UNIVERSIDAD AGRARIA DE LA HABANA
"Fructuoso Rodríguez Pérez"

FACULTAD DE CIENCIAS TÉCNICAS
Departamento de Informática

Métodos matemáticos para la recuperación semántica de información.

Tesis presentada en opción al título de Máster en Biomatemática

Autor: Ing. Adanay Nuñez González

Tutor: Dr.C Alexander Sánchez Díaz

Mayabeque, 2016

AGRADECIMIENTOS

A la persona más especial y maravillosa de mi vida, que me ha hecho lo que soy hoy, a mi amiga por sobre todas las cosas. A mi mamá. Este es tu triunfo también. Gracias por guiarme siempre por los buenos caminos de la vida y sobre todo por el amor que me entregas día a día.

A mi papá, por ser siempre el ejemplo intachable a seguir como profesional, recuerda que a tu paloma todavía le falta mucho por volar. Gracias por toda tu dedicación en mi educación y por el inmenso amor que me has entregado.

A mi novio, por su paciencia en mis mayores momentos de estrés, por siempre estar a mi lado en las buenas y en las malas; por su comprensión, paciencia y amor, dándome ánimo, fuerza y valor para seguir adelante. Te amo.

A toda mi familia, porque siempre me ha apoyado en mis decisiones, en especial a mi madrina pues la considero como mi segunda madre.

A mi suegra y mi cuñada, por apoyarme y aconsejarme en todo momento, por su preocupación y ayuda incondicionalmente en los momentos de agobio.

A mi tutor, que aun cuando pensé que esta batalla estaba perdida me supo orientar y motivar ayudándome a reunir las fuerzas y continuar adelante.

Al claustro de profesores que me formó de manera profesional durante el transcurso de estos dos años de maestría y me transmitieron sus conocimientos, experiencias y valores.

A mis amistades, principalmente a mis compañeros del Departamento.

A todos… ¡Muchas Gracias!

-Adanay-

A mi Mamá...

RESUMEN

La recuperación de información es uno de los principales desafíos hoy día para la comunidad científica. Esta temática de amplio interés ha sido objeto de estudio en esta tesis. Se han abordado algunos de los principales métodos de recuperación de información que existen. Fundamentalmente, los que hacen uso de diferentes técnicas matemáticas para la obtención de resultados relevantes durante un proceso automatizado de búsqueda. La Indexación Semántica Latente es una de estas técnicas, la cual tiene en cuenta la semántica implícita en las relaciones entre términos y documentos. Ha sido empleada con el objetivo de disminuir el efecto de la sinonimia y la polisemia en el momento de realizar una búsqueda. Basado en este método, se desarrolló una aplicación informática, en la cual se implementaron algunos esquemas de ponderación de términos propuestos en la literatura consultada. También se realizó un análisis sobre la influencia de las longitudes de los documentos en el momento de efectuar una consulta. Se definieron para la evaluación de los resultados experimentales métricas de precisión y cobertura. La colección de documentos MEDLINE, usada en varios proyectos internacionales, permitió realizar un análisis de los métodos matemáticos usados. Se evaluó la eficiencia de las métricas de ajuste para la creación de la matriz de términos-documentos usando una función de normalización pivoteada. Se logró así, recuperar documentos según sus probabilidades de relevancia obtenidas aplicando los criterios establecidos en la colección.

Palabras claves: recuperación de información, Indexación Semántica Latente, matriz de términos-documentos, normalización pivoteada, similitud coseno.

ÍNDICE

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Actualmente los motores de búsqueda para recuperar información en Internet utilizan matrices para ubicar dónde ésta se encuentra e incluso para determinar la forma en que los sitios web se relacionan entre sí. La eficacia de Google radica fundamentalmente en la forma en que utiliza estas matrices para determinar cuáles sitios están referenciados en otros. La estructura de la matriz de Google determina las páginas web que coinciden con el tema de la búsqueda y posteriormente, los algoritmos presentan una lista de las páginas recuperadas en un orden de relevancia (Gallardo, 2006). Para la Recuperación de Información (RI) existen tres modelos matemáticos clásicos: el modelo booleano, el modelo vectorial y el modelo probabilístico; cada uno de los cuales presenta disímiles extensiones con el objetivo de su optimización (Alimohammadi, 2010).

Al mismo tiempo, debido a la rápida evolución de la web, arribando a la Web 3.0 o Web Semántica, cada vez se hace más necesario tener métodos eficientes de recuperación de información, es decir, encontrar recursos dentro de grandes colecciones de datos para satisfacer la necesidad de un usuario. Para Berners-Lee y Hendler (2001) la Web Semántica es una "web de datos" a diferencia de una "web de documentos". Para ello el consorcio W3C recomienda el uso del modelo conceptual RDF[1] (*Resource Description Framework*) para describir semánticamente el contenido que brinda una página web a través del uso de metadatos. Por otra parte, con el objetivo de lograr una mayor integración y navegabilidad, el modelo de *Linked Data*[2] (datos entrelazados) propone el uso de URIs (*Uniform Resource Identifier*) para identificar cualquier recurso que sea accesible a través de la Web. La Web Semántica es una expansión de la red tradicional, donde los sistemas informáticos "comprenden" o interpretan el significado de la información. Propone superar las limitaciones de la Web actual mediante la introducción de descripciones explícitas del significado, la estructura interna y la estructura global de los contenidos y servicios disponibles en la Web. Todo esto se logra gracias a los metadatos y a la información enlazada.

Los buscadores actuales permiten localizar cualquier tipo de información alojada en la Web mediante mecanismos sintácticos fundamentalmente. Sin embargo, estos sistemas muestran mejores prestaciones para personas que saben exactamente qué tipo de resultado esperan obtener. En estos casos, si se ingresan palabras las cuales su significado viene dado según el contexto en

[1] https://www.w3.org/RDF/
[2] https://www.w3.org/standards/semanticweb/data

que se utilice, se obtendrá como resultado un gran número de sitios y documentos; los cuales en su mayoría no serán capaces de satisfacer las demandas reales del usuario. Además este tipo de búsqueda tampoco contempla palabras que tienen un significado similar o idéntico entre sí y que pertenecen a la misma categoría gramatical (sinonimia). Por eso se plantea que la búsqueda semántica ayuda a la tradicional introduciéndole mejoras, sin que ello implique su sustitución. Un buscador semántico identifica el contexto o la intención del usuario, es decir, lo que realmente desea buscar, siendo capaz de aplicar algoritmos y criterios para recuperar información de acuerdo a su significado.

Existe un método matemático que se apoya en el Álgebra Lineal, propuesto por Dumais et al. (1988), que contempla ambas tendencias nombrado Indexación Semántica Latente. La misma está basada en el modelo matemático vectorial (Alimohammadi, 2010) siendo su idea principal acotar cada documento y consulta en un pequeño espacio dimensional asociado con los conceptos, pretendiendo así modelar matemáticamente relaciones de asociación entre los documentos y sus términos. Esta técnica parte de una matriz inicial que representa esas relaciones, la cual se construye a partir del peso del término en cada documento; es decir, la medida que cuantifica su relación. Para calcular estos pesos existe una variedad de esquemas de ponderación, en los cuales se tiene en cuenta incluso hasta la longitud de los documentos, pues este es un aspecto fundamental por el cual se ve afectado el proceso de RI. El ruido presente en esta matriz es eliminado mediante la reducción del espacio vectorial. Éste está constituido por aquellos documentos que a pesar de contener algún término de la consulta no son considerados relevantes, debido a que su semántica no es afín con lo que se busca.

En Cuba también se han llevado a cabo investigaciones con respecto a la recuperación de información en la Web obteniéndose buenos resultados. Algunas de estas investigaciones se han llevado a cabo en la Universidad de Ciencias Informáticas (UCI). Allí se desarrolló un motor de búsqueda para la web de dicho centro llamado ORION, el cual recientemente se integró en la plataforma digital de Contenidos Unificados para Búsqueda Avanzada (CUBA). Todo esto forma parte de las acciones que se realizan para la informatización de la sociedad cubana. Ambos resultados son buscadores puramente sintácticos.

Por otra parte, la Dirección de Informatización de la Universidad Agraria de La Habana (UNAH) tiene dentro de sus objetivos el desarrollo de aplicaciones que gestionen el proceso de enseñanza-

aprendizaje usando las tecnologías actuales. Unido a ese propósito su grupo PRONEG (Procesos de Negocios) está realizando la implementación de herramientas informáticas que permitan recrear escenarios de enseñanza aprendizaje centrados en el estudiante (Sánchez Díaz et al., 2015). En vista del cumplimiento de estos objetivos, en el grupo se desarrollan investigaciones sobre un tema muy latente en la actualidad en el mundo del *e-learning*: los Entornos Personales de Aprendizaje (PLE, *Personal Learning Environment*). Adell y Castañeda (2010) definen un PLE como "…el conjunto de herramientas, fuentes de información, conexiones y actividades que cada persona utiliza de forma asidua para aprender", y a su vez estos autores, líderes en este tema definen un grupo de funcionalidades que debería poseer estos entornos, dentro de los cuales la búsqueda de información es un punto crucial. Actualmente las bibliotecas digitales constituyen unas de las mayores fuentes de información a disposición de los usuarios, y especialmente en las universidades, donde continuamente se publican una gran cantidad de artículos. Generalmente las búsquedas realizadas son por autores, palabras claves o el propio título de la obra, sin embargo se pueden realizar búsquedas a partir del contenido implícito en el resumen de los documentos.

La búsqueda sintáctica no es totalmente eficiente en estos escenarios porque no manejan situaciones en las cuales, dentro de la consulta, existen términos con diversos significados, devolviendo información y documentos no deseados. De igual forma al realizar una búsqueda de manera sintáctica no se contemplan aquellos documentos donde aparecen los sinónimos de los términos en cuestión, dejando ocultos aquellos que también podrían ser útiles para el usuario. Por el contrario, la búsqueda semántica identifica el significado de la palabra pues trabaja a través de conceptos, no sólo se basa en la sintaxis de la misma. En esos escenarios se aspiran el desarrollo de aplicaciones informáticas capaces de diferenciar el significado de los términos de búsqueda, y al mismo tiempo procesar su contenido, combinarlo y realizar deducciones lógicas, atendiendo a las demandas de información del usuario (Caicedo, 2012).

La herramienta informática en desarrollo maneja perfiles de usuarios a través de la implementación de grafos o redes sociales. En estas redes los estudiantes pueden intercambiar materiales didácticos, dejar criterios o comentarios sobre revisiones bibliográficas, puntuar obras orientadas por los profesores, etc. Todo esto implica que se debe proveer una herramienta de búsqueda de información que garantice los resultados en función de los intereses del usuario.

El rápido crecimiento de las fuentes de datos en la sociedad actual conlleva el perfeccionamiento de las herramientas de búsqueda de información, mediante técnicas que aprovechen los significados de las palabras y las relaciones entre conceptos. Nuestra universidad no está alejada de esa necesidad, teniendo en cuenta el actual modelo de Educación Superior en Cuba y el uso intensivo de las nuevas tecnologías en los procesos formadores. Es importante que los sistemas automatizados que acceden a la información de catálogos públicos de la UNAH, como el de la biblioteca digital o el del Laboratorio de Tecnologías Educativas (LATED), brinden información relevante. También así se favorece el desarrollo de entornos donde los procesos de enseñanza y aprendizaje se personalicen mediante plataformas informáticas. Es bajo estas premisas y en el contexto universitario actual, que se plantea el **problema científico** de esta tesis mediante la siguiente interrogante:

¿Cómo obtener un mayor número de documentos relevantes en un proceso de búsqueda teniéndose en cuenta las relaciones entre sus conceptos?

Para dar solución al mismo se propone la siguiente **hipótesis**:

Una adecuada ponderación de las relaciones entre términos y documentos y el manejo de dimensiones que representen conceptos en un contexto determinado, así como su tratamiento mediante matrices y espacios vectoriales, permitirán refinar los resultados de las búsquedas de manera que se recuperen los documentos más relevantes.

La investigación se realiza sobre el **objeto de estudio**: técnicas de recuperación de información. Se enmarca como **campo de acción** la recuperación de información a partir de la técnica de Indexación Semántica Latente.

Se establece como **objetivo general**: implementar métodos de recuperación de documentos mediante la aplicación de la técnica Indexación Semántica Latente.

Objetivos específicos:

1. Definir funciones de asociación de términos y documentos usando las funciones de peso que evalúan la frecuencia de distribución de términos y la entropía.
2. Determinar la dimensión del subespacio vectorial que define la Indexación Semántica Latente, de manera que se reduzca el ruido de la matriz términos-documentos y no disminuya el poder de consulta.

3. Implementar todas las etapas del proceso de recuperación de información teniendo en cuenta los algoritmos matemáticos implícitos.
4. Evaluar los métodos propuestos a partir de un corpus de documentos de prueba.

El **aporte práctico** de la tesis es el análisis experimental que se realizó apoyado en una aplicación informática de recuperación de información, mediante la cual se evaluó el comportamiento de la cobertura y precisión de los resultados al aplicar la técnica de Indexación Semántica Latente bajo ciertas condiciones. Éstas fueron dadas por el manejo de diferentes esquemas para representar las relaciones entre los términos y los documentos, y diferentes dimensiones del espacio vectorial de búsqueda de manera tal que fuese posible eliminarse el ruido sin disminuir el poder de consulta.

CAPÍTULO 1: LA RECUPERACIÓN DE INFORMACIÓN A TRAVÉS DE MODELOS MATEMÁTICOS.

La recuperación de información permite identificar los documentos relevantes de los que no lo son mediante su jerarquización en el momento de realizar una búsqueda. La figura 1 sintetiza cómo se realiza este proceso el cual es guiado por un modelo.

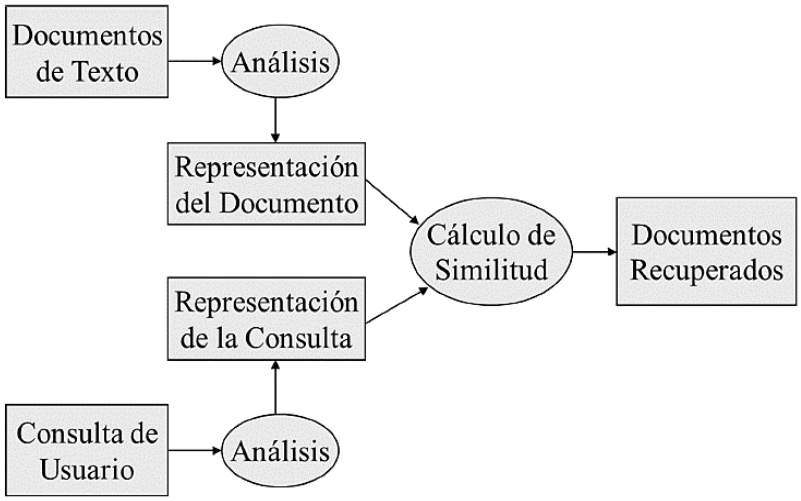

Figura 1: Proceso de recuperación de información. Fuente: García (2013)

El modelo es el marco teórico mediante el cual se diseñan los sistemas de recuperación de información y por eso es la parte fundamental del mismo. La construcción de modelos en esta área surgió en los años 60, pretendiendo ofrecer una base teórica a la gran cantidad de experimentos y desarrollos que se estaban llevando acabo. Las numerosas técnicas diferentes y combinadas que empleaban, junto con la creciente necesidad de evaluar la efectividad de los sistemas, trajo consigo la construcción de patrones más generales en los cuales se pudiera agrupar la variedad de propuestas (Jones, 1997). Muchos de estos modelos se relacionan con diferentes áreas matemáticas y computacionales. En las siguientes secciones se realizará una caracterización de los modelos más utilizados según la literatura consultada, comenzando por los clásicos, los cuales constituyen la base fundamental del resto de los modelos.

1.1. Modelos de Recuperación de Información.

Los modelos clásicos incluyen los 3 más comunes: el booleano basado en la teoría de conjuntos, el vectorial basado en el álgebra y el probabilístico basado en la teoría de probabilidades. Estos modelos han ido evolucionando partiendo de teorías iniciales y han sido extendidos para lograr una

mejor eficiencia y precisión. En este epígrafe se abordará la taxonomía de estos modelos y de algunas de sus extensiones.

Baeza-Yates y Ribeiro-Neto (1999) realizan una caracterización formal de los modelos de recuperación de información:

Definición: Un modelo de RI es una cuádrupla $[\mathbf{D}, \mathbf{Q}, \mathbf{F}, \mathbf{R}(\mathbf{q}_i, \boldsymbol{d}_j)]$, donde:

D es el conjunto de las representaciones lógicas de los documentos de la colección.

Q es el conjunto de las representaciones lógicas de las consultas (necesidades de información del usuario).

F es un marco para modelar los documentos, las consultas y las relaciones entre ambos.

$\mathbf{R}(\mathbf{q}_i, \boldsymbol{d}_j)$ es la función de puntuación (*ranking*), la cual asocia un número real con la representación de una consulta $\mathbf{q}_i$ $(\mathbf{q}_i \in \mathbf{Q})$ y la representación de un documento $\mathbf{d}_j$ $(\mathbf{d}_j \in \mathbf{D})$.

Existen diversas propuestas de clasificación de los modelos. Zhao y Grosky (2002) muestran mediante la figura 2, una clasificación según el modelo clásico del cual ellos desciendan.

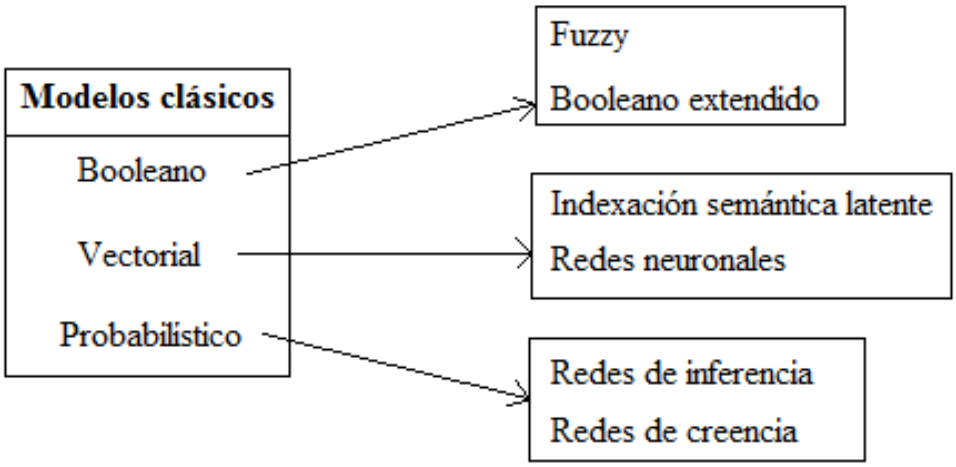

Figura 2: Clasificación de los principales modelos de RI.

Por su parte, Dominich (2000) realiza una síntesis clasificándolos en 5 grupos: modelos clásicos, alternativos, lógicos, los basados en la interactividad y los basados en la Inteligencia Artificial. Los modelos clásicos incluyen el booleano, el vectorial y el probabilístico. Los modelos alternativos están basados en la lógica de Fuzzy. Los modelos lógicos se basan en la lógica formal y la recuperación de la información es un proceso inferencial. Los modelos basados en la interactividad

incluyen posibilidades de expansión del alcance de la búsqueda y hacen uso de retroalimentación por la relevancia de los documentos recuperados. Por último se encuentran los modelos basados en la Inteligencia Artificial donde se encuentran las redes neuronales, algoritmos genéticos y el procesamiento del lenguaje natural.

1.1.1. Booleano

El modelo booleano se basa en la teoría de conjuntos y es la base de los modelos usados para la recuperación de la información debido a su sencillez. En este modelo la representación de la colección de documentos se realiza sobre una matriz binaria términos-documentos, donde los términos han sido extraídos manual o automáticamente de los documentos y representan el contenido de los mismos. Un término puede estar presente o ausente en un documento lo que trae como consecuencia que sólo serán relevantes aquellos documentos que contengan los términos claves por los que está compuesta la consulta. Esto trae consigo que el resultado de una consulta no contendrá documentos que pueden ser relevantes a pesar de no encajar a la perfección con la consulta.

