad de que dado un documento D, sea relevante para la consulta Q_a, y el término X_k esté presente en D. Entonces la fórmula se reescribe como:

$$P(D/+R_a) = \prod_{k=1}^{n} r_{ak}^{xk}(1-r_{ak})^{1-xk} \tag{1.5}$$

Análogamente se puede derivar una expresión similar para $P(D/-R_a)$, siguiendo los mismos pasos, donde:

- $r_{ak} = P(X_k = 1/-R_a)$ es la probabilidad de que dado documento D, esto no es factible a la consulta Q_a, el término k está presente en D.

se concluye que:

$$W_{D/Q_a} = \sum_{k=1}^{n} X_k \times W_{ak} + C \tag{1.6}$$

$$X_k \in 0, 1$$

$$W_a k = \log \frac{r_a k}{1 - r_a k} + \log \frac{1 - S_a k}{S_a k} \tag{1.7}$$

$$C = \log \frac{P(+R_a)^{14}}{P(-R - a} + \sum_{k=1}^{n} \log \frac{1 - r_a k}{1 - S_a k} \tag{1.8}$$

De esta forma se evaluará los pesos para los términos de la consulta (W_{ak}), que también están presentes en los documentos ($X_k = 1$) y C (constante la cual es la misma), pero puede interpretada como el valor de corte para la función de recuperación. Por esta razón, la ecuación final se puede escribir simplemente de la forma:

$$sim(D, Q_a) = W_{D/Q_a} = \sum_{k=1}^{n} X_k \times W_{ak} \tag{1.9}$$

W_{D/Q_a} es la medida de similitud entre la consulta Q_a y el documento D. Note que el W_{ak} es el peso para el término k de la consulta, mientras que el X_k es el peso para el término k en el documento.

1.4. Criterios de Evaluación de un SRI

Un sistema de recuperación de información (SRI) es una terna $S =< D, Rep, E >$, donde D es una colección de documentos, $Rep : D \rightarrow 2^v$ una representación y $E : V^+ \rightarrow 2^D$ una función de búsqueda.

Consulta Se asume que la consulta se formula en lenguaje natural, y, de lo dicho se deduce que el resultado de la consulta consiste en una vector de documentos ordenados en orden decreciente en función de su similitud con la consulta. Una consulta supervisada Qes una pareja donde su primer componente es una expresión y el segundo un subconjunto de $D : (q, r) \rightarrow V^+ \times 2^D$

Precisión y Evocación De forma tradicional se ha conferido mucha importancia a la efectividad de la recuperación, normalmente basada en la relevancia de los documentos, lo cual ha representado un problema, ya que medir la relevancia es un proceso subjetivo y sin confianza. Esto es, diferentes juicios personales asignarían diferentes valores de relevancia a un documento recuperado en respuesta a la búsqueda hecha. La seriedad del problema es la materia de debate, bastantes investigadores señalan que la subjetividad del juicio sobre la relevancia no es suficiente para invalidar el sistema. Muchas medidas de la efectividad de la recuperación han sido propuestas. Las más empleadas, de forma general, son las conocidas como *evocación y precisión* [16]

La evocación son los documentos relevantes recuperados en una búsqueda dada, sobre el número de documentos relevantes para esa búsqueda en la base de datos. Excepto para pequeñas colecciones, este denominador es generalmente desconocido y debe ser estimado por muestreo o por otros métodos. Precisión es el número de documentos relevantes recuperados, sobre el número total de documentos recuperados. El rango de valores de ambos, está comprendido entre 0 y 1.

Formalizando, para un SRI S y una consulta supervisada $Q = (q, r)$, se definen la precisión $P = \#(E(q) \cap r)/\#E(q)$, y la evocación $R = \#(E(q) \cap r)/\#r$. Es común promediar estos valores para un conjunto de consultas supervisadas [2].

Coordinación, evocación y precisión Coordinación. Así como se ha definido el niv-

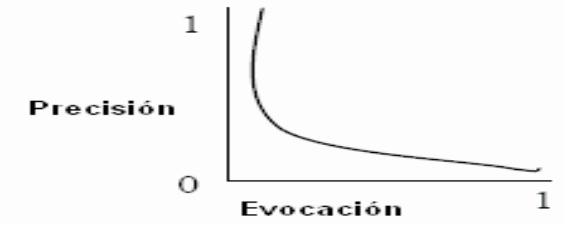

Figura 1.5: Precisión y evocación.

el de coordinación para los términos que participan en una respuesta, en la evaluación consideramos el ranking de los documentos, i.e. los documentos ordenados descendentemente, de acuerdo al valor de similitud con la respuesta.

Gradación (ranking) de una respuesta Sea $S =< D, Rep, E >$ un SRI, *(q, r)* una consulta supervisada con $\#r = nr$ y supóngase que la gradación de E *(q)* es $[D_{qi}]_{i \leq nq}$ con $\#E(q) = nq$. Considérese el primer elemento de este vector, D_{qj} , que ocurre en r, y referido como el nivel de evocación 1.

Evaluación por niveles Para el nivel de evocación 1 $(100 \times (1/nr)\,\%)$, la precisión y evocación son $P = 1/j$ y $R = 1/nr$. Para el nivel de evocación $2(100 \times (2/nr)\,\%)$ $P = 2/j$ y $R = 2/nr$, donde j es la posición en $[D_{qi}]_i$ del segundo documento relevante $(\in r), etc.$

A continuación se muestra un ejemplo de la evaluación de precisión y evocación global [17], por niveles de evocación estándar y F1.

Supongamos la información presentada en la tabla 1.1:

En base a las ecuaciones 1.10 y 1.11, se evalúa la precisión y evocación global.

$$precisión = \frac{\#dedocuemntosrelevantesrecuperados}{\#númerodedocumentosrecuperados} \tag{1.10}$$

Tabla 1.1: Información para ejemplificar precisión y evocación

Rango de documento	relevante	Precisión	Evocación
1		0	0
2	★	1/2	1/3
3	★	2/3	2/3
4		2/4	2/3
5		2/5	2/3

$$evocación = \frac{\#dedocumentosrelevantesrecuperados}{\#dedocuemntosrelevantes} \tag{1.11}$$

En la misma tabla 1.1 se ilustran los cálculos de precisión y evocación para los primeros 5 documentos recuperados, para una consulta en particular;en este caso,la colección contiene 3 documentos relevantes. El primer documento recuperado no es juzgado relevante, así que la precisión y la evocación de este documento son nulos. El segundo documento recuperado por el sistema es relevante, por lo tanto la precisión para los dos primeros documentos es entonces 1 documento relevante recuperado dividido entre 2 documentos recuperados; y la evocación es 1 documento relevante recuperado, dividido entre 3 documentos de la colección que han sido juzgados relevantes.