Las consultas se formulan como expresiones lógicas que combinan los términos claves mediante operadores del Álgebra de Boole (AND, OR, NOT y otros). Todos los términos son considerados de igual importancia por lo que dada una consulta, se tendrán en cuenta todos los términos de la misma para la búsqueda. Por estas razones en este modelo existe poco control sobre la cantidad de documentos que se obtendrán en una búsqueda realizada, lo que ocasiona que se tenga una gran cantidad de resultados como respuesta, o todo lo contrario, solo unos pocos. Este modelo no dice qué tan relevante es un documento de otro, asumiendo que todos tienen el mismo grado de importancia. No ordena los documentos por relevancia, tal como se llevaría a cabo en un modelo basado en pesos o ponderación de los términos. Teniendo en cuanta que la relevancia de un documento es binaria, un documento será relevante o totalmente irrelevante (Manning et al., 2008). Como ejemplo de extensiones de este modelo se puede mencionar al booleano extendido y al modelo *fuzzy*.

1.1.1.1. Booleano Extendido

El modelo booleano extendido es una extensión del modelo booleano tradicional, el cual intenta resolver los problemas que este presenta. Este modelo incluye la funcionalidad de calcular los pesos

de cada término en los documentos. En Salton et al. (1983) se describe con mayor detalle, aunque a continuación se explicarán sus aspectos más relevantes.

Este modelo intenta eliminar la utilización de los pesos binarios en los términos utilizando la función (1.1) con lo cual se eliminan la mayor parte de las desventajas en el modelo booleano tradicional. Esta función establece una relación entre la frecuencia de un término (tf) dentro de un documento y su frecuencia en los documentos de la colección (idf).

$$peso = tf * idf \qquad (1.1)$$

En el cálculo de idf, los valores cercanos a 0 indican que el término posee poco peso y por lo tanto, bajo valor de discriminación. Por otro lado, los valores positivos lejanos a 0 indican que el término es poco frecuente y resulta más adecuado para caracterizar a los documentos donde se encuentran.

Por otra parte utiliza la función de similitud $sim(q, d_j)$ la cual permite definir un orden de relevancia en los documentos recuperados. En este modelo también es posible la elaboración de consultas (q) con operadores booleanos, por lo que define dos fórmulas, una para las consultas con el operador booleano OR (1.2) y otra para consultas con el operador booleano AND (1.3). En este caso el valor p permite la generalización de la distancia debido a que existirán p términos en un espacio de m dimensiones; y w_{ij} representa el peso del término en el documento. Si solo existieran dos términos se podría utilizar la distancia Euclidiana.

$$sim(q_{or}, d_j) = \left(\frac{{w_{ij}}^p + \cdots + {w_{mj}}^p}{m}\right)^{1/p} \qquad (1.2)$$

$$sim(q_{and}, d_j) = 1 - \left(\frac{\left(1 - w_{ij}\right)^p + \cdots + \left(1 - w_{mj}\right)^p}{m}\right)^{1/p} \qquad (1.3)$$

Además es posible lograr una combinación en el momento de hacer las consultas y para esto se implementaría una combinación de estas fórmulas hasta lograr la incorporación de todos los términos contenidos en la consulta.

1.1.1.2. Fuzzy

El modelo *fuzzy* (difuso) intenta solucionar también los problemas del modelo booleano basándose en la lógica difusa. Los documentos son representados de forma similar al modelo booleano, se

asignará un 1 si el término se encuentra en el documento y 0 en caso contrario. Este modelo mantiene las relaciones entre los términos indexados con el objetivo de ampliar el conjunto de términos indexados de un documento con otros términos que se relacionen con estos, lo cual posibilita que se puedan recuperar documentos relevantes adicionales (Baeza-Yates y Ribeiro-Neto, 1999). Para almacenar estas relaciones se utiliza una matriz de correlación cuyas filas y columnas están asociadas a términos indexados y en la posición i, j se almacena la relación que existe entre el término i y el término j. La siguiente función (1.4) se utiliza para definir un factor de correlación normalizado ($0 \leq c_{ij} \leq 1$) entre los términos k_i y k_j, donde n_i es la cantidad de documentos que contienen al término k_i, n_j es la cantidad de documentos que contienen al término k_j, y n_{ij} es la cantidad de documentos que contienen ambos términos.

$$c_{ij} = \frac{n_{ij}}{n_i + n_j - n_{ij}} \qquad (1.4)$$

En este modelo se va a definir un conjunto difuso para cada término indexado k_i. Cada documento tendrá un grado de pertenencia a cada uno de dichos conjuntos. El grado de pertenencia del documento d_j al conjunto difuso asociado al término k_i se calcula a partir de la siguiente función.

$$\mu_{ij} = 1 - \prod_{k_i \in d_j} (1 - c_{ij}) \qquad (1.5)$$

Por lo tanto siempre que exista al menos un término k_i del documento d_j que esté fuertemente relacionado con k_i ($c_{ij} \approx 1$) entonces $\mu_{ij} \approx 1$ y el índice k_i es un índice relevante para el documento d_j. En el caso de que todos los índices del documento d_j estén débilmente relacionados con k_i, entonces k_i no es un buen índice difuso para d_j ($\mu_{ij} \approx 0$).

A diferencia del modelo booleano tradicional, define un *ranking* entre los documentos recuperados mediante una función de pertenencia. El objetivo de esta función (1.6) es evaluar la relevancia de un documento d_j a una consulta q.

$$\mu_{qj} = 1 - \prod_{i=1}^{p} \left(1 - \mu_{c_{ij}}\right) \qquad (1.6)$$

Después de calculada la relevancia de cada documento a la consulta se establece un umbral de relevancia y se recuperan todos los documentos cuyo grado de relevancia sea mayor que el umbral definido.

1.1.2. Vectorial

El modelo vectorial representa cada documento como un grupo de términos, y a su vez la unión de estos grupos representa la colección de documentos. Kiela y Clark (2014) explican que este conjunto de términos es definido como un "espacio" donde cada término representa una dimensión en este espacio, el cual es denominado como "espacio de los documentos". A cada uno de los términos se le asigna un peso en cada documento, por lo que un término puede recibir diversos pesos en dependencia del documento en que se encuentre; si el término no aparece en un documento determinado recibe el valor 0, pues estos pesos son asignados en dependencia de la frecuencia de aparición del término en cada uno de los documentos. Los pesos asignados a los términos dentro de un documento se pueden interpretar como las coordenadas del mismo dentro del espacio de los documentos, el documento es representado como un punto en el espacio de los documentos, es interpretado como un vector desde el origen del espacio hasta el punto definido por sus coordenadas

En este modelo también es definido el "espacio de los términos" para una colección de documentos. En este caso cada documento es una dimensión, cada punto o vector en el espacio es un término y sus coordenadas son los pesos asignados en cada uno de los documentos en los que aparece. Ambos espacios se pueden combinar de manera tal que la colección de documentos sea representada por una matriz términos-documentos donde cada fila es un término en el espacio de documentos, de esta forma el elemento de la *i*-ésima fila y la *j*-ésima columna, es el peso del término *i* en el documento *j*.

La consulta realizada por el usuario es vista en este modelo como un grupo de términos, por lo que es procesada al igual que un documento. De esta forma la consulta puede ser interpretada como otro documento en el espacio de los documentos, en el caso de poseer términos que no se encuentran en la colección, estos representarán una dimensión adicional en el espacio de los documentos.

Para evaluar la similitud entre un documento y una consulta; es decir, para obtener la relevancia del documento con respecto a la consulta, se realiza una comparación de los vectores que los representan. Uno de los métodos más habituales para comparar el grado de similitud es calcular el

coseno del ángulo que forman ambos vectores. Cuanto más se parezcan los vectores, más próximo a 0° será el ángulo que forman y, en consecuencia, más se aproximará a 1 el coseno de ese ángulo. Seco (2009) muestra mediante la figura 3 un ejemplo de cómo se vería gráficamente esta similitud.

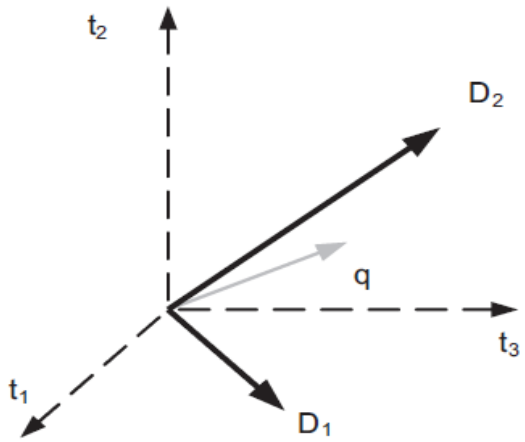

Figura 3: Comparación de similitud en el modelo vectorial. Fuente: Seco (2009)

En este caso D_1 y D_2 son los vectores que representan a dichos documentos y *q* el vector que representa la consulta. Evidentemente el documento D_2 es más relevante para la consulta *q* que el documento D_1. El ángulo que forman los vectores de D_2 y *q* es menor que el que forman los vectores de D_1 y *q*, trayendo como consecuencia que el coseno del primer ángulo sea mayor que el del segundo ángulo.

1.1.2.1. Indexación Semántica Latente

La Indexación Semántica Latente (ISL), conocida en inglés como "*Latent Semantic Indexing*", propuesta por Dumais et al. (1988), es una extensión del modelo vectorial. Consiste en un método matemático apoyado en el Álgebra Lineal. Al igual que el modelo vectorial, este método parte de una matriz donde son asociados los términos y documentos (matriz TxD). Esta técnica de indexación y recuperación de información ha cobrado fama, pues es muy efectiva a la hora de ocuparse de los problemas de sinonimia y polisemia. La idea principal es acotar cada documento y consulta en un pequeño espacio dimensional el cual es asociado con los conceptos (Bradford, 2009).

En la ISL el concepto en cuestión, así como también todos los documentos que están relacionados con él, serán representados por una misma combinación ponderada de variables indexadas. Además

se aplica la factorización matricial conocida como Descomposición en Valores Singulares (DVS) de esta matriz. Esta descomposición constituye un aspecto importante pues permite la reducción de ruido encontrado en la colección. El ruido se considera a la existencia de varios documentos que tengan algún término de una consulta y aun así no son considerados como documentos relevantes debido a que su semántica no es semejante a lo que se busca. En esta técnica los documentos que se consideran que tienen la misma semántica se localizan en una posición cercana en un espacio multidimensional y este proceso se realiza mediante la DVS (Calabuig et al., 2015).

Mediante la figura 4 Botana (2010) representa esquemáticamente las transformaciones llevadas a cabo por medio de la DVS. En la figura 4a, cada término está representado por cuatro dimensiones, tantas como documentos existen en el corpus (d1, d2, d3, d4). En la figura 4b, los términos pasan a estar representados por dos dimensiones abstractas pero de una mayor utilidad funcional. A cada término se le infiere una probabilidad de estar representado en un concepto, por ejemplo, al término t2 se le infiere cierta probabilidad de encontrarse en documento d2 aunque como muestra la figura 4a, esto no se produzca. El autor plantea que la DVS hace el análisis semántico más robusto y capaz de explotar el significado de los textos con fenómenos como la sinonimia y la homonimia.

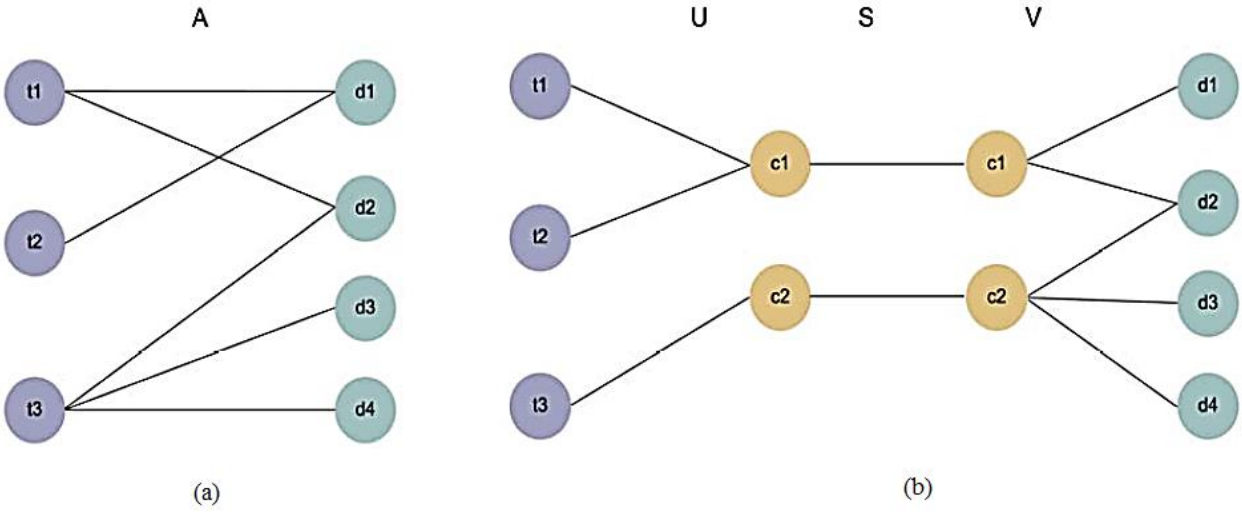

Figura 4: Representación de la DVS de una matriz. Fuente: Botana (2010)

Según Botana (2010) lo relevante es que la matriz original puede ser reconstruida sin emplear todas las dimensiones de la matriz original, solo utilizando las dimensiones que más peso tienen y cuya ponderación está descrita en la matriz diagonal de valores singulares (matriz S). Como los valores de la matriz de valores singulares están ordenados de mayor a menor será fácil tomar sólo los que

más ayuden a construir agrupaciones, reduciendo las dimensiones y volviendo a multiplicar. La multiplicación de las tres matrices resultará otra parecida a la original, pero con la particularidad de que se ha reducido gran parte del ruido que ejercían las dimensiones que no eran del todo relevantes en las relaciones entre términos y documentos. El número de dimensiones (k) que se tomarán en esta nueva multiplicación sigue siendo un tema en estudio. Posteriormente se aplica una función de similitud semántica, en este caso la consulta realizada por el usuario es definida como un conjunto de términos, es decir, es otro vector dentro del espacio de los documentos. Con el análisis semántico latente se pretende modelar matemáticamente estas relaciones de asociación términos-documentos. En la figura 5 se vincula cada paso de la ISL con la rama de la matemática asociada a él.

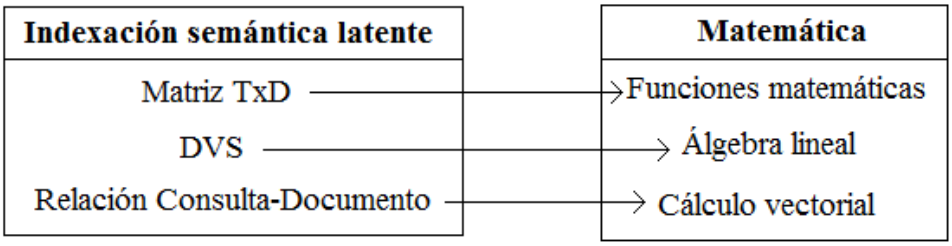

Figura 5: Asociación de la Indexación Semántica Latente con ramas de la matemática

1.1.3. Probabilístico

En el modelo probabilístico el proceso de recuperación de información es intrínsecamente impreciso. Su diferencia radica en el método estadístico y en las premisas bajo las que se constituye su funcionamiento. Existen determinados aspectos que son no deterministas como son la representación que hace una consulta de la necesidad de información del usuario y la representación de los documentos en el sistema (Cacheda, 2008). Al igual que en el modelo booleano se fundamenta en la representación binaria de los documentos. De esta forma cualquier documento tendría la siguiente representación:

$$D_j = (T_1, T_2, \dots, T_M) \qquad (1.7)$$

donde T_i = 0 o 1, indica la ausencia o presencia del *i*-ésimo término respectivamente, y *M* el número de términos de la colección. Además existen dos eventos excluyentes: ω_1, que representa el hecho de que un documento sea relevante, y ω_2 que indica que no lo sea. El modelo asume que dada la consulta planteada por el usuario, se conocen dentro de la colección de documentos dos grupos:

Conjunto de Documentos Relevantes y Conjunto de Documentos Irrelevantes. El objetivo que se persigue es calcular $p(\omega_1|D_j)$ y $p(\omega_2|D_j)$, es decir, la probabilidad de que el documento D_j sea relevante o no relevante, dada una consulta Q, respectivamente. Además, desarrollar una función (1.8) que ofrezca un valor de relevancia para así ordenar los documentos.

$$Sim(D_j, Q) = \frac{p(\omega_1|D_j)}{p(\omega_2|D_j)} \qquad (1.8)$$

En otras palabras, dada una consulta Q y un documento D_j, el modelo probabilístico permite determinar la probabilidad de que el usuario encuentre el documento relevante. Se asume que esta probabilidad de relevancia depende sólo de las representaciones del documento y de la consulta. También se asume que hay un subconjunto de todos los documentos que el usuario prefiere como respuesta a su consulta, al cual se le denomina como conjunto de respuesta ideal (*R*). El conjunto *R* debería maximizar la probabilidad global de relevancia para el usuario. Los documentos que no pertenezcan al conjunto serán considerados como no relevantes para el usuario. Un documento se considera relevante si su probabilidad de ser relevante $p(\omega_1|D_j)$ es mayor que la probabilidad de no ser relevante $p(\omega_2|D_j)$. El modelo es capaz de ordenar los documentos de la colección conforme al orden decreciente de su probabilidad de relevancia con respecto a la consulta del usuario, es decir, se mostrará en primer lugar el documento cuya probabilidad de relevancia sea más alta (Alimohammadi, 2010).

1.1.3.1. Redes de Inferencia

El modelo de redes de inferencia es una extensión del modelo probabilístico basados en redes bayesianas. Una red bayesiana representa un conjunto de variables aleatorias y sus dependencias. Suelen ser ciclos cuyos nodos representan cada una de estas variables. Estos modelos están formados por dos subredes o vectores binarios. La figura 6 muestra un ejemplo de red de inferencia representado por Grossman y Frieder (2012), donde se muestran la consulta y tres documentos, cada cual con sus términos respectivamente.

La red de documentos es una red fija para cada colección, formada por dos tipos de nodos que representan los términos de los documentos y los documentos respectivamente. De un nodo documento salen arcos hacia los nodos de los términos que han sido indexados. Por su parte, la red de consulta es una red que se crea cuando el usuario realiza una consulta, esta contiene nodos

consultas y nodos de términos, de manera que de un nodo término salen arcos hacia los nodos consultas correspondientes (Turtle y Croft, 1989).

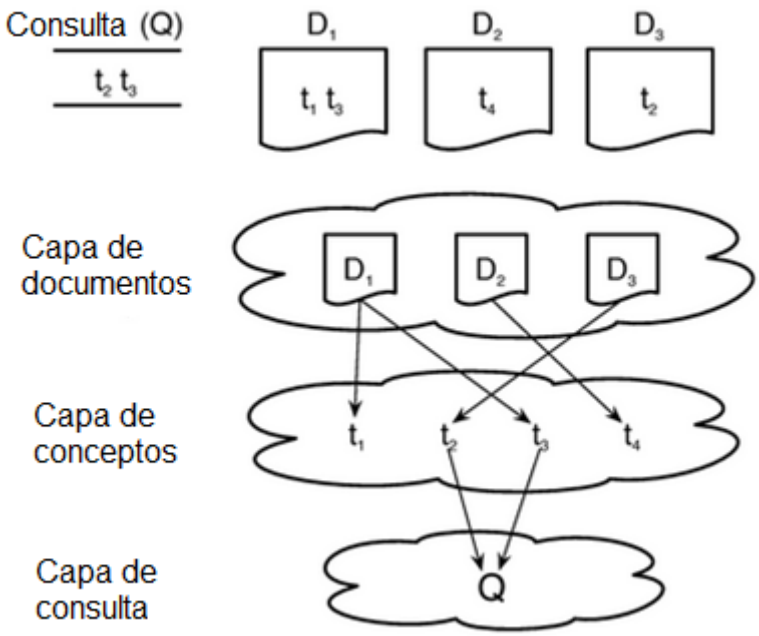

Figura 6: Ejemplo de una red de inferencia.

1.2. La evaluación en la recuperación de información

La información obtenida tras una consulta puede ser evaluada desde 2 puntos de vista: uno algorítmico y otro orientado a los usuarios. En este epígrafe se abordará la evaluación desde el enfoque algorítmico el cual intenta medir si la respuesta (documentos recuperados) es adecuada a la consulta realizada.

Baeza-Yates y Ribeiro-Neto (1999) exponen que tras una consulta, la colección de documentos quedaría dividida como se muestra en la figura 7. En el conjunto de documentos recuperados existirán los documentos relevantes recuperados, es decir aquellos que se han recuperados correctamente; y los no relevantes, los cuales serán recuperados erróneamente. También se encuentran los documentos no recuperados que se dividen en los relevantes rechazados por el sistema de manera errónea y los no relevantes, rechazados de manera correcta por el sistema.

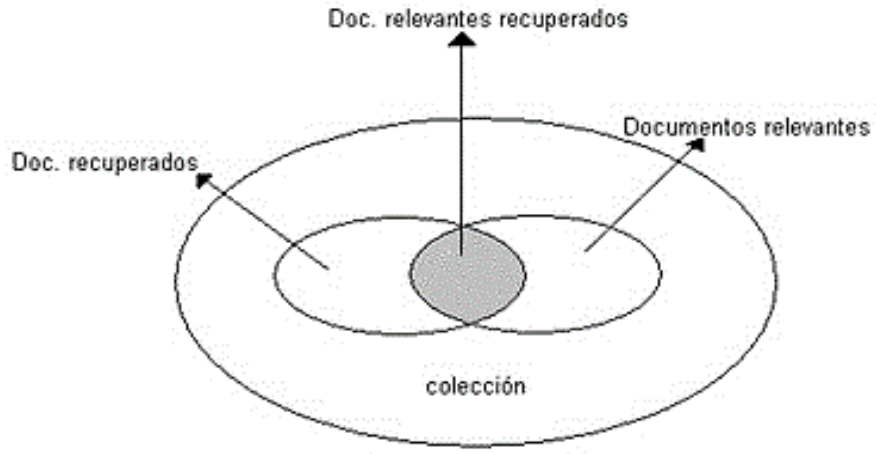

Figura 7: Recuperación de documentos en una colección. Fuente: Baeza-Yates y Ribeiro-Neto (1999)

1.2.1. Relevancia y Pertinencia

Decidir cuáles documentos son relevantes o no, depende de diversos factores, dentro de los que se encuentran quiénes son los expertos que analizan los documentos, los criterios que se utilizan para clasificarlo y los usuarios. La relevancia se refiere a la utilidad de los documentos recuperados, con relación a la satisfacción de los objetivos del usuario. De esta forma 2 usuarios distintos pueden decidir que el mismo documento es y no es, respectivamente, relevante a una consulta dada (Saracevic et al., 1988). Por estos motivos, con colecciones reales de documentos es difícil conocer el conjunto de documentos relevantes para cada consulta.

La relevancia en esta cuestión se puede resumir en dos tendencias teniendo en cuenta lo planteado por Swanson (1988): la relevancia objetiva y la subjetiva o pertinencia. La relevancia objetiva está basada en los sistemas pues define cómo el tema de la información recuperada coincide con el de la consulta. Por su parte la subjetiva o pertinencia es la que tiene en cuenta al usuario. Según Korfhage (1997), la relevancia es la medida de cómo una pregunta se ajusta a un documento mientras que la pertinencia es la medida de cómo un documento se ajusta a una necesidad informativa. La valoración de la pertinencia es mucho más difícil de realizar ya que es el propio usuario el único que sabe si un documento se ajusta a su necesidad o no.

La relevancia de los documentos se realiza de forma manual. Consiste en la exploración de los documentos uno a uno para saber si se adecúan o no como respuesta a una pregunta. El principal problema que presenta este método, es que en colecciones muy grandes, hay que invertir gran

cantidad de tiempo. Además, algunas colecciones son más especializadas que otras, lo que hace necesario contar con un número mayor o menor de especialistas. Para solucionar estos problemas se crean colecciones experimentales donde se establece de antemano qué documentos son relevantes para cada pregunta.

1.2.2. Medidas para calcular la efectividad en la RI

Las dos medidas más frecuentes y básicas para la efectividad en la recuperación de información son precisión y cobertura (*precision, recall*), también en este subepígrafe se explicará una de las medidas alternativas debido a su grado de efectividad. Baeza-Yates y Ribeiro-Neto (1999) y Manning et al. (2008) realizan amplias descripciones sobre estas medidas.

La precisión es la proporción de los documentos recuperados que realmente son relevantes, con respecto al total de los documentos recuperados. Su valor varía entre 0 y 1, siendo 1 si la recuperación fuera perfecta, es decir, si todos los documentos relevantes fueran los recuperados.

$$Precisión = \frac{documentos\ relevantes\ recuperados}{documentos\ recuperados} \qquad (1.9)$$

Para obtener una alta precisión es necesario que los términos aparezcan con frecuencia alta pero en pocos documentos y nula en el resto. Como la salida obtenida en la recuperación es ordenada en función de la relevancia y los documentos más relevantes están al comienzo de la salida, a medida que se avanza en el número de documentos recuperados, la precisión decae.

La cobertura es la proporción de los documentos relevantes recuperados con respecto al total de los documentos que son relevantes en la colección. Esta medida es inversamente proporcional a la precisión. Al igual que la precisión su valor oscila entre 0 y 1 .Si el valor es 1 se está en presencia de una recuperación perfecta, pues indica que se han encontrado todos los documentos relevantes existentes en la colección. Para conseguir una alta recuperación es necesario que los términos aparezcan en muchos documentos de la colección.

$$\text{Cobertura} = \frac{documentos\ relevantes\ recuperados}{documentos\ relevantes} \qquad (1.10)$$

Un buen sistema de RI debe presentar buenas medidas tanto de precisión como de cobertura pues una cobra sentido en función de la otra. Se podría alcanzar una cobertura del 100% recuperando todos los documentos ya que en el conjunto resultado están incluidos todos los documentos

relevantes. Sin embargo, la precisión en ese caso sería muy baja ya que también se están recuperando muchos documentos no relevantes.

Resumiendo, Gelbukh y Sidorov (2010) expresan que la precisión es la relación entre los resultados correctos sobre los resultados obtenidos en total, mientras que la cobertura es la relación entre los resultados correctos sobre los resultados que deberían haber sido obtenidos. La precisión indica qué tan preciso fue la recuperación de información, mientras que la cobertura da a conocer si se recuperaron todos los resultados que debían ser obtenidos. Para un usuario la situación ideal es aquella en la que existe una precisión y recuperación alta. Con el propósito de ponderar y ver cuán lejano se encuentran ambas medidas suele emplearse los valores de ambas métricas combinadas en la medida *F* (Beitzel, 2006).

$$F = \frac{(\beta^2 + 1)PC}{\beta^2 P + C} \qquad (1.11)$$

La medida F se utiliza para corregir el error de la distancia cuando la cobertura (C) y la precisión (P) se compensan. El valor de β es preestablecido, teniendo en cuenta que si $\beta = 1$ se le está dando la misma importancia a P que a C, si $\beta > 1$ se le está dando más importancia a C y si $\beta < 1$ se le está dando más importancia a P. Lo interesante de esta métrica es que un máximo valor de F corresponde al mejor compromiso entre P y C y su valor solamente será alto cuando ambas componentes tengan valores altos.

1.3. Aplicación de la matemática en los buscadores en Internet

Cada buscador en Internet necesita de un algoritmo para determinar el orden de las páginas devueltas por cualquier consulta y aquí es donde las herramientas matemáticas son la clave para la solución. Sin duda, uno de los motores de búsqueda de mayor éxito es Google, y su éxito se atribuye principalmente a su algoritmo de búsqueda.

1.3.1. PageRank de Google

Google utiliza el algoritmo PageRank para decidir la importancia de una página web. También se basa en los sistemas de puntuación (*ranking*) de las publicaciones científicas, es decir, un artículo es importante en la medida que ha sido citado por otros colegas del área. El *ranking* en Google también tiene en cuenta factores relacionados con la proximidad de la búsqueda a las palabras claves de la página, obtenidas en el indexado inverso. Es válido destacar que realmente no se sabe

con exactitud cómo funciona el algoritmo de Google salvo sus propios creadores Lawrence Page y Sergey Brin. La teoría más difundida es la del PageRank (PR), una familia de algoritmos que sirve para asignar de forma numérica la importancia y relevancia de los documentos o páginas web que aparecerán en la lista. Este algoritmo es extremadamente complejo y se actualiza continuamente.

Google evalúa cada página web y calcula su importancia en relación con otras páginas web, donde la importancia de una página web puede ser determinada por la importancia de las páginas que la referencian. Por ejemplo: cada enlace de una página u con destino v le da un voto a v. Pero, si se quiere tener en cuenta la importancia de la página u, se debe balancear los votos que ella ha recibido de otras. Cuando se está parado en una página se le denomina retroenlaces a todos aquellos que hacen referencia a ella, y ultraenlaces a los enlaces que esa página tiene hacia otras. En la figura 8 se muestra un ejemplo expuesto por Gracia (2002) donde se supone que una página u es destino de 100 enlaces por lo que tiene 100 votos. Si esta página tiene 2 enlaces hacia adelante envía un valor de 50 votos a cada uno. En caso que sea v uno de estos 2 destinos y a su vez recibe también 3 votos de otro retroenlace, obtiene un total de 53 votos. Si v, a su vez, tiene dos ultraenlaces, aporta a cada uno de sus destinos 53/2 votos, y así sucesivamente. Esto funciona como un grafo dirigido donde las páginas web son nodos y los arcos son enlaces de una página a otra.

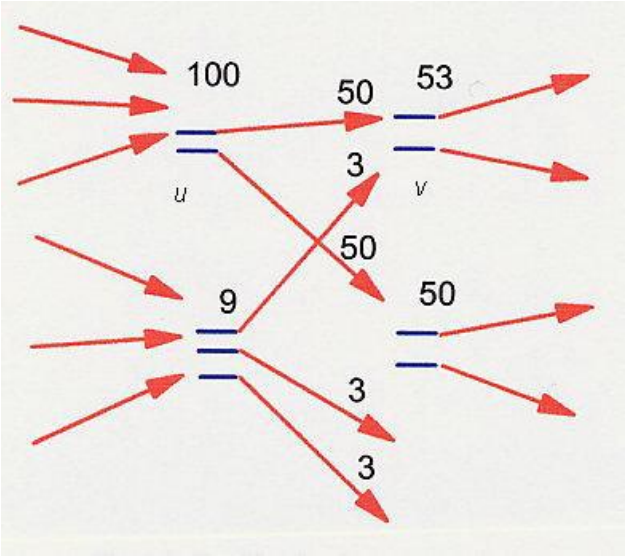

Figura 8: Funcionamiento de la indexación de las páginas en Google. Fuente: Gracia (2002)

El algoritmo inicial del PageRank se puede encontrar en Seco (2009), aunque en la actualidad ha sufrido diversas transformaciones en busca de la perfección. Este autor plantea que:

$$PR(A) = (1-d) + d\sum_{i=1}^{n}\frac{PR(i)}{C(i)} \qquad (1.12)$$

En esta función, A es una página web a la que se le aplica el PageRank, d es un factor de amortiguación que tiene un valor entre 0 y 1, $PR(i)$ son los valores de PageRank que tienen cada una de las páginas i que enlazan a A y $C(i)$ es el número total de enlaces salientes o ultraenlaces de la página i ya sean o no hacia A.

Gracia (2002) y Harikumar y Srinivasan (2015), coinciden en que la siguiente función es una versión aproximada del PageRank, donde W es el conjunto de todas las páginas web, u una página web, F_u el conjunto de páginas web a las que enlaza la página u, y denota por B_u el conjunto de páginas que apuntan hacia u. ($N_u := |F_u|$ el cardinal de F_u)

$$PR: W \to [0, \infty) \qquad (1.13)$$

Esto satisface la ecuación:

$$PR(u) = \sum_{v \in B_u}\frac{R(v)}{N_v} \qquad (1.14)$$

De manera general, Harikumar y Srinivasan (2015) agregan que una matriz de hipervínculo *H*, donde $outdeg(v)$ es el número ultraenlaces de la página v puede ser definida como:

$$H_{uv} = \begin{cases} 1/outdeg(v) & si\ v\ \in B(u) \\ 0 & otra\ forma \end{cases} \qquad (1.15)$$

Gallardo (2006) explica mejor esta cuestión y plantea que la red puede ser descrita mediante un grafo dirigido. De esta forma, una página P_j de la red es un vértice del grafo y existirá una arista entre los vértices P_i y P_j si desde la página P_i hay un enlace a la página P_j. En los grafos son representados los vértices mediante círculos mientras que las aristas son flechas que unen estos círculos. El autor considera una interpretación alternativa, en este caso matricial y conforma la matriz M (véase figura 9) de dimensiones nxn cuyas filas y columnas van etiquetados con los símbolos P_1 ... ... P_n, y sus entradas son ceros y unos. La entrada m_{ij} de la matriz será un 1 si es que hay un enlace de la página P_j a la página P_i y un cero en caso contrario. Esta sería la matriz de adyacencia del grafo, la cual no tiene por qué ser simétrica, pues se plantea que es un grafo dirigido. De esta forma la suma de las entradas correspondientes a la columna de P_j es el número de enlaces

que salen de la página y a su vez la suma de los registros de una fila coincide con el número de enlaces entrantes.

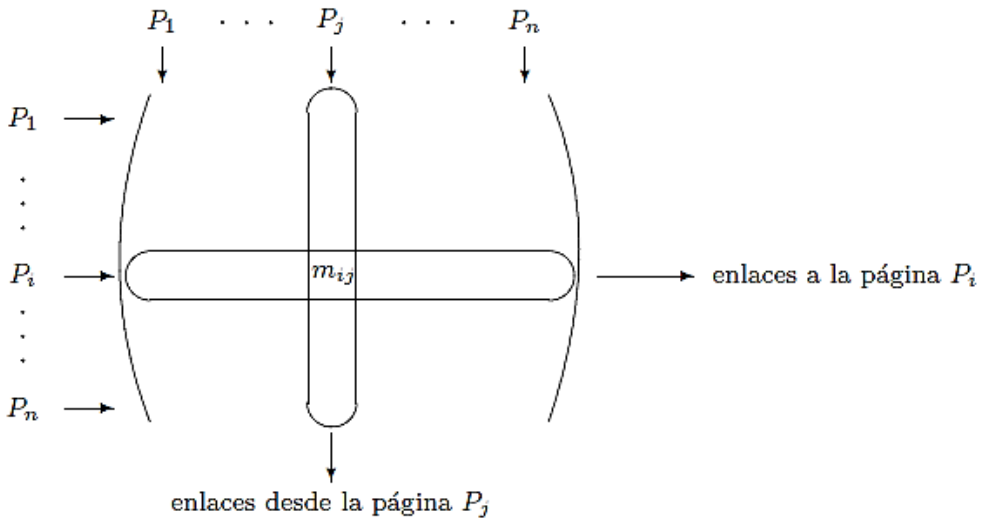

Figura 9: Matriz de enlaces entre las páginas en Internet. Fuente: Gallardo (2006).

La importancia de una página es proporcional a la suma de las importancias de las páginas que enlazan con ella. Gallardo (op. cit.) basándose en este principio, sugiere que puede ser calculada determinando un vector x, en el que tendrán lugar las importancias que se desean calcular, es decir x_j sería la importancia de P_j. Estos valores se obtendrían a partir de la ecuación 1.16, donde λ es la constante de proporcionalidad. De esta forma se transforma la cuestión en un problema de autovalores y autovectores. El vector de importancias x es un autovector de la matriz M.

$$Mx = \lambda x \qquad (1.16)$$

Por regla general, el dominio o página principal acumula un mayor PR que las páginas o secciones internas debido a que la mayoría de los enlaces externos suelen apuntar directamente al dominio. Google únicamente tiene en cuenta los enlaces que conoce, los enlaces desde páginas no indexadas o penalizadas no se tienen en cuenta.

1.4. CUBA: buscador cubano basado en los modelos clásicos de RI

Actualmente nuestro país cuenta con un buscador que permite recuperar información publicada sólo en el dominio (.cu), lo que constituye una alternativa para las personas que hoy sólo tienen acceso a la intranet nacional. La plataforma digital Contenidos Unificados para Búsqueda Avanzada (CUBA) fue presentada julio del 2015 (GranmaTV, 2015). Esta plataforma cuenta con

un diseño que se adapta a los diferentes terminales (ordenadores, tabletas, teléfonos móviles). Dicho buscador opera con el motor de búsqueda ORION, desarrollado por la UCI hace algunos años.

ORION utiliza el *spider mnoGoSearch* que es software libre, lanzado bajo la licencia GPL de la *Free Software Fundation*. Este *spider* integra la mayoría de los conceptos y características asociadas a los motores de búsqueda actuales. Para el cálculo del Page Rank, emplea algoritmos del modelo vectorial y booleano. Ambos algoritmos han sido ampliamente probados y utilizados, obteniéndose muy buenos resultados con los mismos (Barkov, 2009). ORION en su momento garantizó una visible optimización de los tiempos empleados en la búsqueda y recuperación de la información existente en la red interna de la UCI permitiendo encontrar la información con un adecuado grado de relevancia (Nieto y Rueda, 2010).

1.5. Conclusiones parciales

Una vez caracterizado el objeto de estudio de la actual investigación se considera apropiado a los efectos de la investigación que se llevará a cabo, utilizar la técnica Indexación Semántica Latente pues, al pasar al espacio de los conceptos, se pueden recuperar documentos que traten sobre el mismo tema e incluso recuperar documentos aun cuando no contengan los términos presentes en la consulta. Además, el desarrollo de la investigación se fundamenta sobre la base de investigaciones realizadas por Reina et al. (2013) y Archana y Ravichandran (2014), que demuestran ventajas de la utilización de esta técnica en el mundo del *e-learning* obteniendo resultados satisfactorios. Por otra parte, se pudo corroborar la coincidencia de algunos autores que plantean que la relevancia de un documento depende fundamentalmente de la necesidad del usuario que realiza la consulta, lo cual dificulta la determinación de los documentos relevantes en una colección. Esto confirma que la evaluación de los sistemas de recuperación de información es un tanto subjetiva.

CAPÍTULO 2: INDEXACIÓN SEMÁNTICA LATENTE. PRINCIPALES ELEMENTOS DE LA INVESTIGACIÓN

Una vez estudiado el estado de la cuestión se hace imprescindible describir cada una de las etapas de la Indexación Semántica Latente, así como todo el procesamiento necesario que debe realizarse antes de aplicar esta técnica. La figura 10 muestra en forma de resumen estos aspectos y en los próximos epígrafes se abordarán estos temas.

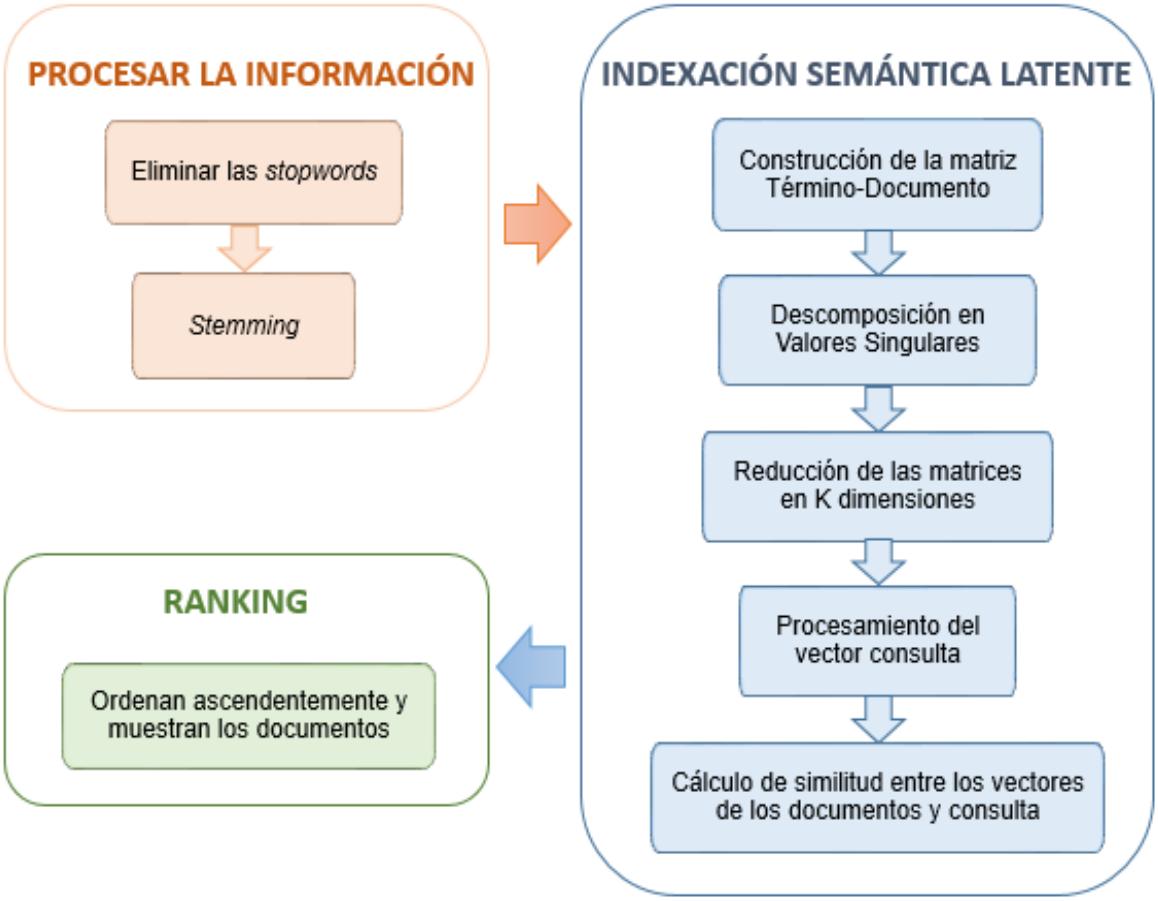

Figura 10: Proceso de recuperación de información.