La evaluación de los niveles de evocación estándar sirve para facilitar la comparación de los modelos propuestos, sobretodo cuando hay varios, un valor simple que representa la eficacia del sistema es útil. La precisión media sobre 11 niveles de evocación estandar (0,0, . . . , 1,0) o 10 niveles de evocación (0,1 . . . , 1,0) de evocación son dos métricas estándares. En nuestro trabajo, utilizamos la precisión media sobre 11 niveles de evoación para evaluar todos los experimentos, es decir, es la evaluación sobre el total de los documentos relevantes recuperados por el modelo propuesto entre los once niveles de evocación, el resultado obtenido es el número de documentos de los cuales se calcula el promedio de los valores de similaridad de cada documento y se obtiene para cada

nivel estandar.

Para el cálculo de F1 se aplica la fórmula 1.12, donde p corresponde a la precisión obtenida y e corresponde a la evocación obtenida en el cálculo anterior. Así el valor de de precisión global es:1/2, la evocación es: 1/3 y fianlmente F1=.39997

$$F1 = \frac{2(p * e)}{p + e} \tag{1.12}$$

Capítulo 2

Ponderación de términos

2.1. Introducción

Un modelo de representación tiene como objetivo satisfacer las necesidades reales y potenciales de información de todos los usuarios, proporcionándoles la información veraz pertinente, justo a tiempo y al menor coste. En particular, contempla una serie de etapas con las que debe cumplir para ser considerado un modelo de representación de RI óptimo [11]:

1. Obtener la representación de los documentos. Generalmente los documentos se presentan utilizando un conjunto más o menos grande de términos índice.

2. Identificar la necesidad informativa del usuario. Se trata de obtener la representación de esa necesidad, y plasmarla formalmente en una consulta acorde con el sistema de recuperación.

3. Búscar los documentos que satisfagan la consulta. Consiste en comparar las representaciones de documentos y la representación de la necesidad informativa para seleccionar los documentos pertinentes.

4. Obtener los resultados y presentarlos al usuario.

5. Evaluar los resultados por parte del usuario.

Una técnica particular que se encuentra en investigación en el grupo de RI de la FCC-BUAP, es precisamente la del punto de transición. Un trabajo derivado del estudio de esta técnica se puede ver en [12], en donde se presenta un mecanismo para reducir los términos de representación de un documento mediante el PT. En este artículo se hace uso de la técnica PT para la obtención de un nuevo mecanismo de ponderación comparándolo con el peso clásico del modelo de espacio vectorial propuesto por Salton [2].

2.2. Planteamiento del problema

En este trabajo de tesis se propone usar un modelo matemático para representar documentos de texto. La idea sustancial del trabajo recae en el concepto denominado Punto de Transición [12]. En este caso, la representación del documento puede realizarse por un número pequeño de términos pertenecientes al vocabulario. El objetivo principal de este trabajo es la definición del modelo de ponderación. Una aproximación inicial podría ser el de establecer de un peso sobre los términos alrededor del punto de transición, de acuerdo a la distancia de los mismos hacia el PT. Este modelo aporta grandes ventajas, ya que permite seguir utilizando conceptos matemáticos en el proceso de RI. Una manera de validar el modelo de representación sería a través de su comparación con otro modelo. Inicialmente se plantea comparar el nuevo modelo con el modelo de espacio vectorial [11] sobre un conjunto de documentos evaluados *a priori*, por ejemplo, el TREC [6]. El TREC es una colección de aproximadamente 250,000 noticias que posee consultas supervisadas. Existen alrededor de 50 preguntas y las respuestas están indicadas en el mismo *corpus*. Se espera establecer un nuevo modelo de ponderación que

permita competir con los niveles de precisión y evocación de los modelos tradicionales, como es el caso del modelo de espacio vectorial.

2.3. Modelos Propuestos

Una forma de representar los documentos es por medio del cálculo de pesos en donde se asigna un valor numérico a cada término del documento, por lo que se propone establecer dos técnicas de cálculo de pesos, con la finalidad de obtener una comparación entre dichas técnicas y así mostrar que pueden ser comparables con el modelo de espacio vectorial, obteniendo como ventaja en reducción de espacio de términos y el trabajo de búsqueda, como se muestra en la tabla .

2.3.1. Uso del Punto de Transición

El Punto de Transición (PT) refiere básicamente a un valor de frecuencia que corresponde a un término en el vocabulario del texto que divide al mismo vocabulario en términos de alta y baja frecuencia. Urbizagástegui [8], por ejemplo, se refiere a este concepto a través de la ley de Zipf [9], y presenta un ejercicio en donde argumenta el hecho de que, existe una vecindad de términos alrededor del punto de transición que describen de manera general el contenido del mismo texto. Este concepto es sumamente importante, ya que dichos términos podrían utilizarse para representar el documento. La fórmula para la obtención del valor de frecuencia del PT se muestra en la ecuación (2.1).

$$PT = \frac{\sqrt{1 + 8 * l_1} - 1}{2}, \tag{2.1}$$

donde l_1 es el número de términos con frecuencia igual a uno.

Después de diversos experimentos, se observa que el PT dado por la fórmula 2.1, puede ser aplicado sólo para colecciones de documentos extensos, ya que si es aplicado en documentos pequeños, su vocabulario reducido, hace que el valor de PT quede fuera de cualquier valor de este vocabulario, por lo que se aplica una técnica denominada cálculo por inspección la cual consiste de calcular, por inspección, la frecuencia más baja, de las altas, que no se repita.