Es oportuno caracterizar también en este capítulo las tecnologías, métodos y herramientas necesarias para el desarrollo de la investigación. Además es esencial la caracterización de la colección experimental de documentos, pues permitirá confirmar la efectividad de la técnica y aprobará su despliegue en un escenario real.

2.1. Métodos de investigación

Para el desarrollo de la investigación se tuvieron en cuenta los siguientes métodos:

Histórico-Lógico: se utilizó con el objetivo de analizar los modelos matemáticos existentes para la RI. Para esto se tuvo en cuenta la evolución y desarrollo de los mismos con el transcurso del tiempo, lo cual dio lugar a que fueran caracterizadas en el capítulo algunas de sus extensiones. Esto permitió en la investigación decidir cuál modelo sería el más factible para la etapa de implementación.

Inductivo-deductivo: Se tuvo en cuenta para el análisis de las diversas aplicaciones de la ISL para la recuperación de información, es decir, en el estudio de casos particulares donde esta técnica proyectaba resultados beneficiosos.

Sistémico: se aplicó para explicar cada una de las etapas de la ISL teniendo en cuenta la confección y las transformaciones de sus componentes iniciales (matriz de términos-documentos, consulta) y las relaciones entre ellos.

Matemático – Estadístico: Se empleó mediante la implementación de la técnica ISL, resolviendo funciones logarítmicas, realizando operaciones básicas con matrices, descomposiciones, calculando distancia entre vectores y otras acciones que se describen en próximos epígrafes.

2.2. Colecciones de documentos experimentales

Tal como se explicó en el epígrafe 1.2, para la evaluación de un sistema de recuperación de existen colecciones experimentales que son utilizadas a nivel mundial. El grupo de Recuperación de Información de la Universidad de Glasgow[3] tiene publicadas una serie de colecciones experimentales públicas en internet con el objetivo de que otros investigadores evalúen sus algoritmos a partir de ellas. Este grupo de investigación forma parte del Departamento de Ciencias de la Computación de dicha universidad. Tiene un amplio programa de investigación, basada en la teoría y la experimentación; y está dirigido por el profesor Keith van Rijsbergen, un informático teórico, informatólogo e investigador del procesamiento del lenguaje natural y de la disciplina Recuperación de Información. El grupo juega un papel de liderazgo en la comunidad internacional de recuperación de información y ha establecido tendencias en muchos aspectos de la investigación; constituyendo uno de los centros de investigación más antiguos e importantes del mundo que abordan esta temática. Los miembros del grupo también han sido ampliamente

[3] http://ir.dcs.gla.ac.uk

involucrados en la organización de grandes conferencias, talleres y cursos de verano en el área de recuperación de información (Group-IR, 2016). La colección que se utilizó en la experimentación fue la MEDLINE (Lindberg, 2000). Esta colección es una muestra de la base de datos bibliográfica de ciencias de la vida y de información biomédica. También incluye información bibliográfica para los artículos de revistas académicas que cubren la medicina, enfermería, farmacia, odontología, medicina veterinaria y cuidados de salud. MEDLINE también cubre gran parte de la literatura en Biología y Bioquímica, así como campos tales como la evolución molecular. Esta colección ya tiene establecida una serie de consultas y también están identificados los documentos que son relevantes para cada una (juicios de relevancia). Se utilizaron para la experimentación 550 resúmenes de artículos y 19 consultas. La figura 11 muestra la estructura de un documento, la figura 12 de una consulta y la figura 13 de un juicio de relevancia (C: consulta, DR: documento relevante); todas son muestras de la colección MEDLINE.

```
.I 468
.W
3553. diffuse nodular amyloidosis of the lungs
in a 59-year-old man, who 12 yr. previously
had ceased working in an aluminium
factory and who had died from respiratory
insufficiency, the postmortem examination
revealed disseminated nodules in the lungs.
the nodules were sharply defined, of
a greyish color and of a doughy consistency
and could easily be removed from their
capsules.  the diagnosis of amyloidosis was
made on the basis of the histological
findings, the staining qualities of the material,
its intraseptal and periarteriolar
localization, and the existence of plasmocytic
infiltrates and foreign-body giant-cell
granulomas.  amyloid deposits were also found
in the spleen, kidney and some
coronary branches.  in regard to the pathogenesis,
it is suggested that the exposure
to aluminium products for more than 10 yr. might
have constituted a predisposing
factor.
```

Figura 11: Estructura de un documento de la colección MEDLINE

```
.I 11
.W
 blood or urinary steroids in human breast or prostatic neoplasms.
```

Figura 12: Estructura de una consulta de la colección MEDLINE

```
C   DR
8   52
8   60
8   61
8   123
8   190
8   251
8   262
8   263
8   264
8   265
8   266
```

Figura 13: Estructura de un juicio de relevancia de la colección MEDLINE

Tanto a los documentos como a las consultas, antes de comenzar a conformar la matriz TxD, es necesario aplicarles algunas técnicas de procesamiento del lenguaje natural, las cuales se explican en el siguiente subepígrafe. Estas técnicas se utilizan con el objetivo eliminar las palabras irrelevantes en un documento lo cual permitirá posteriormente analizar solo aquellos términos que verdaderamente tienen importancia y aportan información, ya que no todas las palabras poseen el mismo nivel de significado para representar al documento.

2.2.1. Técnicas de procesamiento de los documentos.

La primera técnica que se aplicó para reducir ruido lingüístico dentro de las colecciones de documentos consiste en la eliminación de estructuras que no van a aportar ningún beneficio. Los términos que ocurren en casi todos los documentos de una colección no son buenos para la recuperación de información por su reducido poder para discriminar documentos. Con este objetivo se plantea eliminar las palabras vacías (*stop words*), dentro de las cuales se encuentran los pronombres, adjetivos comunes, adverbios, artículos preposiciones, conjunciones, etc. (Blanchard, 2007). Una ventaja adicional de la eliminación de estos términos es la reducción del tamaño de almacenamiento necesario para soportar el corpus. En el anexo 1 se encuentra la lista de palabras vacías que se tuvieron en cuenta para el procesamiento de los textos.

Posteriormente, se aplica otra de las técnicas más utilizadas: el *stemming*, la cual se basa en eliminar las desinencias o terminaciones (sufijos, plurales, etc.) y conservar las raíces de los términos de manera que palabras de la misma familia sean representadas por un mismo término; por ejemplo: "*cognitive*", "*cognition*" y "*cognitively*" pueden ser reducidos a "*cognit*". Esta técnica de reducción permite detectar variantes morfológicas de un mismo término. El algoritmo clásico es el de Porter

(1980) cuya versión original está para la lengua inglesa. Las reglas que forman los algoritmos clásicos de *stemming* dependen del idioma de las colecciones de documentos a procesar.

2.3. Construcción de la matriz Términos-Documentos

Un paso fundamental en la ISL es la confección de la matriz términos-documentos (matriz TxD) pues existe una variedad de funciones (ponderaciones) o combinaciones de las mismas para determinar el peso de los términos en cada uno de los documento. Esto es válido también para los términos de la consulta realizada por el usuario, pues es procesada como si fuera un documento más. Según Chisholm y Kolda (1999) de manera general, el peso de un término ($\boldsymbol{W_{ij}}$) viene dado por 3 diferentes tipos de pesos como se muestra en la función 2.1. El peso del término de manera local ($\boldsymbol{L_{ij}}$) consiste en su frecuencia de aparición en un documento dado. El peso global ($\boldsymbol{G_i}$) viene dado por la frecuencia de aparición del término en toda la colección de documentos; y por último la normalización de la longitud del documento ($\boldsymbol{N_j}$), pues en documentos extensos los términos suelen tener una mayor frecuencia de aparición, lo cual no hace que el documento sea más importante.

$$\boldsymbol{W_{ij} = L_{ij} * G_i * N_j} \quad (2.1)$$

Para cada uno de estos tipos de pesos existen numerosas funciones y derivaciones de las mismas, propuestas por una variedad de autores dentro de los cuales se encuentran Chisholm y Kolda (1999), Kumar y Srinivas (2009), Nassar et al. (2010), Zaman y Brown (2010) y Singh y Singh (2015), los cuales presentan según sus investigaciones, disímiles criterios al respecto. A continuación se fundamentan las más utilizadas según la bibliografía consultada y son las que se implementarán para un posterior análisis.

2.3.1. Funciones de peso local ($\boldsymbol{L_{ij}}$)

Los pesos de los términos usualmente suelen ser normalizados con el propósito de que el peso de un término dependa de su frecuencia de ocurrencia relativa hacia otros términos en el mismo documento, no de su frecuencia absoluta (tf_{ij}). El peso de un término por su frecuencia absoluta tiende a favorecer a los documentos más largos dentro de la colección (Greengrass, 2000). Una de las funciones que se utiliza es la Normalización Máxima (2.2), esta consiste en dividir la frecuencia absoluta entre la frecuencia del término que más ocurre en el documento ($tfmax$).

$$L_{ij} = \frac{tf_{ij}}{tfmax} \qquad (2.2)$$

El resultado de esta función es la obtención de valores entre 0 y 1, evitando los valores extremadamente elevados. Existe una variación de esta función la cual es denominada como "aumentada" (2.3), en este caso los valores obtenidos oscilan entre 0,5 y 1, y se obtiene a partir de la siguiente fórmula:

$$L_{ij} = 0{,}5 + \left(0{,}5 * \left(\frac{tf_{ij}}{tfmax}\right)\right) \qquad (2.3)$$

Ambas funciones dependen exclusivamente de la frecuencia del término que mayor ocurrencia presenta dentro del documento, lo que constituye una desventaja cuando exista el caso de un término con una desproporcionada elevada frecuencia de aparición.

Como una alternativa se puede utilizar la función logarítmica (2.4), la cual disminuye la importancia de la frecuencia del término en colecciones con documentos de longitudes muy variables y al mismo tiempo reduce el efecto de un término con una frecuencia inusualmente alta dentro de un documento.

$$L_{ij} = log\left(tf_{ij}\right) + 1 \qquad (2.4)$$

Hacía varios años atrás que Singhal et al. (1996a) identificaron que el cálculo de los pesos utilizando la máxima normalización no es un método óptimo para la normalización de la frecuencia del término porque lo significante es la frecuencia del término relativa hacia las frecuencias de todos los otros términos dentro del documento, no solo relativa al término de mayor frecuencia. Por estas razones el autor propone la siguiente función:

$$L_{ij} = \frac{1 + log\left(tf_{ij}\right)}{1 + log(promedio\ tf)} \qquad (2.5)$$

2.3.2. Funciones de peso global (G_i)

Para analizar el comportamiento del término en toda la colección, Greengrass (2000) sugiere que una de las funciones más abarcadoras parte de la frecuencia inversa del documento (idf). Esta función caracteriza la aparición del término dentro de una colección de documentos, es una medida de cómo está el término distribuido en la colección. En ella se tiene en cuenta la cantidad de documentos de la colección (N) y la cantidad de documentos en los que aparece el término (n).

$$G_i = idf = \ln(N/n) \quad (2.6)$$

Por consiguiente, mientras menos documentos contengan el término el valor de idf aumentará, por el contrario, si todos los documentos contienen ese término tomará valor 0 demostrando que un término que ocurra en todos los documentos de la colección probablemente no sea usado para distinguir documentos acerca de un tema en particular, de otros documentos que aborden otras temáticas.

Uno de los esquemas de ponderación más sofisticado según Dumais (1992) lo constituye el cálculo de la entropía pues está basada en la información teórica de las ideas. Esta función tiene en cuenta la distribución de los términos en los documentos y asigna valores entre 0 y 1 para el término que aparece en un solo documento. Si un término aparece una vez en todos los documentos, entonces se le asigna 0 como el valor del peso. Si un término aparece una vez en un solo documento, entonces se le asigna 1 como el valor del peso. Cualquier otra combinación de frecuencias producirá un peso entre 0 y 1.

$$G_i = 1 + \sum_{j=1}^{N} \frac{\frac{tf_{ij}}{F_i} \log \frac{tf_{ij}}{F_i}}{\log N} \quad (2.7)$$

Esta función tiene en cuenta la frecuencia del término i en toda la colección (F_i) y la cantidad de documentos presentes en la misma (N). La entropía tiende a dar mayor valor a los términos que aparecen menos veces y en un pequeño número de documentos.

2.3.3. Funciones de normalización de la longitud del documento (N_j)

El factor de normalización se utiliza para corregir las diferencias en la longitud de los documentos, para que de esta manera sean recuperados con independencia de su longitud. Dentro de una colección, los documentos presentan una diversidad de tamaño y por supuesto esto influye en el cálculo de los pesos de los términos, pues mientras más extenso sea el documento, más probabilidad existe que la ocurrencia de un término en ese documento sea mayor. Kolda (1998) expone dos razones principales por lo que se requiere el uso de la normalización de pesos:

- Frecuencias de términos muy altas: los documentos largos suelen utilizar los mismos términos en repetidas ocasiones y como resultado los factores término-frecuencia pueden ser grandes para documentos muy extensos.

- Número de términos: los documentos extensos tienen una gran cantidad de términos y esto aumenta las coincidencias entre una consulta y un documento largo, aumentando las posibilidades de recuperación de documentos largos en preferencia sobre los documentos más breves.

Una de las normalizaciones utilizadas en este aspecto es la "normalización coseno" la cual consiste en la división de cada uno de los pesos de los términos entre la distancia Euclidiana del vector, pues el vector normalizado tiene una distancia y su proyección sobre cualquier eje en el espacio de los documentos es el coseno del ángulo entre el vector y dicho eje (Greengrass, 2000).

$$N_j = \frac{1}{\sqrt{\sum (L_{ij} G_i)^2}} \quad (2.8)$$

Greengrass afirma que dicha normalización tiende a favorecer a documentos cortos, especialmente a aquellos que abordan solamente un tema en específico, esto suele suceder cuando los documentos extensos dentro de la colección abordan múltiples temas y sólo un tema responde a la consulta realizada. En este caso el factor de normalización penaliza en exceso a los pesos de los documentos grandes, otorgándoles una desventaja a la hora de realizar una búsqueda. Cuando esto sucede es necesario equilibrar el factor de normalización para aumentar la posibilidad de recuperar un documento de cierta longitud.

Por su parte la **normalización pivote** intenta corregir el problema anterior (Singhal et al., 1996a). El principio básico es corregir las diferencias dadas por la longitud de los documentos, entre la probabilidad de que un documento sea relevante y la probabilidad que sea recuperado dicho documento. Este método consta de dos pasos: primero se calcula el factor de normalización, por ejemplo mediante el coseno, y posteriormente se calcula el nuevo factor de normalización. Usando otro factor de normalización (coseno) un grupo de documentos es recuperado, y las curvas de probabilidad de recuperación y probabilidad de relevancia se representan en un eje de coordenada $(x; y)$ en función de las longitudes promedios de los documentos (véase figura 14). Se puede observar que ambas curvas se interceptan en un punto. Los documentos hacia el lado izquierdo de ese punto, en general tienen una mayor probabilidad de ser recuperados que de ser relevantes, y los documentos hacia el lado derecho tienen una mayor probabilidad de ser relevantes que de ser recuperados. Por ello es necesario rectificar el factor de normalización.

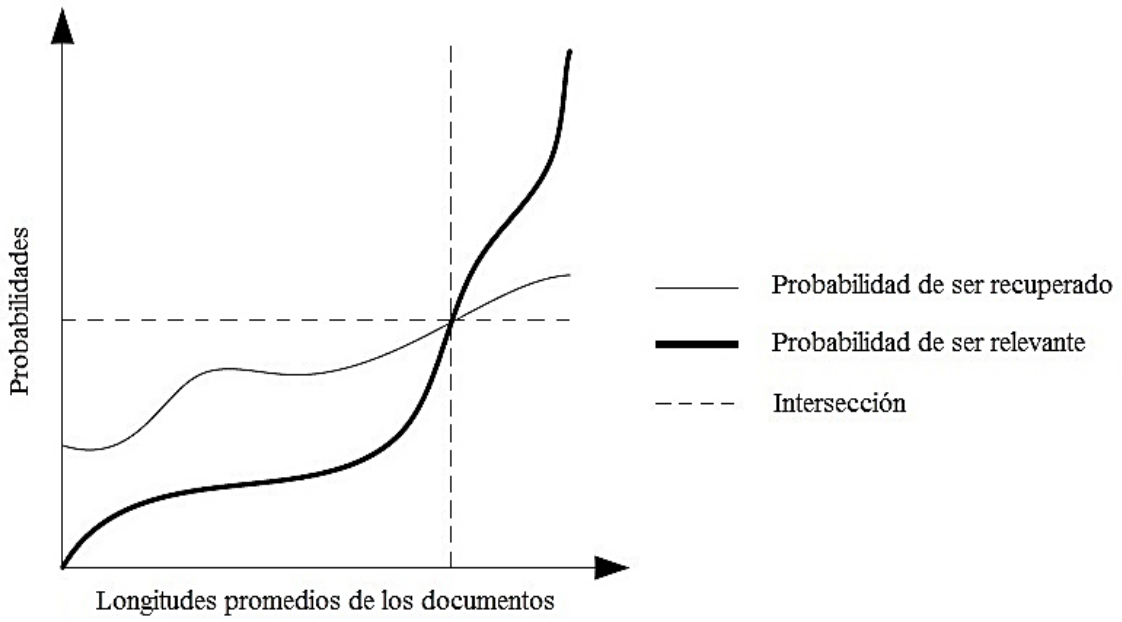

Figura 14: Probabilidad de relevancia y recuperación en función de las longitudes promedios de los documentos.

El factor de normalización es rotado e inclinado sobre ese punto, con el propósito de que aumente o descienda el mismo, para lograr una mejor combinación y equilibrio entre las probabilidades de relevancia y las de recuperación (Singhal et al., 1996b). Para llevar a cabo esta rectificación se representa por el eje x el factor de normalización original (coseno) y por el eje y el factor de normalización rectificado. Cabrera Diego (2012) explica todo el proceso partiendo de la ecuación principal de la recta (2.9) conformada por la pendiente (m) y el punto de intercepción en la ordenada (n). Esta ecuación se lleva a la situación actual, en función de N_1 que es la recta que representa la normalización original (recta identidad) y N_2 es la recta que representa la normalización rectificada (2.10). En este caso N_1 está formada por los valores de la normalización original (2.11).

$$y = mx + n \qquad (2.9)$$

$$N_2 = mN_1 + n \qquad (2.10)$$

$$N_1 = b \qquad (2.11)$$

La figura 15 muestra como quedarían ambas rectas representadas en un plano, las cuales se interceptan en el punto p denominado pivote.

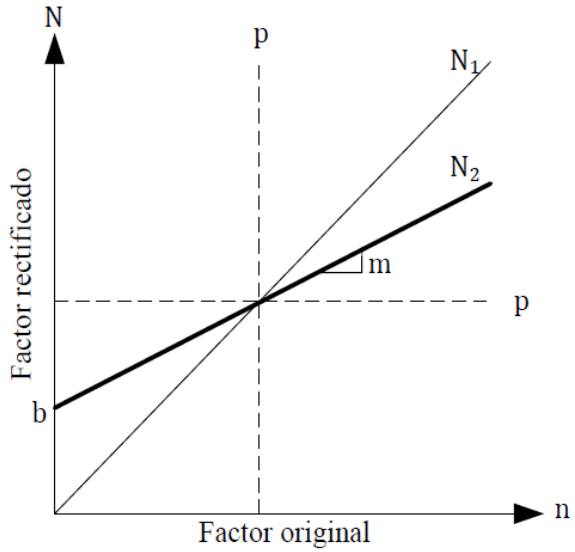

Figura 15: Representación gráfica de los factores de normalización. Fuente:Cabrera Diego (2012)

La ecuación de la recta N_2 tiene dos parámetros (m, n) que se desconocen en un inicio. Con respecto a m su valor se calcula empíricamente y debe encontrarse entre 0 y 1, pues si su valor es 0 sería una recta horizontal y normalizaría todos los documentos de la misma forma; y si su valor fuera 1 sería una recta paralela N_1 y no resolvería ningún problema pues se estarían normalizando los documentos igual que la forma anterior. En cuanto al valor de n, debe permitir que la recta N_2 cruce por el mismo punto por el cual N_1 cruza el pivote. Por tanto, es necesario sustituir la ecuación 2.11 en la ecuación 2.10 quedando de la siguiente forma:

$$N_2 = (m * b) + n \qquad (2.12)$$

Considerando que el punto donde se intersecan las dos rectas de normalización pertenece a la recta normalizada y tiene por coordenadas $(p; p)$, la ecuación (2.12) se podría plantear de la siguiente manera forma para $b = p$.

$$p = (m * p) + n \qquad (2.13)$$

Simplificando y despejando la ecuación (2.13) queda:

$$n = (1 - m)p \qquad (2.14)$$

Sustituyendo finalmente la ecuación (2.14) en la ecuación (2.12) se tiene:

$$N_2 = (m * b) + (1 - m)p \qquad (2.15)$$

El valor del pivote (p) no tiene una fórmula definida. Los autores del método (Singhal et al., 1996a) recomiendan como valor p, el promedio de todos los pesos del documento a normalizar; y como pendiente 0.2. Por lo tanto la normalización quedaría:

$$N_j = \frac{1}{m * b + (1 - m) * p} \quad (2.16)$$

2.3.4. Ponderaciones de pesos más utilizadas

Teniendo en cuenta la literatura consultada en el desarrollo de la investigación se tendrán en cuenta 3 funciones pesos propuestas. Según Salton y Buckley (1988) el método más utilizado de ponderación de término documento se obtiene con el producto de las funciones frecuencia del término (tf) y la frecuencia inversa de los documentos (idf) con una normalización coseno.

$$w_{ij} = \frac{tf * idf}{\sqrt{\sum (tf * idf)^2}} \quad (2.17)$$

Ambos autores también proponen la combinación de frecuencia local "aumentada" de los términos normalizados en conjunto con la frecuencia inversa de los documentos y normalizados por el coseno como el mejor esquema de ponderación de términos.

$$w_{ij} = \frac{0{,}5 + \left(0{,}5 * \left(\frac{tf_{ij}}{tfmax}\right)\right) * idf}{\sqrt{\sum \left(0{,}5 + \left(0{,}5 * \left(\frac{tf_{ij}}{tfmax}\right)\right) * idf\right)^2}} \quad (2.18)$$

Por otra parte, Araujo (2006) expone que se obtienen buenos resultados utilizando la combinación de peso local logaritmo y entropía como peso global, sin tener en cuenta la longitud de los documentos.

$$w_{ij} = \left(log(tf_{ij}) + 1\right) * \left(1 + \sum_{j=1}^{N} \frac{\frac{tf_{ij}}{F_i} \log \frac{tf_{ij}}{F_i}}{\log N}\right) \quad (2.19)$$

Sin embargo Singhal et al. (1996a) proponen la función 2.20, normalizando mediante la función pivote.

$$w_{ij} = \frac{tf * idf}{m * b + (1 - m) * p} \quad (2.20)$$

2.4. Descomposición en Valores Singulares

Una vez conformada la matriz A_{txd}, suponiendo que se desee nombrar así la matriz TxD, donde se asocian los términos y documentos; mediante la técnica matemática Descomposición en Valores Singulares o DVS se realiza la descomposición de esta matriz, a partir del cálculo de valores propios, valores singulares y vectores propios. Esta descomposición emplea directamente un gran número de contenidos básicos del Álgebra Lineal. Este método plantea que toda matriz *A* cuadrada o no, simétrica o no, puede ser descompuesta en el producto de estas 3 matrices (Mora, 2011):

$$A_{txd} = U_{txn} S_{nxn} V_{dxn}^{T} \quad (2.21),$$

donde las columnas de la matriz U están compuesta por los vectores singulares ortonormales de AA^T que representa el espacio de los términos, la matriz S es una matriz diagonal $(n\ x\ n)$ donde los valores de la diagonal están compuesta por los valores singulares de A ordenados descendentemente, y por último V es una matriz ortogonal compuesta por los vectores singulares ortonormales de $A^T A$ que representan al espacio de los documentos (figura 16).

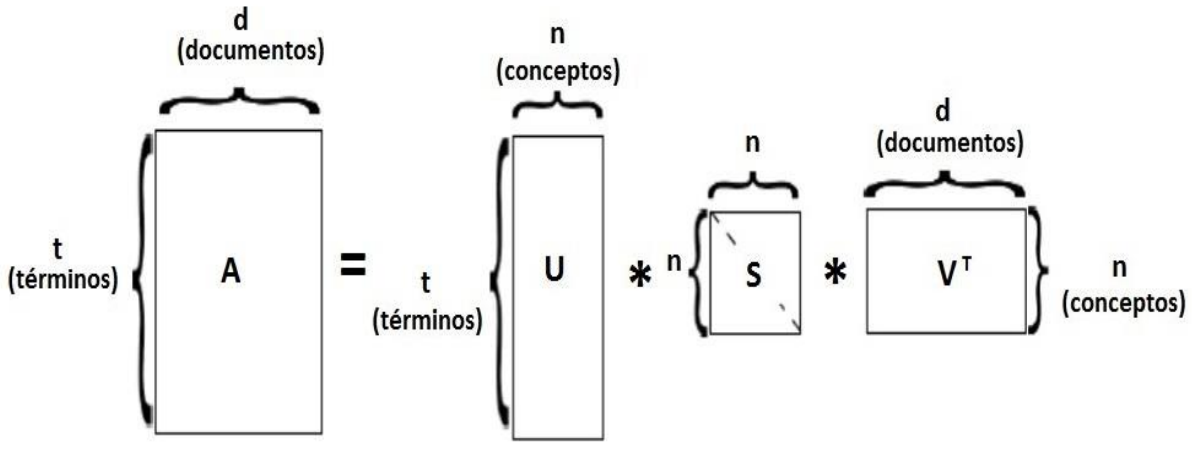

Figura 16: Descomposición en Valores Singulares.

Para la obtención de estas 3 matrices es necesario conocer cómo hallar los valores singulares de una matriz. Según la definición de Kolman y Hill (2006): "Si A es una matriz $m\ x\ n$, los valores singulares (σ) de *A* son las raíces cuadradas de los autovalores o valores propios (λ) de $A^T A$ ". Los

valores propios de una matriz se calculan resolviendo el polinomio originado por la siguiente ecuación:

$$\lambda = \det(X - \lambda I) = 0 \quad (2.22),$$

donde X, en este caso es la matriz resultante de $A^T A$ y la matriz I es la matriz identidad. Una vez calculados los valores propios y siguiendo la definición anterior se obtienen los valores singulares de A:

$$\sigma_i = \sqrt{\lambda_i} \quad (2.23).$$

De esta forma se puede conformar la matriz S, ubicando estos valores singulares de A en su diagonal de manera descendente.

Por otra parte las columnas de la matriz V se forman a partir de los vectores singulares ortonormales. Una vez calculados los valores propios, se calculan los vectores propios dándole solución a la siguiente ecuación:

$$(X - \lambda I)v = 0 \quad (2.24).$$

Estos vectores propios son divididos por su norma, en este caso la norma Euclidiana.

$$\| X\| = \sqrt{\sum X_i^2} \quad (2.25).$$

De esa manera queda conformada la matriz V, con cada uno de los vectores v_i. En este caso se calcula su transpuesta, cuya matriz resultante será la utilizada en esta descomposición.

Para calcular los vectores singulares ortonormales que conformar la matriz U se utiliza la siguiente expresión:

$$v_i = \frac{Av_i}{\sigma_i} \quad (2.26).$$

Finalmente, quedan conformadas las tres matrices cuyo producto reconstruye la matriz original.

2.5. Reducción de las dimensiones del espacio: valor *k*

Una vez realizada la descomposición de la matriz en el producto de las matrices U, S y V^T, es necesario la reducción de la dimensión para eliminar el ruido existente en la matriz, es decir los elementos que no brindan información discriminante. Para esto se transforman estas matrices a un

espacio de k dimensiones (véase figura 17) obteniendo U_k, S_k y V_k^T, para conformar así una matriz que representa la asociatividad término-documento implícita en la matriz original, es decir, la semántica latente implícita, garantizando eliminar el ruido presente en esta.

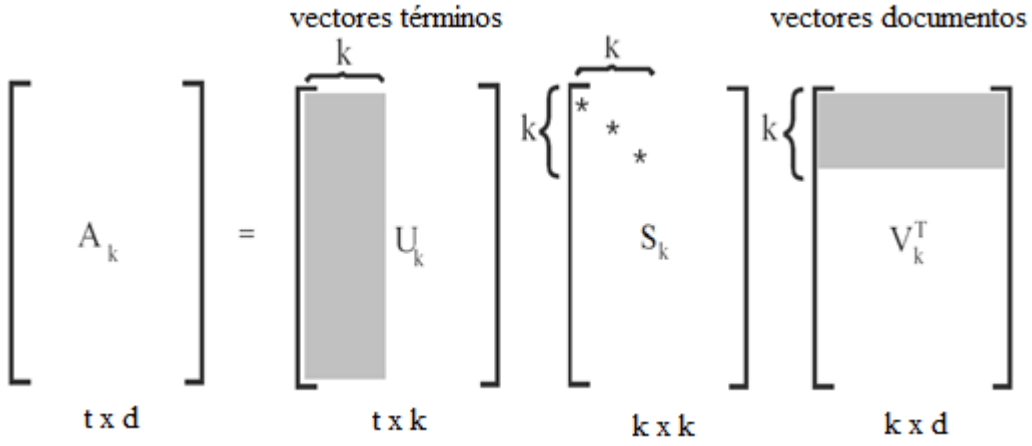

Figura 17: Reducción de las matrices a k dimensiones.

Es importante señalar que el valor de *k* aún es un problema interesante que no se encuentra bien definido, pues es necesario seleccionar un valor lo suficientemente pequeño como para no modelar el ruido comprendido en la matriz original, pero a su vez debe ser lo suficientemente grande para no dejar fuera información relevante contenida en la matriz original (Saracevic et al., 1997). Mientras una reducción en *k* dimensiones puede eliminar gran parte del ruido, conservar muy pocas dimensiones puede traer consigo la pérdida de información importante. Experimentos realizados por Deerwester et al. (1990) usando una colección de prueba arrojaron resultados favorables para valores de *k* menores que 100. Afirmando los criterios anteriormente mencionados Grossman y Frieder (2012) expresan que el valor de *k* es el número de columnas que permanecen después de la DVS y es determinado vía experimentación.

2.6. Análisis del vector consulta

Al igual que un documento, la consulta es un conjunto de palabras o términos, y su representación en dicho espacio es determinada mediante el cálculo del vector. Este vector se construye a partir de los pesos de los términos y se procesa como si fuera un documento más. La consulta de un usuario debe ser ubicada en el espacio k-dimensional, utilizando la siguiente función:

$$q = q^T U_k S_k^{-1} \quad (2.27)$$

De esta forma el vector de la consulta puede ser comparado con todos los documentos existentes y establecer una puntuación (*ranking*) por su similaridad con la misma.

2.7. Similitud coseno

Para el cálculo de la similitud entre cada uno de los vectores que representan los documentos (V_k^T) y el vector consulta (q), se utiliza la similitud coseno. En este caso el vector consulta es definido como $q = (c_{1q}, c_{2q}, \dots, c_{kq})$ donde c representa cada una de las coordenadas del vector en el espacio de *k* dimensiones. El vector del documento es representado como $d_j = (v_{1j}, v_{2j}, \dots, v_{kj})$. El modelo vectorial evalúa el grado de similitud del documento d_j con relación a la consulta q como la correlación entre los vectores d_j y q (Zhao y Grosky, 2002).

$$Sim(d_j, q) = \frac{(\sum_{i=1}^{t} v_{ij} * c_{iq})}{\sqrt{\sum_{i=1}^{t} v_{ij}^2} * \sqrt{\sum_{i=1}^{t} c_{iq}^2}} \quad (2.28)$$

2.8. Tecnologías y herramientas utilizadas para la implementación

En este epígrafe se caracterizan las tecnologías que se utilizaron en la implementación de la aplicación. Se describen las herramientas y librerías que permitieron un mejor procesamiento de los documentos y el trabajo con las distintas funciones del Álgebra Lineal.

2.8.1. Lenguaje de programación Java

La implementación se realizó utilizando el lenguaje de programación Java[4]. Éste es un lenguaje orientado a objetos de propósito general mediante el cual se puede construir cualquier tipo de proyecto. Java también posee mecanismos para garantizar la seguridad durante la ejecución comprobando, antes de ejecutar código, que este no viola ninguna restricción de seguridad del sistema donde se va a ejecutar. De forma general es un lenguaje simple, tipificado estáticamente, compilado, independiente de la arquitectura, multi-procesos, robusto, seguro y ampliable (Beitzel, 2006). Dentro de las clases implementadas en Java se utilizaron *StringTokenizer*, la cual permitió la manipulación del texto de los documentos permitiendo almacenarlo en una lista de términos independiantes. Para la declaración de las listas se utilizó la clase *ArrayList*, y las clases

[4] *https://www.**java**.com*

BufferedReader y *FileReader* permitieron cargar tanto la información de los documentos como de las consultas, las cuales estaban en diferentes ficheros de texto.

2.8.2. Entorno de desarrollo NetBeans

Como entorno de desarrollo integrado (IDE) se utilizó NetBeans[5] el cual es multilenguaje completo y modular con soporte para *Java Standard Edition*, *Java Enterprise Edition* y *Java Micro Edition*. Es una plataforma gratis y *open source* que permite construir aplicaciones completas para el cliente, ya sea de escritorio, web, móvil o de empresa. Dentro de sus principales características se puede mencionar que presenta mejoras en el editor de código, la instalación y actualización es simple, presenta características visuales para el desarrollo web (Korfhage, 1997).

2.8.3. Librería JAMA

JAMA es un paquete básico de Álgebra Lineal para Java. Ofrece un conjunto de clases para construir y manipular matrices reales. Tiene el propósito de proporcionar funcionalidades suficientes para problemas rutinarios en una forma natural y comprensible. JAMA cuenta con 5 clases fundamentales que permiten realizar diversas operaciones con matrices.

La clase principal *Matrix* ofrece las operaciones fundamentales del álgebra lineal numérica. Cuenta con varios constructores que permiten crear matrices bidimensionales. Varios métodos que facilitan el acceso a matrices secundarias y elementos de la matriz. Incluye algunas funciones que permiten aplicar la aritmética básica, incluyendo la suma y multiplicación de matrices, las normas de la matriz, cálculo de determinantes, inversas y otras funciones matriciales (Korfhage, 2008).

2.8.4. Snowball

Snowball es un pequeño lenguaje de procesamiento de cadenas (*strings*) que permite implementar fácilmente algoritmos de *stemming*. Se puede generar código en ANSI, C y Java. Las páginas de Snowball contienen *stemmers* para 12 idiomas (incluido el español, catalán, euskera y el inglés). Todas las explicaciones, sin embargo, son dadas en inglés. (Snowball, 2014). La colección de documentos utilizada en esta investigación se encuentra en idioma ingles por lo que la clase *englishStemmer* fue la utilizada para el análisis de los términos.

[5] *https://**netbeans**.org*

2.9. Conclusiones parciales

Se caracterizó detalladamente cada una de las etapas por las que estará guiado el proceso de recuperación de información, haciendo énfasis en las correspondientes a la ISL, siendo esenciales la construcción de la matriz TxD y la determinación del valor de *k* para la reducción de las dimensiones. Con respecto a la matriz TxD, una vez analizadas las distintas funciones de pesos propuesta en la literatura, se decidió evaluar sus comportamientos e influencias en la obtención de resultados relevantes. En cuanto a la reducción del espacio a *k* dimensiones se dispuso utilizar en el experimento valores entre 10 y 130; y evaluar en cuanto a precisión y cobertura los resultados obtenidos. Por otra parte se definieron las herramientas y tecnologías que se utilizarán en la implementación de la aplicación informática, especificando en cada caso los componentes principales de cada una y su utilización y beneficios en el desarrollo de la misma.

CAPÍTULO 3: RESULTADOS Y DISCUSIÓN

Utilizando las herramientas y tecnologías mencionadas en el capítulo anterior, se desarrolló un sistema de recuperación de información utilizando la técnica ISL como extensión del modelo vectorial. La aplicación desarrollada permite recuperar un grupo de documentos los cuales son relevantes en su mayoría a esa consulta, teniendo en cuenta los juicios de relevancia establecidos para la colección utilizada.

En este capítulo se muestran los resultados de experimentos utilizando la colección MEDLINE, con el objetivo de determinar cuáles combinaciones de las funciones de peso y valores del parámetro *k* brindan mejores resultados. Para la experimentación se utilizó un ordenador con las siguientes prestaciones: microprocesador Core i3, 4GB de memoria RAM y 1TB de almacenamiento. Para evaluar los resultados obtenidos se calcularon los valores de precisión y cobertura para cada consulta.

3.1. Aplicación informática implementada

El software implementado abarca todas las etapas del proceso de recuperación de información representado en la figura 10 que se encuentra en el capítulo 2. Para esto se implementaron todas las funciones y métodos necesarios, los cuales permitieron realizar los experimentos pertinentes con la finalidad de determinar, empíricamente, el esquema de ponderación de términos y el valor del parámetro *k* más adecuado, es decir, que permita obtener un gran número de documentos relevantes. La figura 18 muestra la estructura básica de la aplicación, donde se pueden observar las clases implementadas.

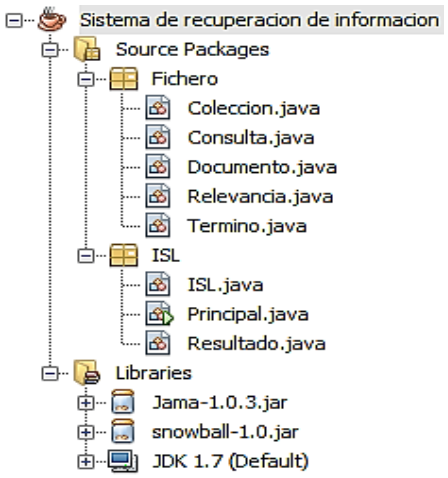

Figura 18: Clases implementadas en el lenguaje Java

A cada una de las consultas se le aplicaron también estas técnicas en el momento de su utilización. Para la construcción de la matriz TxD (6464 x 550), se implementaron algunas de las funciones locales, globales y de normalización, con la intención de valorar los resultados obtenidos. La figura 19 muestra el código de la función de la entropía, una de las funciones de peso global que se tuvieron en cuenta.

```
public Matrix G_Entropia() {
    double tf, F, entropia, sumatoria;
    int fila = Matriz_inicial.getRowDimension();

    Matrix Matriz_temp = new Matrix(fila, 1);
    for (int i = 0; i < Matriz_inicial.getRowDimension(); i++) {
        sumatoria = 0;
        F = 0;
        for (int h = 0; h < Matriz_inicial.getColumnDimension(); h++) {
            F = F + Matriz_inicial.get(i, h);
        }
        for (int j = 0; j < Matriz_inicial.getColumnDimension(); j++) {
            tf = Matriz_inicial.get(i, j);
            if (tf != 0) {
                sumatoria = sumatoria + ((tf / F) * Math.log10(tf / F)) / Math.log10(D);
            }
        }
        entropia = 1 + sumatoria;
        Matriz_temp.set(i, 0, entropia);
    }
    return Matriz_temp;
}
```

Figura 19: Función de peso global: entropía, implementada en el lenguaje Java

Para la obtención de los resultados se tuvieron en cuenta solo aquellos que su distancia con respecto al vector consulta fuera superior a 0,5; pues mientras más próximo a 1 sea este valor, más similitud existirá en entre la consulta y el documento. En las próximas secciones se explicarán los detalles de la experimentación realizada y sus resultados.