A continuación, se abordará un ejemplo para el cálculo del PT. Dado un texto no preprocesado (ver tabla 2.1), se somete a una serie de pasos como lo es el preproceso (eliminación de las palabras cerradas), el cual genera como resultado el texto que se observa en la tabla 2.2, posteriormente se obtienen las frecuencias de los términos, que a su vez forman el vocabulario de dicho texto (ver tabla 2.3). El valor de frecuencia obtenido para el punto de transición, usando la técnica de inspección, es 3 mientras que el valor de PT usando la ecuación 2.1 es de 10.18.

Esto demuestra la utilidad de la técnica de inspección, y dado que la mayoría de los documentos a utilizar en las pruebas son pequeños, usaremos esta técnica. Por otro lado si se deseara obtener un conjunto de términos alrededor de PT sería más fácil hacerlo. La tabla 2.4 muestra una vecindad del 25 % de PT para este mismo ejemplo.

2.3.2. Modelo $IDPT_{ij} \cdot DPTC_i$

Esta técnica se aplica sobre los términos alrededor del punto de transición que se encuentran en una vecindad del 25 %, de acuerdo a la distancia de los mismos hacia el PT. Lo anterior, se muestra evaluado por las fórmulas mostradas en las ecuaciones (2.2), (2.3), (2.4).

El objetivo es obtener términos altamente representativos de cada documento y de la colección. Así, W_{ij} es el peso que le corresponde al término i en el documento j.

Tabla 2.1: Noticia No. SP94-0000006

VERACRUZ.- Más de 70 taxistas se manifestaron hoy en las principales avenidas de esta ciudad para demandar al Gobierno mayor seguridad. Armando Rodríguez, quien encabezó la manifestación, señaló que se exigirá se encuentre al responsable del homicidio del seor Porfirio Aguilar, taxista asesinado recientemente. Informó que el número de asaltos se incrementó a cinco por semana, la mayoría de ellos con violencia. El líder de los taxistas destacó que los secuestros de unidades del transporte público y asesinatos de choferes se ha incrementado en forma alarmante y las autoridades no han hecho nada al respecto. Los vehículos de los taxistas escribieron en todos los cristales de sus unidades denuncias de ilícitos cometidos en su contra y demandas de solución a su problema. Esta es la segunda manifestación masiva de choferes, luego del asesinato de Porfirio Aguilar. Descriptores: Columna-México-Hoy Columna

Tabla 2.2: Documento preprocesado

SP94-0000006 veracruz taxistas manifestaron principales avenidas ciudad demandar gobierno mayor seguridad armando rodríguez encabez manifestación sealó exigirá encuentre responsable homicidio porfirio aguilar taxista asesinado recientemente informó número asaltos incrementó cinco semana mayoría violencia líder taxistas destacó secuestros unidades transporte público asesinatos choferes incrementado forma alarmante autoridades hecho respecto vehículos taxis escribieron cristales unidades denuncias ilícitos cometidos demandas solución problema esta segunda manifestación masiva choferes asesinato porfirio aguilar descriptores columna méxico columna

Tabla 2.3: Vocabulario de SP94-0000006

palabras	Frecuencias
aguilar	2
alarmante	1
armando	1
asaltos	1
asesinado	1
asesinato	1
asesinatos	1
autoridades	1
avenidas	1
choferes	2
cinco	1
ciudad	1
columna	2
cometidos	1
cristales	1
demandar	1
. . .	. . .

Tabla 2.4: Vocabulario representativo de SP94-0000006

Frecuencias	palabras
3	taxistas
2	columna
2	aguilar
2	porfirio
2	manifestación
2	choferes
2	unidades

$$W_{ij} = IDPT_{ij} * DPTC_i \tag{2.2}$$

$$IDPT_{ij} = \frac{1}{\left| PT_j - F(t_{ij}) \right|^2} \tag{2.3}$$

$$DPTC_i = \left| (PT - F(t_i)) \right| \tag{2.4}$$

Donde $IDPT_{ij}$ es la distancia inversa del término i al punto de transición, PT_j, del documento j. Se eleva al cuadrado el denominador para asignar un valor de potencia cuadrática a los términos en función de su cercanía al PT_j.

Y $DPTC_i$ es la distancia de la frecuencia de aparición del término i del vocabulario del *corpus* al punto de transición PT evaluado sobre el *corpus* completo.

2.3.3. Modelo $IDPT_{ij} \cdot idf_k$

La segunda técnica del cálculo de pesos está basada también en el PT por documento y se presenta en la ecuación 2.5.

$$W = IDPT_{ij} \cdot idf_k \tag{2.5}$$

Donde $IDPT_{ij}$ es la distancia inversa del término al punto de transición PT del documento (como se menciona en el punto 2.3.2). e

$$idf_k = \log_2(\frac{2 \times M}{df_k}) \tag{2.6}$$

Siendo df_k el número de documentos que usan el término k y M el número de documentos en la colección usada. Como se puede observar, el cálculo de idf_k es idéntico al presentado por Salton en su modelo de espacio vectorial.

2.4. Representación de la consulta

Se requiere un apartado especial para discutir la representación de la consulta. Así, a continuación se discute este tema para ambos modelos planteados. Los usuarios que consultan el sistema de recuperación para buscar información deben traducir su necesidad informativa en una consulta adecuada al sistema de recuperación. Esto supone utilizar un conjunto de términos que expresen semánticamente su necesidad. En sistemas tradicionales también es habitual utilizar operadores booleanos para conectar varios criterios de búsqueda por campos diferentes.

2.4.1. Representación de la consulta para el modelo $IDPT_{ij} \cdot DPTC_i$

Se asigna un peso a cada término de la consulta usando el cálculo definido en la sección 2.3.2, a excepción de que $IDPT_{ij}$ se asume con un valor de 1, debido principalmente a la cantidad reducida de términos usualmente contenidos en una consulta.

2.4.2. Representación de la consulta para el Modelo $IDPT_{ij} \cdot idf_k$

Se asigna un peso a cada término de la consulta usando el cálculo de la sección 2.3.3. correspondiente a idf_k, a excepción de que $IDPT_{ij}$ se asume con un valor de 1, al igual que el cálculo de la subsección 2.4.1.