3.2. Normalización de la longitud de los documentos

Para la construcción de la matriz TxD se tuvieron en cuenta diversos criterios de los autores mencionados en el epígrafe 2.3. Primeramente, se realizó un análisis sobre la influencia de la longitud del documento en la obtención de resultados relevantes en la colección y en función de

esto se definió para cada documento su longitud, dada por la cantidad de palabras contenidas en cada uno. Una vez obtenidos estos valores (véase anexo 2) se pudo apreciar que existía una gran variabilidad en las longitudes de los documentos por lo que fue necesario analizar cuál de las dos normalizaciones (coseno o pivote) era más factible utilizar en el desarrollo del sistema.

Con este objetivo, la colección de documentos se ordenó de manera ascendente teniendo en cuenta la longitud y se dividió en subgrupos de 10 documentos, obteniéndose un total de 55 subgrupos. Para cada uno de ellos se calculó la media de la longitud de los documentos contenidos en él (véase anexo 3). Para construir la matriz TxD se tuvo en cuenta como esquema de ponderación de términos la frecuencia absoluta (*tf*) como peso local; como función global la frecuencia inversa del documento (*idf*), y la normalización coseno. Se seleccionaron 100 documentos del total de recuperados, por cada consulta, y se agruparon en un total de 1900 pares *consulta-documento*. Partiendo de estos resultados se calculó la probabilidad de que un documento D, que pertenece a un subgrupo dado, fuese recuperado:

$$P(D \in subgrupo \,|D\ sea\ recuperado)$$

Posteriormente, de estos 1900 pares de *consulta-documento* se seleccionaron los que son relevantes, quedando solamente 196 pares *consulta-documento* y se calculó la probabilidad que un documento D, que pertenezca a un subgrupo fuese relevante:

$$P(D \in subgrupo \,|D\ sea\ relevante)$$

La figura 20 muestra los resultados de ambas probabilidades, la de recuperación y la de relevancia, en función de las longitudes promedios de los documentos. Se puede apreciar el efecto negativo de la normalización coseno en la recuperación de documentos relevantes de gran longitud. Mediante esta ponderación de términos se tiende a recuperar documentos cortos de menor relevancia, mientras que los documentos de mayor longitud y en ocasiones de mayor relevancia que los documentos cortos, no son recuperados. Por esta cuestión se utiliza en la implementación del sistema desarrollado en la tesis, la Normalización Pivote (*NP*) para reducir este efecto negativo de la normalización coseno.

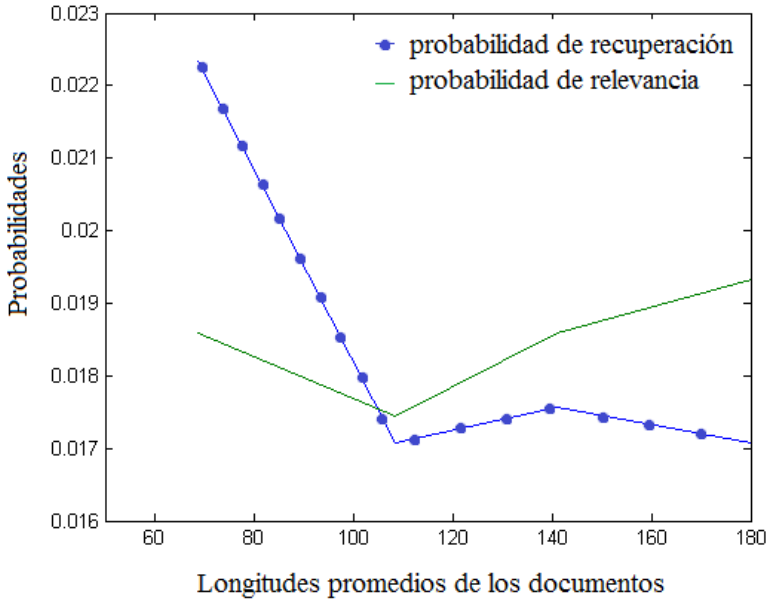

Figura 20: Probabilidad de recuperación y de relevancia con respecto a la longitud de los documentos.

3.3. Ponderación de términos y reducción de las dimensiones

Como se había explicado anteriormente en el epígrafe 2.5, el valor de *k* se calcula empíricamente. Por tales motivos se realizó un experimento con diversos valores de *k*, con un incremento de 10, hasta un valor máximo de k=130, pues como se había expresado, este valor permitirá que no se modele el ruido comprendido en la matriz original, pero a su vez que no se deje fuera información relevante contenida en la misma. Para cada valor de *k* se procesaron las 19 consultas. Los documentos recuperados para cada una de las consultas fueron analizados y se cuantificaron los que realmente eran relevantes para cada una, según los juicios de relevancia de la colección. Para un inicio se tuvo en cuenta la frecuencia absoluta del término (*tf*) como valor de peso local, la frecuencia inversa del documento (*idf*) como peso global y la normalización pivote. En el anexo 4 se encuentran los resultados alcanzados utilizando este esquema.

En los resultados se pueden apreciar para cada consulta (C) el total de documentos relevantes (D Rel) existentes en la colección; y utilizando distintos valores de *k* se observa la cantidad de documentos recuperados (D Rec) y la cantidad de documentos relevantes recuperados (DR Rec). Con el propósito de conocer cuál valor de *k* aporta mejores resultaros, se calcularon las medias de estos valores. La figura 21 muestra cómo a medida que se tiene en cuenta un mayor número de

dimensiones, existe una notable disminución en la cantidad de documentos recuperados y en el número de documentos relevantes recuperados.

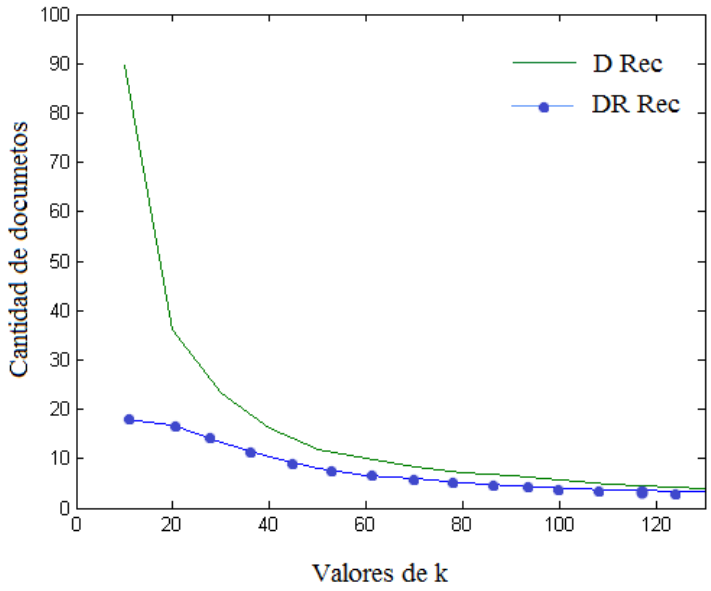

Figura 21: Documentos recuperados a partir de diversos valores de *k* (*tf* * *idf* * *NP*)

A pesar de que para el 52% de las consultas se obtuvieron todos los documentos relevantes utilizando k = 10, este valor de dimensión no es el más óptimo, pues aproximadamente el 80 % de los documentos recuperados no son relevantes. Sin embargo, tomando valores de k = 50 o superiores, el total de documentos que no son relevantes disminuye progresivamente tomando valores por debajo del 42%, pero a su vez para cada consulta se recuperan un mínimo de documentos relevantes. Tal es así que para k = 130, solo se obtiene el 18% de los documentos relevantes aproximadamente.

Al existir, según la literatura consultada, una variedad de funciones para calcular los pesos de los términos y con el propósito de verificar cuánto afectaba en la obtención de los resultados; se repitió este mismo procedimiento utilizando otro esquema de ponderación de términos, manteniendo siempre la normalización pivote. Para el segundo experimento se aplicó la función logarítmica (*log*) para calcular el peso local, y la entropía como función de peso global. Los resultados se encuentran en el anexo 5 y son representados gráficamente en la figura 22. Al analizar los resultados obtenidos anteriormente, utilizando estos 2 esquemas de ponderación, se puede observar que existe una similitud entre ellos; obteniéndose mayor cantidad de documentos relevantes en las dimensiones más pequeñas. Pero no solo es importante que se recuperen la mayor cantidad de

documentos relevantes, sino que del conjunto de documentos recuperados la mayoría sean relevantes. Por tales motivos se decidió realizar un análisis más profundo evaluando los resultados alcanzados en la recuperación.

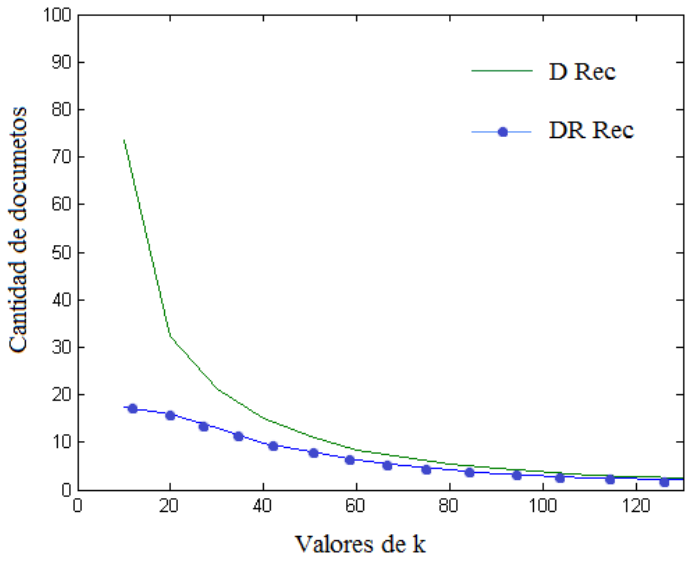

Figura 22: Documentos recuperados a partir de diversos valores de *k* (*log* * *entropía* * *NP*)

3.4. Evaluación de la recuperación de información

A partir de la obtención de los documentos relevantes recuperados para cada consulta, teniendo en cuenta las medidas de precisión (P) y cobertura (C) abordadas en el subepígrafe 1.2.2, se evalúa de manera práctica la aplicación propuesta. Para ello se tienen en cuenta los resultados obtenidos a partir de la utilización de ambos esquemas de ponderación de términos en el epígrafe anterior. Se prosigue calculando estas medidas evaluadoras para cada una de las consultas y se calcula la media de cada medida para cada dimensión.

En el anexo 6 se encuentran los resultados obtenidos a partir de los documentos recuperados utilizando el primer esquema de ponderación de términos. En la figura 23 se representa el comportamiento de la precisión y la cobertura en función de las distintas dimensiones. Como se puede observar, a pesar de que la precisión se mantiene estable entre un 60% y 75% para valores mayores que k = 30 aproximadamente, esto no significa que se obtengan buenos resultados para estas dimensiones, pues resulta todo lo contrario debido a que los valores de la cobertura disminuyen en gran medida.

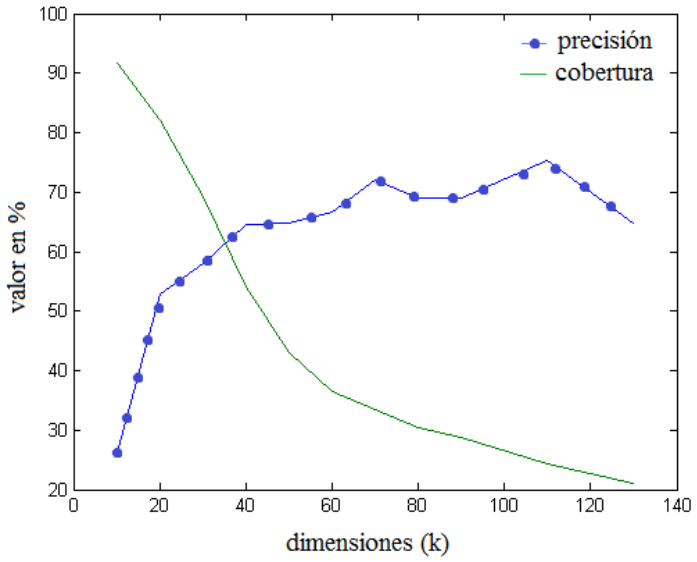

Figura 23: Precisión y cobertura a partir de diversos valores de *k* (esquema*: tf * idf * NP*)

Ambas medidas se deben compensar, debe existir un balance en sus valores y en este caso a medida que aumentan el número de dimensiones decrece notablemente el valor de la cobertura lo cual significa que se recuperaran cada vez menos documentos y se dejan de recuperar en gran medida documentos relevantes. Esto justifica los valores de precisión, pues existe mayor probabilidad de que en ese mínimo grupo de documentos recuperados, casi todos sean relevantes. Para este esquema los mejores resultados se obtienen en la dimensión k = 38 aproximadamente, donde la precisión y la cobertura se interceptan y toman valores alrededor del 62 %.

En el anexo 7 se encuentran los resultados obtenidos a partir de los documentos recuperados utilizando el segundo esquema de ponderación de términos. La figura 23 representa el comportamiento de ambas medidas para este esquema. Como se puede apreciar, similar al análisis anterior, existe una tendencia hacia la pérdida de documentos en la recuperación a medida que aumentan las dimensiones. En este caso los mejores resultados se obtienen en el punto donde se interceptan la precisión y la cobertura, aproximadamente para dimensiones entre 30 y 40. En este punto el valor de estas medidas oscilan alrededor del 64%.

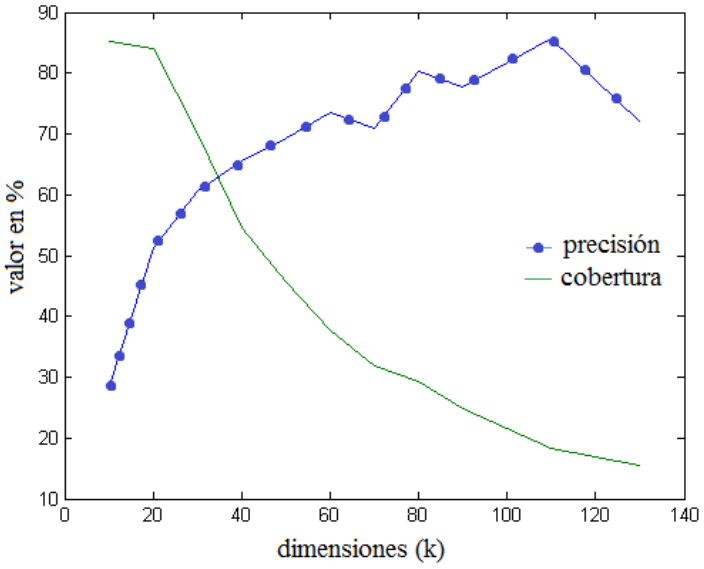

Figura 24: Precisión y cobertura a partir de diversos valores de *k* (esquema: *log * entropía * NP*)

Ambos esquemas de ponderación, a pesar que están compuestos por distintas funciones de peso local y global, reflejan resultados similares, recomendándose de esta manera utilizar dimensiones con valores entre 20 y 40 aproximadamente. Teniendo en cuenta solo estas dimensiones se recuperan resultados que serán aceptados satisfactoriamente por los usuarios.

3.5. Conclusiones parciales

Se implementó una aplicación informática donde se tuvieron en cuenta algunas de las funciones más utilizadas, de peso local, global y de normalización de la longitud del documento. Esta aplicación permitió realizar una experimentación con el objetivo de verificar dos aspectos fundamentales: la influencia de la longitud de los documentos y la influencia de la reducción de las dimensiones del espacio vectorial, según el método ISL, en la obtención de resultados relevantes. Se comprobó que al existir una variabilidad en las longitudes de los documentos no es recomendable utilizar la normalización coseno pues mediante ella se dejan de recuperar documentos de mayor longitud y a su vez de mayor relevancia, dado que son penalizados con respecto a los de menor longitud. Por otra parte, se deben tener en cuenta las dimensiones que tomen valores aproximadamente entre 20 y 40, las cuales permite obtener valores elevados de precisión y cobertura.

CONCLUSIONES

El desarrollo de esta investigación permitió caracterizar los principales modelos matemáticos existentes y analizar algunas de sus extensiones. Se definió la utilización de la Indexación Semántica Latente dado a que no sólo intenta resolver los problemas de polisemia y sinonimia en el momento de efectuar una consulta, sino que utiliza la semántica implícita en las relaciones entre los términos y los documentos. Se identificó cada una de sus etapas, destacándose la normalización de la longitud del documento como un elemento crucial en la construcción de la matriz TxD. Se demostró que la normalización coseno tiende a penalizar demasiado los documentos extensos, favoreciendo la recuperación de los documentos más cortos dentro de la colección. Por tales motivos es recomendable la utilización de la normalización pivote en colecciones donde las longitudes de los documentos varíen en gran medida, pues equilibra la probabilidad de que un documento sea relevante con la probabilidad de que sea recuperado.

Por otra parte de se implementó una aplicación que abarca cada una de las etapas del proceso de recuperación de información. Este software contempla las funciones de ponderación de términos más utilizadas según los criterios de autores consultados. Además, permitió la realización de la experimentación mediante la cual se determinó el rango de los valores aproximados de la dimensión del subespacio vectorial, definido como valor *k*. Según las observaciones se obtienen buenos resultados con dimensiones entre 20 y 40, pues a partir de este último valor se comienza a menguar el poder de consulta, es decir, se dejan de recuperar de manera progresiva documentos relevantes.

También se evaluó la aplicación desarrollada teniendo en cuenta las métricas de precisión y cobertura. Para dimensiones pequeñas, inferiores que 20, se obtienen elevados porcientos de cobertura pero muy poca precisión pues se recupera un gran volumen de documentos que no son relevantes. Sin embargo para dimensiones superiores que 40, se obtienen elevados valores de precisión pero muy bajo valores de cobertura lo que indica que se están dejando de recuperar una gran cantidad de documentos relevantes. Ambas medidas se compensan y toman valores alrededor del 60% cuando las dimensiones se encuentran entre 20 y 40.

De esta forma se cumplieron los objetivos planteados en el proyecto de esta tesis. Se podrá utilizar la solución informática en nuestro contexto universitario.

RECOMENDACIONES

Continuar el trabajo realizado mediante un análisis más profundo de la heurística de la Indexación Semántica Latente. Definir un procedimiento que permita construir una colección entrenante para el Repositorio de Objetos de Aprendizaje de la UNAH y la biblioteca digital. También es interesante investigar sobre qué criterios se deben utilizar para la creación de una colección representativa de prueba, y si existen propiedades particulares de documentos, consultas o juicios de relevancia que podrían requerir un tipo diferente de normalización. Por otra parte sería conveniente definir cómo se pueden conformar los juicios de relevancia necesarios para una colección de prueba teniendo en cuenta los criterios de usuarios y expertos.

REFERENCIAS BIBLIOGRÁFICAS

ADELL, J. & CASTAÑEDA, L. 2010. Los Entornos Personales de Aprendizaje (PLEs): una nueva manera de entender el aprendizaje. *In:* ROIG VILA, R. & FIORUCCI, M. (eds.) *Claves para la investigación en innovación y calidad educativas. La integración de las Tecnologías de la Información y la Comunicación y la Interculturalidad en las aulas.* Alcoy: Marfil.

ALIMOHAMMADI, D. 2010. *Mathematics for Classical Information Retrieval: Roots and Applications*, Lulu. com.

ARAUJO, B. S. 2006. Aprendizaje Automático: conceptos básicos y avanzados. *Aspectos prácticos utilizando el software Weka. ISBN,* 10**,** 84-8322.

ARCHANA, R. & RAVICHANDRAN, M. 2014. Classification of documents in E-Learning using Multidimensional Latent Semantic Analysis. *International Journal of Innovative Research in Science, Engineering and Technology,* 3**,** 1304-1308.

BAEZA-YATES, R. & RIBEIRO-NETO, B. 1999. *Modern information retrieval*, ACM press New York.

BARKOV, A. 2009. *mnoGoSearch 3.3.9 reference manual* [Online]. Available: <http://www.mnogosearch.org/doc33/msearch-intro.html#features>. [Accessed febrero 2015].

BEITZEL, S. M. 2006. *On understanding and classifying web queries.* Tesis Doctoral, Citeseer.

BERNERS-LEE, T. & HENDLER, J. 2001. Publishing on the semantic web. *Nature,* 410**,** 1023-1024.

BLANCHARD, A. 2007. Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information,* 29**,** 308-316.

BOTANA, G. D. J. 2010. *La técnica del Análisis de la Semántica Latente (LSA/LSI) como modelo informático de la comprensión del texto y el discurso.* Tesis doctoral, Universidad Autónoma de Madrid.

BRADFORD, R. 2009. Why LSI? Latent Semantic Indexing and Information Retrieval. *Content Analyst Company*.

CABRERA DIEGO, L. A. 2012. *TF-IDF para la obtención automática de términos y su validación mediante Wikipedia.* Tesis de Grado, Universidad Nacional Autónoma de México.

CACHEDA, M. 2008. Introduction to the Classic Models of Information Retrieval. *Revista General de Información y Documentación,* 18**,** 365-374.

CAICEDO, C. H. 2012. Virtualización organizacional, Web semántica y redes sociales. *Visión Electrónica,* 6**,** 134 - 159.

CALABUIG, J. M., RAFFI, L. M. G. & SÁNCHEZ-PEREZ, E. A. 2015. Álgebra lineal y descomposición en valores singulares. *Modelling in Science Education and Learning,* 8**,** 133-144.

CHISHOLM, E. & KOLDA, T. G. 1999. New Term Weighting Formulas for the Vector Space Method in Information Retrieval. *Computer Science and Mathematics Division*.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science,* 41**,** 391.

DOMINICH, S. 2000. A unified mathematical definition of classical information retrieval. *Journal of the American Society for Information Science,* 51**,** 614-624.

DUMAIS, S. 1992. Enhancing performance in latent semantic indexing (LSI) retrieval. Technical.

DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K., DEERWESTER, S. & HARSHMAN, R. Using Latent Semantic Analysis to improve access to textual information. The SIGCHI conference on Human factors in Computing Systems, 1988. 281-285.

GALLARDO, P. F. 2006. El secreto de Google y el álgebra lineal. *Matemáticas en la ciencia y la cultura contemporáneas,* 6**,** 36.

GARCÍA, O. C. 2013. *Análisis de Redes y búsqueda en la Web: PageRank* [Online]. Universidad de Granada: Soft Computing and Intelligent Information Systems. Available: http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/RedesSistemasCompejos/Seminario02-PageRank.pdf [Accessed octubre 2015].

GELBUKH, A. & SIDOROV, G. 2010. *Procesamiento automático del español con enfoque en recursos léxicos grandes,* México, Instituto Politécnico Nacional.

GRACIA, J.-M. 2002. Algebra Lineal tras los buscadores de Internet. *John Wiley & Sons, Inc. New York*.

GRANMATV 2015. El Google cubano. *Crisol de la Nacionalidad Cubana TV Granma,* 2015.

GREENGRASS, E. 2000. Information Retrieval: A Survey.

GROSSMAN, D. A. & FRIEDER, O. 2012. *Information retrieval: Algorithms and heuristics*, Springer Science & Business Media.

GROUP-IR. 2016. *The Information Retrieval Group* [Online]. Inglaterra: Universidad de Glasgow Available: http://ir.dcs.gla.ac.uk/ [Accessed diciembre 2015].

HARIKUMAR, S. & SRINIVASAN, M. 2015. Google's PageRank Algorithm : An analysis, implementation and relevance today. *Department of Electrical Engineering. Indian Institute of Technology Madras* [Online]. Available: http://www.ee.iitm.ac.in/~ee11b075/files/dsa-termpaper.pdf.

JONES, K. S. 1997. *Readings in information retrieval*, San Francisco, Morgan Kaufmann.

KIELA, D. & CLARK, S. A systematic study of semantic vector space model parameters. Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL, 2014. 21-30.

KOLDA, T. G. 1998. Limited-Memory Matrix Methods with Applications.

KOLMAN, B. & HILL, D. R. 2006. *Álgebra lineal*, Pearson Educación.

KORFHAGE, R. R. 1997. *Information storage and retrieval*, John Wiley & Sons, Inc. New York.

KORFHAGE, R. R. 2008. Information storage and retrieval.

KUMAR, C. A. & SRINIVAS, S. 2009. A Note on the Effect of Term Weighting on Selecting Intrinsic Dimensionality of Data. *Cybernetics and Information Technologies*, 9**,** 5-12.

LINDBERG, D. 2000. Internet access to the National Library of Medicine. *Effective clinical practice: ECP*, 3**,** 256.

MANNING, C. D., RAGHAVAN, P. & SCHÜTZE, H. 2008. *Introduction to information retrieval*, Cambridge university press Cambridge.

MORA, H. M. 2011. SVD: descomposición en valores singulares: resultados, aplicaciones, cálculo.

NASSAR, M. O., KANAAN, G. & AWAD, H. A. H. 2010. Comparison between Different Global Weighting Schemes. *Proceeding of the International MultiConference of Engineers and Computer Scientist*, 1.

NIETO, A. M. & RUEDA, E. R. 2010. ORION: Un motor de búsquedas para la Web de la UCI.

PORTER, M. F. 1980. An algorithm for suffix stripping. *Program*, 14**,** 130-137.

REINA, D. G., MARTINEZ-TORRES, M. R., TORAL, S. L. & BARRERO, F. 2013. Análisis de la oferta educativa en el ámbito de los MOOCs.

SALTON, G. & BUCKLEY, C. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management,* 24**,** 513-523.

SALTON, G., FOX, E. A. & WU, H. 1983. Extended Boolean information retrieval. *Communications of the ACM,* 26**,** 1022-1036.

SÁNCHEZ DÍAZ, A., NUÑEZ GONZÁLEZ, A., ROQUE ALAYÓN, Y., LÓPEZ PADRÓN, A., FERNÁNDEZ DE CASTRO, A. & JORGE DE MOURA, D. 2015. Recuperación de información en escenarios de aprendizaje personalizado para las carreras de perfil agropecuario. *Revista Ciencias Técnicas Agropecuarias,* 24**,** 63-70.

SARACEVIC, T., KANTOR, P., CHAMIS, A. Y. & TRIVISON, D. 1997. A Study of Information Seeking and Retrieving: 1. Background and Methodology. *Readings in Information Retrieval. San Francisco: Morgan Kaufmann***,** 175-190.

SARACEVIC, T., KANTOR, P. B., CHAMIS, A. Y. & TRIVISON, D. 1988. A study of information seeking and retrieving: Part 1, background and methodology. *Journal of the American Society for Information Science,* 39**,** 161-176.

SECO, D. 2009. *Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales.* Tesis Doctoral, Universidad A Coruña.

SINGH, V. K. & SINGH, V. K. 2015. Vector Space Model: An Information Retrieval System. *International Journal of Advanced Engineering Research and Studies,* 4**,** 141-143.

SINGHAL, A., BUCKLEY, C. & MITRA, M. Pivoted document length normalization. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 1996a. ACM, 21-29.

SINGHAL, A., SALTON, G., MITRA, M. & BUCKLEY, C. 1996b. Document length normalization. *Information Processing & Management,* 32**,** 619-633.

SNOWBALL. 2014. *Snowball* [Online]. Available: http://snowball.tartarus.org/ [Accessed noviembre 2015].

SWANSON, D. R. 1988. Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science,* 39**,** 92-98.

TURTLE, H. & CROFT, W. B. Inference networks for document retrieval. Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval, 1989. ACM, 1-24.

ZAMAN, A. & BROWN, C. G. Latent semantic indexing and large dataset: Study of term-weighting schemes. Digital Information Management (ICDIM), 2010 Fifth International Conference on, 2010. IEEE, 1-4.

ZHAO, R. & GROSKY, W. I. 2002. Narrowing the semantic gap-improved text-based web document retrieval using visual features. *Multimedia, IEEE Transactions on,* 4**,** 189-200.

ANEXOS

Anexo 1: Listado de palabras vacías (*stop words*) utilizadas en el procesamiento de los documentos.

a, a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, aren't, around, as, aside, ask, asking, associated, at, available, away, awfully, b, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c, c'mon, c's, came, can, can't, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn't, course, currently, d, definitely, described, despite, did, didn't, different, do, does, doesn't, doing, don't, done, down, downwards, during, e, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, f, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, g, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, h, had, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, he's, hello, help, hence, her, here, here's, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i, i'd, i'll, i'm, i've, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn't, it, it'd, it'll, it's, its, itself, j, just, k, keep, keeps, kept, know, knows, known, l, last, lately, later, latter, latterly, least, less, lest, let, let's, like, liked, likely, little, look, looking, looks, ltd, m, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, n, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, o, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, p, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, q, que, quite, qv, r, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, s, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldn't, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t, t's, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that's, thats, the, their, theirs, them, themselves, then, thence, there, there's, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they'd, they'll, they're, they've, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to,

together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, u, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, uucp, v, value, various, very, via, viz, vs, w, want, wants, was, wasn't, way, we, we'd, we'll, we're, we've, welcome, well, went, were, weren't, what, what's, whatever, when, whence, whenever, where, where's, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, who's, whoever, whole, whom, whose, why, will, willing, wish, with, within, without, won't, wonder, would, would, wouldn't, x, y, yes, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, z, zero

Anexo 2: Longitudes de los documentos dada por la cantidad de palabras.

Doc 1: 101 p	Doc 111: 97 p	Doc 221: 60 p	Doc 331: 287 p	Doc 441: 98 p
Doc 2: 233 p	Doc 112: 218 p	Doc 222: 123 p	Doc 332: 248 p	Doc 442: 98 p
Doc 3: 133 p	Doc 113: 376 p	Doc 223: 104 p	Doc 333: 130 p	Doc 443: 121 p
Doc 4: 140 p	Doc 114: 248 p	Doc 224: 65 p	Doc 334: 120 p	Doc 444: 266 p
Doc 5: 132 p	Doc 115: 128 p	Doc 225: 90 p	Doc 335: 63 p	Doc 445: 129 p
Doc 6: 157 p	Doc 116: 212 p	Doc 226: 53 p	Doc 336: 107 p	Doc 446: 142 p
Doc 7: 128 p	Doc 117: 151 p	Doc 227: 45 p	Doc 337: 60 p	Doc 447: 126 p
Doc 8: 86 p	Doc 118: 371 p	Doc 228: 185 p	Doc 338: 100 p	Doc 448: 156 p
Doc 9: 67 p	Doc 119: 89 p	Doc 229: 169 p	Doc 339: 161 p	Doc 449: 92 p
Doc 10: 94 p	Doc 120: 132 p	Doc 230: 166 p	Doc 340: 101 p	Doc 450: 71 p
Doc 11: 117 p	Doc 121: 176 p	Doc 231: 95 p	Doc 341: 142 p	Doc 451: 91 p
Doc 12: 223 p	Doc 122: 203 p	Doc 232: 97 p	Doc 342: 145 p	Doc 452: 82 p
Doc 13: 71 p	Doc 123: 216 p	Doc 233: 126 p	Doc 343: 108 p	Doc 453: 125 p
Doc 14: 211 p	Doc 124: 88 p	Doc 234: 163 p	Doc 344: 82 p	Doc 454: 131 p
Doc 15: 210 p	Doc 125: 100 p	Doc 235: 76 p	Doc 345: 77 p	Doc 455: 223 p
Doc 16: 68 p	Doc 126: 223 p	Doc 236: 142 p	Doc 346: 293 p	Doc 456: 107 p
Doc 17: 109 p	Doc 127: 95 p	Doc 237: 77 p	Doc 347: 266 p	Doc 457: 216 p
Doc 18: 64 p	Doc 128: 122 p	Doc 238: 72 p	Doc 348: 232 p	Doc 458: 283 p
Doc 19: 92 p	Doc 129: 68 p	Doc 239: 102 p	Doc 349: 138 p	Doc 459: 215 p
Doc 20: 44 p	Doc 130: 123 p	Doc 240: 162 p	Doc 350: 107 p	Doc 460: 223 p
Doc 21: 20 p	Doc 131: 204 p	Doc 241: 121 p	Doc 351: 152 p	Doc 461: 99 p
Doc 22: 69 p	Doc 132: 144 p	Doc 242: 177 p	Doc 352: 98 p	Doc 462: 134 p
Doc 23: 186 p	Doc 133: 218 p	Doc 243: 331 p	Doc 353: 140 p	Doc 463: 157 p
Doc 24: 186 p	Doc 134: 92 p	Doc 244: 213 p	Doc 354: 70 p	Doc 464: 72 p
Doc 25: 251 p	Doc 135: 128 p	Doc 245: 164 p	Doc 355: 180 p	Doc 465: 202 p
Doc 26: 162 p	Doc 136: 101 p	Doc 246: 115 p	Doc 356: 137 p	Doc 466: 177 p
Doc 27: 28 p	Doc 137: 283 p	Doc 247: 122 p	Doc 357: 90 p	Doc 467: 131 p

Doc 28: 103 p
Doc 29: 382 p
Doc 30: 62 p
Doc 31: 101 p
Doc 32: 240 p
Doc 33: 74 p
Doc 34: 37 p
Doc 35: 383 p
Doc 36: 125 p
Doc 37: 33 p
Doc 38: 140 p
Doc 39: 71 p
Doc 40: 133 p
Doc 41: 300 p
Doc 42: 86 p
Doc 43: 31 p
Doc 44: 180 p
Doc 45: 72 p
Doc 46: 93 p
Doc 47: 111 p
Doc 48: 123 p
Doc 49: 132 p
Doc 50: 73 p
Doc 51: 109 p
Doc 52: 57 p
Doc 53: 43 p
Doc 54: 136 p
Doc 55: 61 p
Doc 56: 171 p
Doc 57: 111 p
Doc 58: 208 p
Doc 59: 82 p
Doc 60: 200 p
Doc 61: 87 p
Doc 62: 92 p
Doc 63: 143 p
Doc 64: 247 p
Doc 65: 257 p
Doc 138: 76 p
Doc 139: 90 p
Doc 140: 76 p
Doc 141: 57 p
Doc 142: 219 p
Doc 143: 154 p
Doc 144: 200 p
Doc 145: 63 p
Doc 146: 164 p
Doc 147: 81 p
Doc 148: 179 p
Doc 149: 197 p
Doc 150: 182 p
Doc 151: 263 p
Doc 152: 152 p
Doc 153: 107 p
Doc 154: 241 p
Doc 155: 255 p
Doc 156: 68 p
Doc 157: 141 p
Doc 158: 228 p
Doc 159: 204 p
Doc 160: 146 p
Doc 161: 93 p
Doc 162: 213 p
Doc 163: 107 p
Doc 164: 130 p
Doc 165: 218 p
Doc 166: 151 p
Doc 167: 80 p
Doc 168: 128 p
Doc 169: 92 p
Doc 170: 87 p
Doc 171: 81 p
Doc 172: 227 p
Doc 173: 92 p
Doc 174: 149 p
Doc 175: 98 p
Doc 248: 88 p
Doc 249: 97 p
Doc 250: 53 p
Doc 251: 187 p
Doc 252: 182 p
Doc 253: 125 p
Doc 254: 100 p
Doc 255: 220 p
Doc 256: 116 p
Doc 257: 86 p
Doc 258: 173 p
Doc 259: 287 p
Doc 260: 163 p
Doc 261: 124 p
Doc 262: 89 p
Doc 263: 102 p
Doc 264: 315 p
Doc 265: 167 p
Doc 266: 150 p
Doc 267: 207 p
Doc 268: 65 p
Doc 269: 86 p
Doc 270: 121 p
Doc 271: 142 p
Doc 272: 151 p
Doc 273: 65 p
Doc 274: 211 p
Doc 275: 253 p
Doc 276: 133 p
Doc 277: 97 p
Doc 278: 287 p
Doc 279: 105 p
Doc 280: 231 p
Doc 281: 65 p
Doc 282: 300 p
Doc 283: 213 p
Doc 284: 171 p
Doc 285: 80 p
Doc 358: 450 p
Doc 359: 203 p
Doc 360: 47 p
Doc 361: 244 p
Doc 362: 151 p
Doc 363: 52 p
Doc 364: 140 p
Doc 365: 175 p
Doc 366: 79 p
Doc 367: 97 p
Doc 368: 317 p
Doc 369: 36 p
Doc 370: 188 p
Doc 371: 89 p
Doc 372: 139 p
Doc 373: 86 p
Doc 374: 64 p
Doc 375: 96 p
Doc 376: 121 p
Doc 377: 165 p
Doc 378: 357 p
Doc 379: 324 p
Doc 380: 372 p
Doc 381: 95 p
Doc 382: 68 p
Doc 383: 92 p
Doc 384: 168 p
Doc 385: 327 p
Doc 386: 137 p
Doc 387: 167 p
Doc 388: 324 p
Doc 389: 82 p
Doc 390: 311 p
Doc 391: 130 p
Doc 392: 126 p
Doc 393: 205 p
Doc 394: 180 p
Doc 395: 222 p
Doc 468: 132 p
Doc 469: 198 p
Doc 470: 40 p
Doc 471: 156 p
Doc 472: 254 p
Doc 473: 649 p
Doc 474: 159 p
Doc 475: 78 p
Doc 476: 180 p
Doc 477: 167 p
Doc 478: 172 p
Doc 479: 127 p
Doc 480: 280 p
Doc 481: 135 p
Doc 482: 217 p
Doc 483: 86 p
Doc 484: 113 p
Doc 485: 394 p
Doc 486: 126 p
Doc 487: 202 p
Doc 488: 144 p
Doc 489: 29 p
Doc 490: 266 p
Doc 491: 114 p
Doc 492: 122 p
Doc 493: 64 p
Doc 494: 125 p
Doc 495: 155 p
Doc 496: 170 p
Doc 497: 90 p
Doc 498: 69 p
Doc 499: 388 p
Doc 500: 160 p
Doc 501: 105 p
Doc 502: 315 p
Doc 503: 110 p
Doc 504: 194 p
Doc 505: 128 p

Doc 66: 127 p
Doc 67: 107 p
Doc 68: 143 p
Doc 69: 71 p
Doc 70: 290 p
Doc 71: 216 p
Doc 72: 99 p
Doc 73: 229 p
Doc 74: 55 p
Doc 75: 286 p
Doc 76: 186 p
Doc 77: 118 p
Doc 78: 190 p
Doc 79: 107 p
Doc 80: 518 p
Doc 81: 286 p
Doc 82: 142 p
Doc 83: 136 p
Doc 84: 187 p
Doc 85: 168 p
Doc 86: 181 p
Doc 87: 90 p
Doc 88: 115 p
Doc 89: 177 p
Doc 90: 203 p
Doc 91: 132 p
Doc 92: 143 p
Doc 93: 109 p
Doc 94: 129 p
Doc 95: 154 p
Doc 96: 57 p
Doc 97: 172 p
Doc 98: 374 p
Doc 99: 184 p
Doc 100: 369 p
Doc 101: 89 p
Doc 102: 225 p
Doc 103: 55 p
Doc 176: 159 p
Doc 177: 125 p
Doc 178: 69 p
Doc 179: 132 p
Doc 180: 133 p
Doc 181: 124 p
Doc 182: 134 p
Doc 183: 219 p
Doc 184: 90 p
Doc 185: 59 p
Doc 186: 94 p
Doc 187: 168 p
Doc 188: 163 p
Doc 189: 120 p
Doc 190: 166 p
Doc 191: 40 p
Doc 192: 171 p
Doc 193: 126 p
Doc 194: 225 p
Doc 195: 147 p
Doc 196: 187 p
Doc 197: 265 p
Doc 198: 236 p
Doc 199: 276 p
Doc 200: 326 p
Doc 201: 263 p
Doc 202: 107 p
Doc 203: 142 p
Doc 204: 205 p
Doc 205: 198 p
Doc 206: 316 p
Doc 207: 199 p
Doc 208: 340 p
Doc 209: 199 p
Doc 210: 102 p
Doc 211: 94 p
Doc 212: 274 p
Doc 213: 126 p
Doc 286: 146 p
Doc 287: 375 p
Doc 288: 193 p
Doc 289: 235 p
Doc 290: 196 p
Doc 291: 270 p
Doc 292: 213 p
Doc 293: 338 p
Doc 294: 127 p
Doc 295: 189 p
Doc 296: 188 p
Doc 297: 117 p
Doc 298: 384 p
Doc 299: 132 p
Doc 300: 221 p
Doc 301: 168 p
Doc 302: 122 p
Doc 303: 133 p
Doc 304: 183 p
Doc 305: 230 p
Doc 306: 180 p
Doc 307: 67 p
Doc 308: 159 p
Doc 309: 128 p
Doc 310: 402 p
Doc 311: 173 p
Doc 312: 214 p
Doc 313: 291 p
Doc 314: 134 p
Doc 315: 138 p
Doc 316: 80 p
Doc 317: 183 p
Doc 318: 38 p
Doc 319: 109 p
Doc 320: 65 p
Doc 321: 195 p
Doc 322: 275 p
Doc 323: 222 p
Doc 396: 89 p
Doc 397: 115 p
Doc 398: 103 p
Doc 399: 192 p
Doc 400: 220 p
Doc 401: 243 p
Doc 402: 116 p
Doc 403: 119 p
Doc 404: 103 p
Doc 405: 69 p
Doc 406: 83 p
Doc 407: 209 p
Doc 408: 200 p
Doc 409: 171 p
Doc 410: 172 p
Doc 411: 165 p
Doc 412: 152 p
Doc 413: 122 p
Doc 414: 174 p
Doc 415: 205 p
Doc 416: 145 p
Doc 417: 137 p
Doc 418: 279 p
Doc 419: 123 p
Doc 420: 281 p
Doc 421: 237 p
Doc 422: 190 p
Doc 423: 107 p
Doc 424: 142 p
Doc 425: 406 p
Doc 426: 190 p
Doc 427: 323 p
Doc 428: 175 p
Doc 429: 92 p
Doc 430: 186 p
Doc 431: 134 p
Doc 432: 142 p
Doc 433: 125 p
Doc 506: 114 p
Doc 507: 145 p
Doc 508: 209 p
Doc 509: 131 p
Doc 510: 211 p
Doc 511: 142 p
Doc 512: 192 p
Doc 513: 134 p
Doc 514: 82 p
Doc 515: 52 p
Doc 516: 58 p
Doc 517: 191 p
Doc 518: 133 p
Doc 519: 160 p
Doc 520: 143 p
Doc 521: 74 p
Doc 522: 111 p
Doc 523: 58 p
Doc 524: 119 p
Doc 525: 82 p
Doc 526: 84 p
Doc 527: 58 p
Doc 528: 195 p
Doc 529: 170 p
Doc 530: 147 p
Doc 531: 110 p
Doc 532: 131 p
Doc 533: 299 p
Doc 534: 166 p
Doc 535: 142 p
Doc 536: 287 p
Doc 537: 145 p
Doc 538: 55 p
Doc 539: 115 p
Doc 540: 132 p
Doc 541: 148 p
Doc 542: 260 p
Doc 543: 204 p