Capítulo 3

Pruebas

3.1. *Corpus* de prueba (TREC)

TREC *(Text Retrieval Conference)* constituye uno de los esfuerzos más significativos de investigación experimental en recuperación de información (RI). El patrocinio de elaboración de estas conferencias se encuentra bajo la *National Institute of Standards and Technology* (NIST) y de la *Defense Advanced Research Projects Agency* (DARPA). Dichas conferencias comenzaron en 1992 (TREC-1), y vienen celebrándose con periodicidad anual hasta la fecha. De manera particular en 1996, se celebró TREC-5, en cual se utilizó una colección de aproximadamente 250,000 noticias en español. El TREC-5 posee alrededor de 50 consultas supervisadas y las respuestas son indicadas en el mismo *corpus* [6]. La idea es poder establecer comparaciones fiables entre los distintos sistemas empleados por los investigadores en TREC-5, dado que todos operan con las mismas colecciones y las mismas consultas, y presentan sus resultados en la misma forma; obviamente, utilizan sistemas y técnicas diferentes. En base a lo anterior se creó *corpora* basado en el TREC-5 para la realización de las pruebas, comprendido de cinco *corpora*(como se muestra en la tabla 3.1), todos éstos con un número de noticias correspondientes al Diario el Norte de Guadalajara, las cuales constan de noticias

relevantes y no relevantes (como se muestra en la tabla 3.2), también en la tabla 3.3 se muestra la dimensión del vector de representación para cada *corpus*, en la cual se observa que las técnicas tienen dimensiones iguales debido a que ambas manejan el cálculo de inspección del PT:

Tabla 3.1: *corpora* utilizados

Corpus	Consultas
corpus 1	cosulta10 y consulta11
corpus 2	consulta1 y consulta3
corpus 3	consulta14 y consulta15
corpus 4	consulta4 y consulta5
corpus 5	consulta24 y consulta25

Tabla 3.2: Detalle de los *corpora*

Corpus	Total de noticias	Total de noticias relevantes	Total de noticias no relevantes	Tamaño *corpus*
corpus 1	933	206 y 105	727 y 828	2.65 MB
corpus 2	1117	211 y 164	906 y 953	3.35 MB
corpus 3	816	281 y 7	535 y 809	2.20 MB
corpus 4	448	97 y 257	351 y 191	1.26 MB
corpus 5	1239	131 y 359	1108 y 880	3.87 MB

Tabla 3.3: Número de términos usados en la representación.

Corpus	$\text{IDPT}_{ij} \cdot idf_k$	$\text{IDPT}_{ij} \cdot DPTC_i$	MEV
1	1789	1789	35,144
2	2281	2281	41,878
3	1827	1827	33,566
4	1714	1714	40,651
5	2508	2508	45,715

Los resultados obtenidos en los experimentos realizados sobre estos *corpora* se muestran en la siguiente sección.

3.2. Pruebas y resultados

Una vez que se han definido las dos técnicas en la sección 2.3.2 y 2.3.3, se realiza una serie de pruebas con los diferentes *corporas*, calculando la precisión y evocación global, precisión y evocación por niveles, y el cálculo de F1.

En particular al *corpus* 1 y en función con la consulta 10 y 11 (que se muestra en la tabla 3.4), se obtienen los resultados mostrados en la tabla 3.5, 3.6 y en las gráficas 3.1 y 3.2, en las cuales se observa que el modelo de espacio vectorial supera a las dos técnicas propuestas debido a su alto nivel de evocación, sin embargo en la consulta 10, la técnica de $IDPT_{ij} \cdot DPTC_i$ supera en nivel de precisión a ambas; en cuanto a la consulta 11, la técnica $IDPT_{ij} \cdot idf_k$ muestra una alta precisión en los datos con respecto a las otras dos técnicas.

Tabla 3.4: Consultas 10 y 11

Consulta 10 :México es importante país de tránsito en la guerra antinarcótica.
Consulta 11 :Derechos a las aguas de los ríos en la región fronteriza entre México y los Estados Unidos.

Tabla 3.5: Consulta 10

	IDPT$_{ij} \cdot idf_k$	IDPT$_{ij} \cdot DPTC_i$	MEV
Evocación	.2378	.0145	**.8786**
Precisión	.0045	**.6**	.2350
F1	.0088	.0283	**.3708**

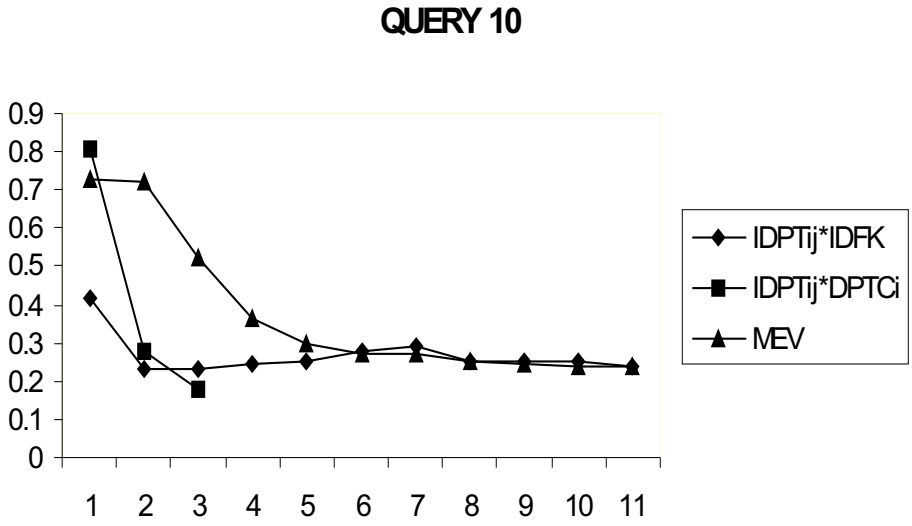

Figura 3.1: Precisión por niveles de evocación estándar para la consulta 10.

Tabla 3.6: Consulta 11

	$\text{IDPT}_{ij} \cdot idf_k$	$\text{IDPT}_{ij} \cdot DPTC_i$	MEV
Evocación	.3428	.0476	**.9809**
Precisión	**.2553**	.1612	.1473
F1	**.2926**	.0734	.2561

Para el *corpus* 2, en función con las consultas 1 (ver tabla3.7), se observa que el modelo de espacio vectorial supera a las otras técnicas debido a su alto nivel de evocación, sin embargo, para ambas técnicas el nivel de evocación es muy bajo, por lo que la técnica de $IDPT_{ij} \cdot DPTC_i$ no arrojó documentos relevantes para ninguna consulta, cabe mencionar que también uno de los motivos sea la inexistencia de pesos para los términos de dichas consultas. Por otro lado, en la consulta 3 (ver tabla3.7) se observa que la técnica $IDPT_{ij} \cdot idf_k$ supera en niveles de precisión y evocación al modelo de espacio vectorial. Estos resultados se muestran en las tablas 3.8, 3.9 y gráficas 3.3 y 3.4.

Tabla 3.7: Consultas 1 y 3

Consulta 1 :Oposición Mexicana al TLC.
Consulta 3 :Polución en el Distrito Federal de México

Tabla 3.8: Consulta 1

	IDPT$_{ij} \cdot idf_k$	IDPT$_{ij} \cdot DPTC_i$	MEV
Evocación	.0142	- -	.7203
Precisión	.15	- -	.3568
F1	.0259	- -	.4772

Tabla 3.9: Consulta 3

	IDPT$_{ij} \cdot idf_k$	IDPT$_{ij} \cdot DPTC_i$	MEV
Evocación	.0975	- -	.6097
Precisión	.5161	- -	.3184
F1	.1640	- -	.4183

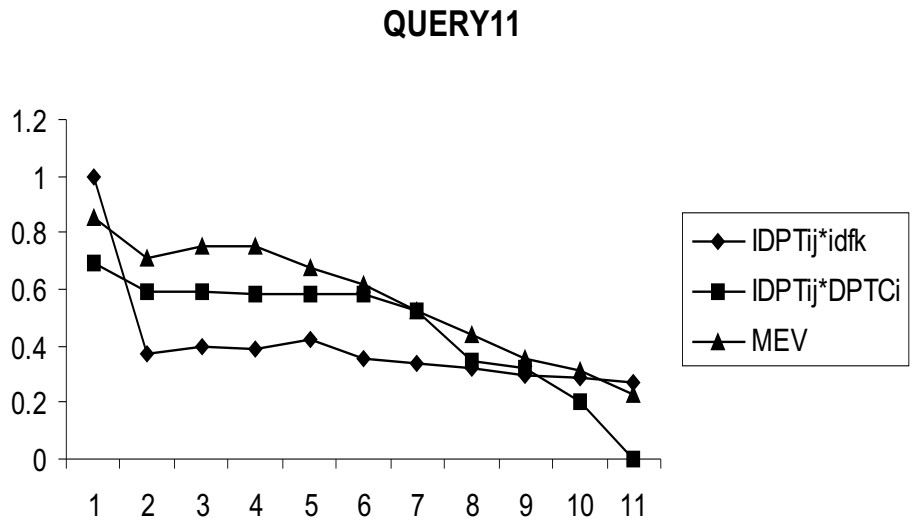

Figura 3.2: Precisión por niveles de evocación estándar para la consulta 11.

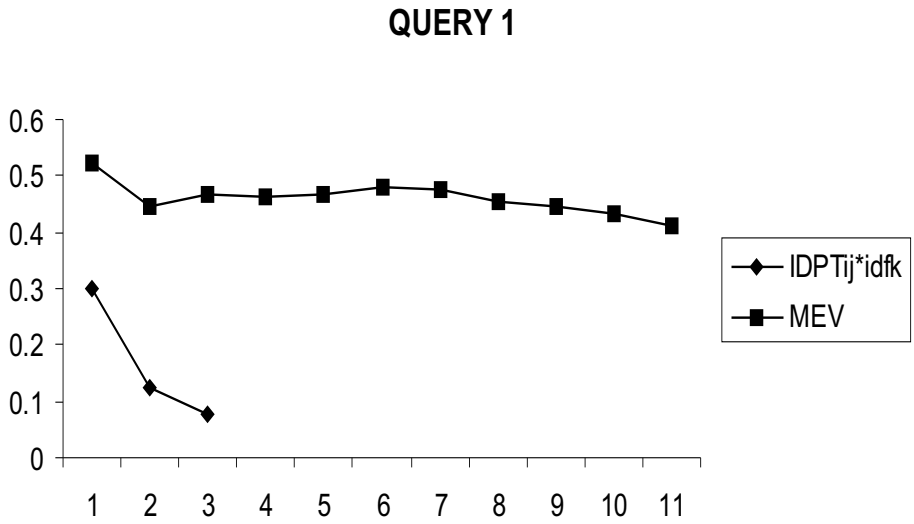

Figura 3.3: Precisión por niveles de evocación estándar para la consulta 1.

Para el *corpus* 3, en función con la consulta 14 (ver tabla 3.10), se observa que la precisión de la técnica $IDPT_{ij} \cdot DPTC_i$ es menor que la de $IDPT_{ij} \cdot idf_k$, pero la primera técnica antes mencionada la supera en niveles de evocación, ya que es menor, sin embargo, el modelo de espacio vectorial muestra altos niveles de precisión y evocación, pero no supera a ambas técnicas (ver tabla 3.11 y la gráfica 3.5). Por otro lado, en la consulta 15 (ver tabla 3.10) se observa que los niveles de evocación y precisión son muy bajos, por lo que la técnica de $IDPT_{ij} \cdot DPTC_i$, no obtuvo resultados; con respecto al modelo de espacio vectorial, tiene un alto índice de evocación y un nivel muy bajo de precisión, i.e. solo arrojó una cantidad pequeña de documentos relevantes.(ver tabla 3.12 y la gráfica 3.6)

Tabla 3.10: Consultas 14 y 15

Consulta 14 :El monopolio petrolero PEMEX tiene mucha influencia en México.
Consulta 15 :La disputa sobre la pesca ha ocasionado la captura de barcos de pesca de los Estados Unidos.

Tabla 3.11: Consulta 14

	IDPT$_{ij} \cdot idf_k$	IDPT$_{ij} \cdot DPTC_i$	MEV
Evocación	.4661	.0284	**.9644**
Precisión	**.8506**	.7272	.6791
F1	.6022	.0546	**.7969**

Tabla 3.12: Consulta 15

	IDPT$_{ij} \cdot idf_k$	IDPT$_{ij} \cdot DPTC_i$	MEV
Evocación	.1428	- -	**.8571**
Precisión	**.0204**	- -	.0175
F1	**.0357**	- -	.0342

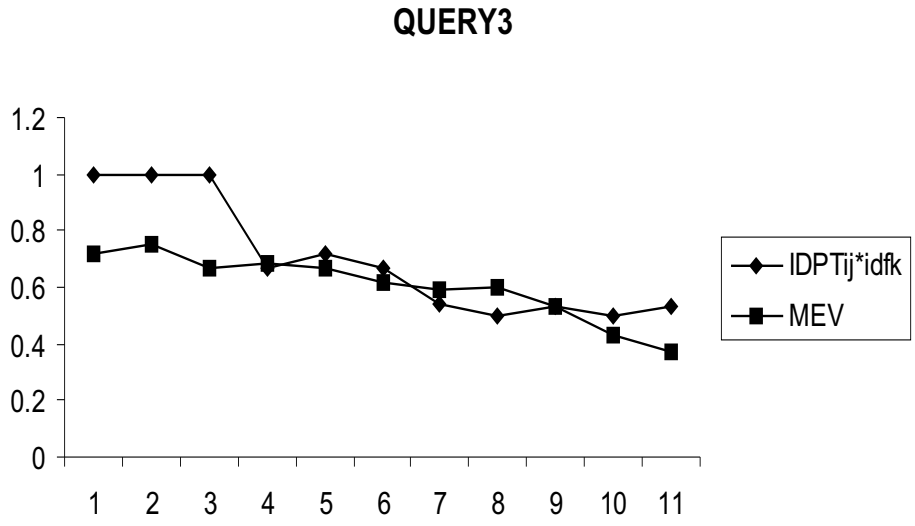

Figura 3.4: Precisión por niveles de evocación estándar para la consulta 3.

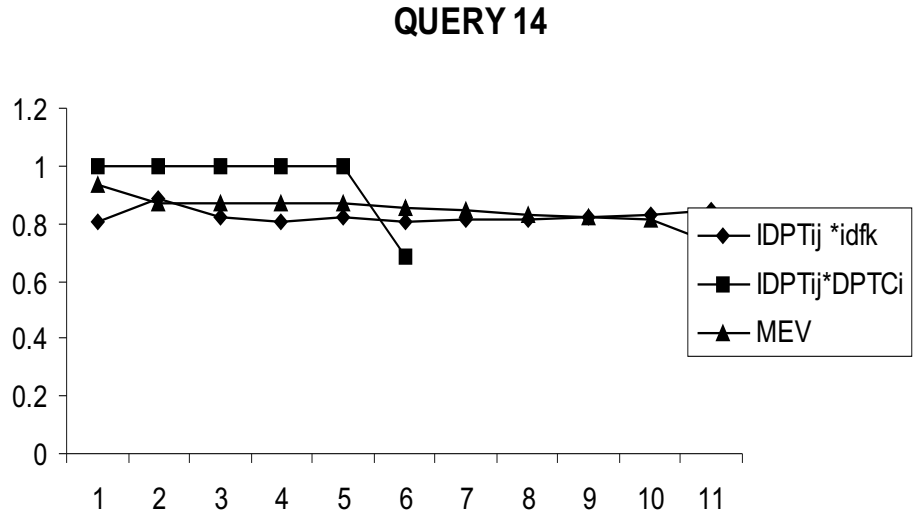

Figura 3.5: Precisión por niveles de evocación estándar para la consulta 14.

Para el *corpus* 4, en función con la consulta 5 (ver tabla 3.13), se observa que el modelo de espacio vectorial supera a ambas técnicas debido a su alto nivel de evocación, sin embargo, puede ser comparable con la técnica de $IDPT_{ij} \cdot DPTC_i$, por su nivel de precisión alto. Por otro lado la consulta 4 no arrojó ningun tipo de dato con respecto a las tres técnicas. (ver tabla 3.14 y la gráfica 3.7)

Tabla 3.13: Consultas 4 y 5

Consulta 4 :Papel de México en la OEA.
Consulta 5 :Maquiladoras en la economía mexicana

Tabla 3.14: Consulta 5

	IDPT$_{ij} \cdot idf_k$	IDPT$_{ij} \cdot DPTC_i$	MEV
Evocación	.0077	.1958	**.8599**
Precisión	.1333	**.6333**	.5815
F1	.0145	.2991	**.6938**

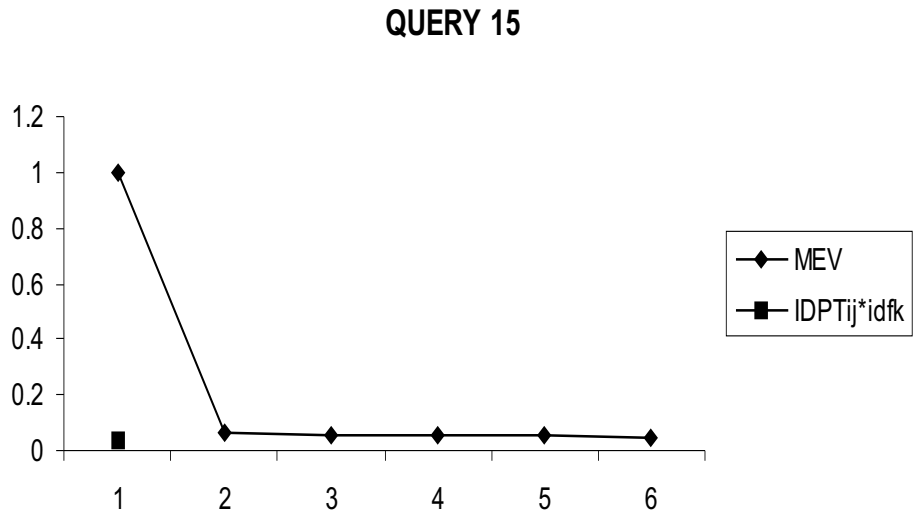

Figura 3.6: Precisión por niveles de evocación estándar para la consulta 15.

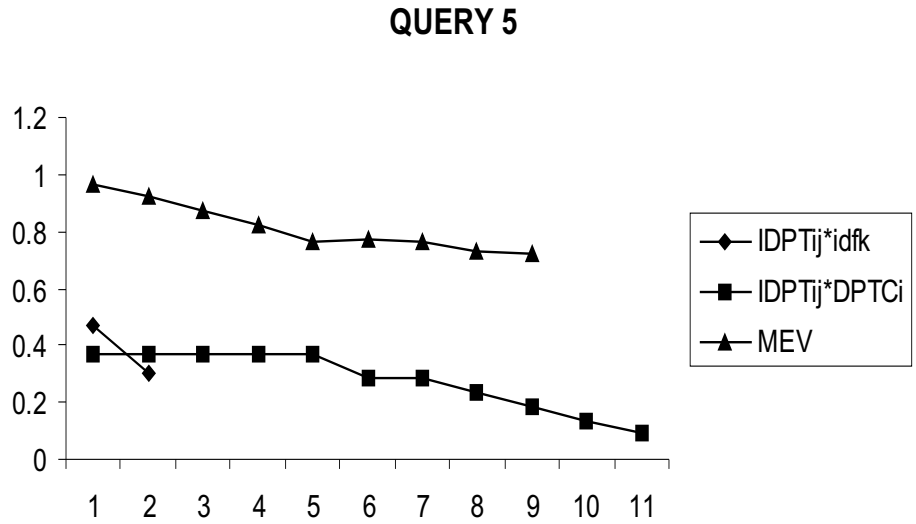

Figura 3.7: Precisión por niveles de evocación estándar para la consulta 5.

Y por último, se analiza el *corpus* 5, con respecto a la consulta 25, que fue la única que proporciona datos con bajo nivel de precisión en ambas técnicas a excepción del modelo de espacio vectorial. (ver tabla 3.15, 3.16 y gráficas 3.8).

Tabla 3.15: Consultas 24 y 25

Consulta 24 :Prevención de SIDA en México. Consulta 25 :Programa de Privatización de Empresas Públicas Mexicanas.

Tabla 3.16: Consulta 25

	$\text{IDPT}_{ij} \cdot idf_k$	$\text{IDPT}_{ij} \cdot DPTC_i$	MEV
Evocación	.0083	.0055	**.9693**
Precisión	.1071	.4	**.5378**
F1	.0154	.0098	**.6917**

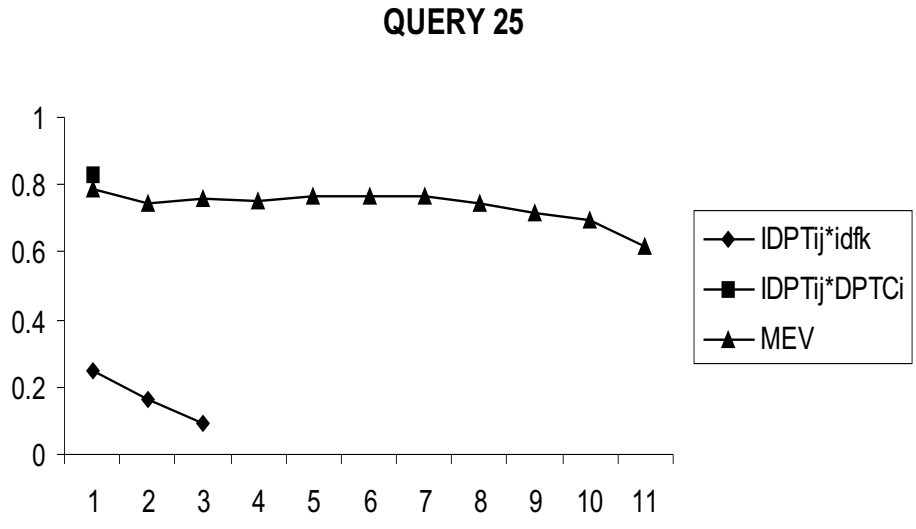

Figura 3.8: Precisión por niveles de evocación estándar para la consulta 25.

Con el objetivo de extender los resultados obtenidos para cualquier consulta y sobre cualquier *corpus* se realizó una prueba estadística de análisis de varianza (ANOVA); utilizando el paquete STATISTICA.

Se tomaron como individuos las consultas realizadas y como variable, una variable cualitativa (técnicas), con tres niveles; el nivel 1 $IDPT_{ij} \cdot idf_k$, el nivel 2 $IDPT_{ij} \cdot DPTC_i$ y el nivel 3 el MEV. Las variables respuesta son evocación, precisión y F1.

La prueba de hipótesis planteada para las tres características fue la siguiente:

$$H_0\text{: } \mu 1 = \mu 2 = \mu 3$$

$$H_1\text{: Algunas son diferentes}$$

Lo que se desea probar es que la media de las tres técnicas sean iguales contra que son diferentes para cada una de las características.

Tabla 3.17: Tabla de ANOVA, Evocación

Causas de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	P
Efecto	2	17,623	8,817	.177	0.8387
Error	21	1,043.27	49,679		

Tomando un nivel de significación de $\alpha = 0{,}05$ se obtuvo que $P > \alpha$, es decir no se rechaza H_0; lo que implica que las medias de las tres técnicas en la población son iguales.

Tabla 3.18: Tabla de ANOVA, Precisión.

Causas de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	P
Efecto	2	115.583	57.792	1.2231	0.3144
Error	21	992.25	47.250		

De igual manera tomando como nivel de significación a $\alpha = 0{,}05$, se obtuvo que $P > \alpha$ por lo que no se rechaza H_0, nuevamente la media de las tres técnicas en la población son iguales.

Tabla 3.19: Tabla de ANOVA, F1.

Causas de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	P
Efecto	2	9,003.25	4,501.625	4.0177	0.033
Error	21	23,529.25	1,120.44		

Tomando $\alpha = 0{,}05$ se obtuvo que $P < \alpha$, por lo que se rechaza la hipótesis de nulidad, esto quiere decir que la media de las tres técnicas son diferentes con respecto a la característica F1.

Analizando la media de las tres técnicas con respecto a F1 se obtuvo lo siguiente:

Tabla 3.20: Analisis de la media de las tres propuestas.

Técnicas	{1}	{2}	{3}
	M=109.13	M=68.875	M=110.75
idf_k {1}		0.25480	0.923573
$DPTC_i$ {2}	0.25480	0	0.20688
MEV {3}	0.923563	0.20680	

De los resultados obtenidos se puede concluir que la F1 obtenida utilizando la técnica $IDPT_{ij} \cdot idf_k$ difiere de la F1 obtenida utilizando la técnica $IDPT_{ij} \cdot DPTC_i$. Por otra parte, la F1 que se obtuvo utilizando el modelo MEV difiere significativamente de la F1 obtenida con la técnica $IDPT_{ij} \cdot DPTC_i$; sin embargo no hay diferencia significativa entre la técnica $IDPT_{ij} \cdot idf_k$ y MEV.

Conclusiones

Se definieron dos técnicas para el cálculo de pesos basadas en el Punto de transición. Mediante un estudio cuidadoso se realizaron una serie de pruebas con diferentes *corpora* extraídos del TREC-5, en las cuales se puede observar la comparabilidad que tienen con respecto al modelo de espacio vectorial (MEV). En la técnica $IDPT_{ij} \cdot DPTC_i$ se observó que los niveles de evocación son muy bajos debido a que, en varias consultas, el peso de los términos que las conforman es cero, ya que dicho terminó se encuentra fuera de la vecindad del punto de transición. De las diez consultas realizadas, la técnica $IDPT_{ij} \cdot DPTC_i$ ofreció mayor nivel de precisión solamente en dos; de igual manera se observó variabilidad con respecto a la precisión en las otras dos técnicas. Se realizó un análisis de varianza para determinar cual de las tres técnicas es más precisa, llegándose a la conclusión que no hay diferencia significativa entre las medias de la precisión en cada una de ellas.

Con referencia a la segunda técnica $IDPT_{ij} \cdot idf_k$ se observa que en muchos de los casos los resultados superan a la técnica $IDPT_{ij} \cdot DPTC_i$, debido a que sus niveles de evocación son ligeramente mas altos ya que el tratamiento de los pesos de la consulta son diferentes a cero. De los resultados obtenidos puede observarse que la técnica $IDPT_{ij} \cdot idf_k$ no difiere significativamente del MEV, y es de destacar que almacena menor número de términos en su vector de representación, sin embargo para poder extender resultados a cualquier *corpora* y usando cualquier consulta se debe realizar mayor número de

experimentos para darle mayor robustez a la misma.

Se han cumplido los objetivos de este trabajo los cuales son, la propuesta e implementación de dos técnicas, ambas brindan una reducción significativa en el tiempo de ejecución, debido al manejo de vectores de representación significativamente pequeos, para los *corpus* analizados se obtuvieron los siguientes resultados (tabla 3.21).

Tabla 3.21: Reducción significativa de términos .

Corpus	No. De términosTotales.	% de almacenamiento en idf_k y $DPTC_i$	% de almacenamiento en MEV
1	35,144	5.09 %	100 %
2	41,878	5.4 %	100 %
3	33,566	5.44 %	100 %
4	40,651	4.21 %	100 %
5	45,715	5.48 %	100 %

Perspectivas

Después de los experimentos realizados en éste trabajo se observó que:

- En la técnica de $IDPT_{ij} \cdot DPTC_i$, el tratamiento de la consulta podría ser mejorado, en base a un análisis del comportamiento de la misma.
- Se propone aplicar las técnicas desarrolladas en este trabajo en otros modelos como, por ejemplo, el modelo booleano o probabilístico, con la finalidad de lograr un refinamiento.
- Se propone realizar un análisis comparativo con otras técnicas de reducción de términos.
- Se propone realizar o estudiar un cálculo de inspección diferente, con la finalidad de ampliar la vecindad de términos significativos de los textos.

Bibliografía

[1] G. Salton,*Automatic Text Processing*,Addison-Wesley,(1989).

[2] Jiménez H. and Pinto D.,*Notas de Academia Recuperación de Información*,Octubre,(2003).

[3] *The Pagerank citation Ranking Bringing Order to WEB*, January,(1998).

[4] Sergey Brin and Lawrence Page,*The Anatomy of a Large-Scale Hypertextual Web Search Engine*,Computer Science Department, Stanford University, Stanford, CA 94305, USA.

[5] T. Stanley",*Meta-Searching on the Web*, http://www.ariadne.ac.uk/issue12/search-engines/,(1998).

[6] http://trec.nist.gov/Text Retrieval Conference (TREC) última revisión 13/sep/04.

[7] Lee, J. H. 1994. P*roperties of extended boolean models in information retrieval*, En memorias del Annual ACM Conference on Research and Development in Information Retrieval, (SIGIR Dubln, Irlanda), 182-190.

[8] Urbizagátegui A. R., *Las implicaciones de la ley de Zipf en la indización automática*,Universidad de California Riverside,(1999).

[9] Zipf George K., *Humand Behavior and the Principle of Least-Effort*, Addison-Wesley, Cambridge MA,(1949).

[10] L. Codina, *Teoría de Recuperación de Información: modelos fundamentales y aplicación a la gestión documental*, Information World en Español, vol. 38, 18-22, Octubre 1995.

[11] Angel F., Rodríguez Zazo, Figuerola G., Alonso J.L. and Gómez R., *Recuperación de Información utilizando el Modelo Vectorial*, Departamento de informática y automática, Universidad de Salamanca, 2002, Mayo.

[12] Moyotl E., Reyes B. and Jiménez H., *Reducción de términos índice usando el Punto de Transición.*

[13] Méndez Rodríguez Eva M and Moreiro González José A. , *Lenguaje natural e Indización automatizada*,Departamento de Biblioteconomía y Documentación,Universidad Carlos III de Madrid (España), 1997.

[14] Sparck Jones Karen , *Information Retrieval Experiment book* published in 1981.

[15] Paes Cardoso Olinda Nogueira, *Recuperación de Información* Universidad federal de lavras depto. de ciencias de la computación.

[16] http://irsweb.blogspot.com/2005/03/el-modelo-del-espacio-vectorial-i.html.

[17] Alvarez Carmen,*Modelos de lenguaje en recuperación de información*,Universidad de Montreal, Febrero 2004.