Doc 104: 321 p	Doc 214: 204 p	Doc 324: 307 p	Doc 434: 120 p	Doc 544: 168 p
Doc 105: 88 p	Doc 215: 128 p	Doc 325: 139 p	Doc 435: 325 p	Doc 545: 142 p
Doc 106: 68 p	Doc 216: 135 p	Doc 326: 391 p	Doc 436: 185 p	Doc 546: 62 p
Doc 107: 419 p	Doc 217: 91 p	Doc 327: 88 p	Doc 437: 93 p	Doc 547: 142 p
Doc 108: 175 p	Doc 218: 273 p	Doc 328: 121 p	Doc 438: 490 p	Doc 548: 161 p
Doc 109: 57 p	Doc 219: 164 p	Doc 329: 220 p	Doc 439: 168 p	Doc 549: 206 p
Doc 110: 203 p	Doc 220: 102 p	Doc 330: 154 p	Doc 440: 195 p	Doc 550: 87 p

Anexo 3: Longitud promedio de cada subgrupo (*bin*) de documentos.

Bin 1: 33.2	Bin 12: 91.8	Bin 23: 127.3	Bin 34: 164.6	Bin 45: 216.7
Bin 2: 49.9	Bin 13: 94.2	Bin 24: 129.7	Bin 35: 167.8	Bin 46: 221.3
Bin 3: 57.6	Bin 14: 97.6	Bin 25: 132.0	Bin 36: 171.3	Bin 47: 228.3
Bin 4: 62.8	Bin 15: 100.7	Bin 26: 133.5	Bin 37: 175.8	Bin 48: 241.9
Bin 5: 66.6	Bin 16: 104.1	Bin 27: 136.8	Bin 38: 181.1	Bin 49: 258.7
Bin 6: 69.8	Bin 17: 107.5	Bin 28: 140.8	Bin 39: 185.9	Bin 50: 274.0
Bin 7: 74.2	Bin 18: 111.2	Bin 29: 142.2	Bin 40: 190.3	Bin 51: 286.7
Bin 8: 80.0	Bin 19: 116.3	Bin 30: 144.6	Bin 41: 196.6	Bin 52: 307.3
Bin 9: 83.9	Bin 20: 120.6	Bin 31: 149.7	Bin 42: 202.0	Bin 53: 327.9
Bin 10: 87.4	Bin 21: 122.8	Bin 32: 154.7	Bin 43: 205.7	Bin 54: 374.3
Bin 11: 89.6	Bin 22: 125.4	Bin 33: 160.6	Bin 44: 211.6	Bin 55: 450.7

Anexo 4: Documentos recuperados utilizando diversos valores de k para la función: tf * idf * normalización pivote

$$w_{ij} = \frac{tf * idf}{m * b + (1 - m) * p}$$

		k=10		k=20		k=30		k=40		k=50		k=60		k=70		k=80		k=90		k=110		k=130	
C	D Rel	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec
1	37	64	37	38	33	30	28	20	19	18	18	12	12	11	11	9	9	8	8	3	3	1	1
2	16	83	16	26	14	21	12	12	7	8	6	7	6	4	3	4	3	3	2	2	2	2	2
3	22	47	21	39	20	25	17	18	14	10	8	10	8	10	8	10	8	8	7	5	5	4	4
4	23	88	19	49	15	22	12	17	10	12	9	9	5	6	3	5	3	5	2	2	1	2	1
5	26	55	26	34	25	23	20	19	18	15	14	11	10	9	8	9	8	8	7	6	5	6	5
6	13	45	13	22	11	14	9	10	7	8	6	9	6	7	5	6	5	6	5	6	5	6	5
7	15	149	15	24	13	14	8	10	7	9	7	9	7	8	7	8	7	7	7	6	6	5	5
8	11	37	10	26	7	25	7	19	6	18	5	18	5	14	4	13	4	13	4	8	4	4	7
9	28	104	25	33	22	24	16	17	12	10	5	7	4	7	4	4	3	5	3	2	2	2	2
10	24	78	23	40	22	20	12	10	8	9	8	9	6	8	6	7	5	4	3	5	3	5	3
11	18	37	18	32	18	26	17	19	14	15	11	12	9	8	5	7	4	5	3	5	3	3	3
12	9	136	9	66	9	32	9	15	7	13	7	11	6	10	6	8	6	7	6	7	6	6	5
13	21	31	21	23	21	11	9	11	9	11	9	10	9	10	9	9	8	9	8	6	6	6	6
14	16	123	16	37	16	26	14	13	7	7	5	6	4	5	4	5	4	3	3	4	4	4	4
15	29	127	25	44	27	31	24	23	18	13	12	10	9	10	9	6	5	7	5	5	3	5	3
16	13	126	13	29	13	27	13	25	12	18	8	15	8	13	8	11	7	10	6	5	4	5	4
17	21	148	19	49	18	25	15	13	10	5	3	3	2	2	2	0	0	0	0	0	0	0	0
18	15	173	13	36	14	24	13	19	12	16	10	16	10	14	9	13	8	11	7	9	5	7	4
19	4	52	1	39	0	22	0	15	0	10	0	7	0	2	1	2	1	4	2	3	2	2	1
Media		89,6	**17,9**	36,1	**16,7**	23,3	**13,4**	16,1	**10,4**	11,8	**7,9**	10,1	**6,6**	8,3	**5,9**	7,1	**5,1**	6,5	**4,6**	4,7	**3,6**	3,9	**3,4**
DT		44,9	**7,8**	10,9	**7,6**	5,58	**6,3**	4,4	**4,8**	3,9	**4,0**	3,6	**2,9**	3,5	**2,7**	3,4	**2,5**	3,1	**2,4**	2,3	**1,7**	2,0	**1,9**

C: consultas. **DRel**: documentos relevantes para cada consulta según los juicios de relevancia de la colección.
DRec: documentos recuperados. **DR Rec**: documentos relevantes recuperados.

Anexo 5: Documentos recuperados utilizando diversos valores de k para la función: logaritmo * entropía * normalización pivote

$$w_{ij} = \frac{\left(log(tf_{ij}) + 1\right) * \left(1 + \sum_{j=1}^{N} \frac{\frac{tf_{ij}}{F_i} \log \frac{tf_{ij}}{F_i}}{\log N}\right)}{m * b + (1 - m) * p}$$

		k=10		k=20		k=30		k=40		k=50		k=60		k=70		k=80		k=90		k=110		k=130	
C	D Rel	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec	D Rec	DR Rec
1	37	62	35	36	29	27	25	17	16	13	12	7	6	5	4	3	2	1	1	1	1	0	0
2	16	71	16	23	14	17	11	9	7	6	6	5	5	5	5	3	3	3	3	1	1	1	1
3	22	46	20	33	18	17	14	13	11	10	8	9	8	9	8	6	6	5	5	4	4	3	3
4	23	81	22	33	15	24	15	18	12	13	10	5	4	5	3	2	2	1	1	2	2	0	0
5	26	40	25	27	23	24	21	20	18	15	13	14	12	9	8	5	5	3	3	2	2	1	1
6	13	49	13	31	13	21	11	16	11	12	9	9	7	7	6	6	5	6	5	5	5	5	5
7	15	62	12	25	13	13	8	9	7	10	7	8	7	7	7	7	7	6	6	5	5	5	5
8	11	44	10	31	9	21	9	15	5	15	5	11	4	10	3	7	4	6	3	2	2	2	2
9	28	69	24	31	23	19	12	13	9	4	3	1	1	1	1	1	1	1	1	1	1	0	0
10	24	86	22	33	15	17	10	10	7	7	4	3	1	3	1	3	1	2	1	2	1	3	2
11	18	37	18	24	17	21	15	19	14	16	13	13	11	5	4	3	3	3	3	3	3	2	2
12	9	124	9	39	9	22	9	14	7	12	7	11	7	11	7	10	7	9	7	7	5	5	4
13	21	32	21	27	21	22	20	15	14	14	13	12	11	12	11	12	11	10	10	6	6	5	5
14	16	127	14	38	15	19	13	12	9	8	6	7	5	4	3	4	3	4	3	3	2	3	2
15	29	107	29	37	28	30	24	17	13	12	11	10	9	9	9	5	5	2	2	1	1	2	2
16	13	87	13	30	13	27	13	24	13	18	12	12	10	11	9	9	8	6	5	6	5	4	4
17	21	130	18	49	12	25	6	16	4	7	2	5	1	2	0	1	0	1	0	0	0	0	0
18	15	92	13	30	14	24	13	19	11	17	10	15	10	14	9	13	8	13	8	7	4	5	3
19	4	54	0	35	2	11	1	9	1	5	1	2	1	4	1	2	2	2	1	1	1	1	1
Media		73,6	**17,5**	32,2	**15,9**	21,1	**13,1**	15	**9,9**	11,2	**8**	8,3	**6,3**	7	**5,2**	5,3	**4,3**	4,4	**3,5**	3,1	**2,6**	2,4	**2,2**
DT		31,2	**7,9**	6,2	**6,6**	4,8	**6,0**	4,1	**4,3**	4,2	**3,9**	3,7	**3,7**	3,6	**3,3**	3,5	**2,9**	2,7	**2,3**	2,2	**1,8**	1,9	**1,7**

C: consultas. **DRel**: documentos relevantes **DRec**: documentos recuperados. **DR Rec**: documentos relevantes recuperados.

Anexo 6: Precisión y cobertura a partir de los mejores resultados obtenidos, según la reducción de las dimensiones, siguiendo el esquema: tf * idf * normalización pivote.

$$w_{ij} = \frac{tf * idf}{m * b + (1 - m) * p}$$

	k=10		k=20		k=30		k=40		k=50		k=60		k=70		k=80		k=90		k=110		k=130	
C	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)
1	57,8	100	86,8	89,1	93,3	75,6	95	51,3	48,6	65,4	100	32,4	100	29,7	100	24,3	100	21,6	100	08,1	100	02,7
2	19,2	100	53,8	87,5	57,1	75	58,3	43,7	75	37,5	85,7	37,5	75	18,7	75	18,7	66,6	12,5	100	12,5	100	12,5
3	44,6	95,4	51,2	90,9	68	77,2	77,7	63,6	80	36,3	80	36,3	80	36,3	80	36,3	87,5	31,8	100	22,7	100	18,1
4	21,5	82,6	30,6	65,2	54,5	52,1	58,8	43,4	75	39,1	55,5	21,7	50	13,0	60	13,0	40	08,6	50	04,3	50	04,3
5	47,2	100	73,5	96,1	86,9	76,9	94,7	69,2	93,3	53,8	90,9	38,4	88,8	30,7	88,8	30,7	87,5	26,9	83,3	19,2	83,3	19,2
6	28,8	100	84,6	62,8	64,2	69,2	70	53,8	75	46,1	66,6	46,1	71,4	38,4	83,3	38,4	83,3	38,4	83,3	38,4	83,3	38,4
7	10,1	100	54,1	86,6	57,1	53,3	70	46,6	77,7	46,6	77,7	46,6	87,5	46,6	87,5	46,6	100	46,6	100	40	100	33,3
8	27,0	90,9	26,9	63,6	28	63,6	31,5	54,5	27,7	45,4	27,7	45,4	28,5	36,3	30,7	36,3	30,7	36,3	50	36,3	57,1	36,3
9	24,0	89,2	66,6	78,5	66,6	57,1	70,5	42,8	50	17,8	57,1	14,2	57,1	14,2	75	10,7	60	10,7	100	07,1	100	07,1
10	29,4	95,8	91,6	68,7	60	50	80	33,3	88,8	33,3	66,6	25	75	25	71,4	20,8	75	12,5	60	12,5	60	12,5
11	48,6	100	56,2	100	65,3	94,4	73,6	77,7	73,3	61,1	75	50	62,5	27,7	57,1	22,2	60	16,6	60	16,6	100	16,6
12	6,61	100	13,6	100	28,1	100	46,6	77,7	53,8	77,7	54,5	66,6	60	66,6	75	66,6	85,7	66,6	85,7	66,6	83,3	55,5
13	67,7	100	91,3	100	81,8	42,8	81,8	42,8	81,8	42,8	90	42,8	90	42,8	88,8	38,0	88,8	38,0	100	28,5	100	28,5
14	13,0	100	43,2	100	53,8	87,5	53,8	43,7	71,4	31,2	66,6	25	80	25	80	25	100	18,7	100	25	100	25
15	19,6	86,2	61,3	93,1	77,4	82,7	78,2	62,1	92,3	41,3	90	31,0	90	31,0	83,3	17,2	71,4	17,2	60	10,3	60	10,3
16	10,3	100	44,8	100	48,1	100	48	92,3	44,4	61,5	53,3	61,5	61,5	61,5	63,6	53,8	60	46,1	80	30,7	80	30,7
17	12,8	90,4	36,7	85,7	60	71,4	76,9	47,6	60	14,2	66,6	09,5	100	09,5	0	0	0	0	0	0	0	0
18	7,51	86,6	38,8	93,3	54,1	86,6	63,1	80	62,5	66,6	62,5	66,6	64,2	60	61,5	53,3	63,6	46,6	55,5	33,3	57,1	26,6
19	1,92	25,0	0	0	0	0	0	0	0	0	0	0	50	25	50	25	50	50	66,6	50	50	25
M	26,2	91,7	52,9	82,2	58,1	69,2	64,6	54,1	64,7	43,1	66,6	36,6	72,1	33,5	69	30,3	68,9	28,7	75,4	24,3	64,7	21,1
DT	18,7	17,2	25,8	23,8	21,8	23,7	22,5	20,7	23,4	19,3	23,6	18,3	18,8	16,2	23,2	16,7	26,0	17,6	26,5	17,1	27,0	14,2

C: consulta. **M**: media **P**: precisión. **C**: cobertura. **DT:** desviación típica

Anexo 7: Precisión y cobertura a partir de los mejores resultados obtenidos, según la reducción de las dimensiones, siguiendo el esquema: logaritmo * entropía * normalización pivote.

$$w_{ij} = \frac{\left(log\left(tf_{ij}\right) + 1\right) * \left(1 + \sum_{j=1}^{N} \frac{\frac{tf_{ij}}{F_i} \log \frac{tf_{ij}}{F_i}}{\log N}\right)}{m * b + (1 - m) * p}$$

	k=10		k=20		k=30		k=40		k=50		k=60		k=70		k=80		k=90		k=110		k=130	
C	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)	P (%)	C (%)
1	56,4	94,5	80,5	78,3	92,5	67,5	94,1	43,2	92,3	32,4	85,7	16,2	80	10,8	66,6	05,4	100	2,7	100	02,7	0	0
2	22,5	100	60,8	87,5	64,7	68,7	77,7	43,7	100	37,5	100	31,2	100	31,2	100	18,7	100	18,7	100	06,2	100	06,2
3	43,4	90,9	54,5	81,8	82,3	63,6	84,6	50	80	36,3	88,8	36,3	88,8	36,3	100	27,2	100	22,7	100	18,1	100	13,6
4	27,1	95,6	45,4	65,2	62,5	65,2	66,6	52,1	76,9	43,4	80	17,3	60	13,0	100	08,6	100	04,3	100	08,6	0	0
5	62,5	96,1	85,1	88,4	87,5	80,7	90	69,2	86,6	50	85,7	46,1	88,8	30,7	100	19,2	100	11,5	100	07,6	100	03,8
6	26,5	100	41,9	100	52,3	84,6	68,7	84,6	75	69,2	77,7	53,8	85,7	46,1	83,3	38,4	83,3	38,4	100	38,4	100	38,4
7	19,3	80	52	86,6	61,5	53,3	77,7	46,6	70	46,6	87,5	46,6	100	46,6	100	46,6	100	40	100	33,3	100	33,3
8	22,7	90,9	29	81,8	42,8	81,8	33,3	45,4	33,3	45,4	36,3	36,3	30	27,2	57,1	36,3	50	27,2	100	18,1	100	18,1
9	34,7	85,7	74,1	82,1	63,1	42,8	69,2	32,1	75	10,7	100	03,5	100	03,5	100	03,5	100	03,5	100	03,5	0	0
10	25,5	91,6	45,4	62,5	58,8	41,6	70	29,1	57,1	16,6	33,3	04,1	33,3	04,1	33,3	04,1	50	04,1	50	04,1	66,6	08,3
11	48,6	100	70,8	94,4	71,4	83,3	73,6	77,7	81,2	72,2	84,6	61,1	80	22,2	100	16,6	100	16,6	100	16,6	100	11,1
12	7,2	100	23,0	100	40,9	100	50	77,7	58,3	77,7	63,6	77,7	63,6	77,7	70	77,7	77,7	77,7	71,4	55,5	80	44,4
13	65,6	100	77,7	100	90,9	95,2	93,3	66,6	92,8	61,9	91,6	52,3	91,6	52,3	91,6	52,3	47,6	64,5	100	28,5	100	23,8
14	11,0	19,5	39,4	93,7	68,4	81,2	75	56,2	75	37,5	71,4	31,2	75	18,7	75	18,7	75	18,7	66,6	12,5	66,6	12,5
15	27,1	100	75,6	96,5	80	82,7	76,4	44,8	91,6	37,9	90	31,0	100	31,0	100	17,2	100	06,8	100	03,4	100	06,8
16	14,9	100	43,3	100	48,1	100	54,1	100	66,6	92,3	83,3	76,9	81,8	69,2	88,8	61,5	83,3	38,4	83,3	38,4	100	30,7
17	13,8	85,7	24,4	57,1	24	28,5	25	19	28,5	09,5	20	04,7	0	0	0	0	0	0	0	0	0	0
18	14,1	86,6	46,6	93,3	54,1	86,6	57,8	73,3	58,8	66,6	66,6	66,6	64,2	60	61,5	53,3	61,5	53,3	57,1	26,6	60	20
19	0	0	5,7	50	9	25	11,1	25	20	25	50	25	25	25	100	50	50	25	100	25	100	25
M	28,5	85,1	51,3	84,1	60,7	70,1	65,6	54,5	69,4	45,7	73,4	37,7	70,9	31,8	80,3	29,2	77,8	24,9	85,7	18,2	72,2	15,5
DT	18,7	27,4	22,0	15,4	21,9	22,7	22,7	21,8	22,3	22,9	23,1	23,2	29,3	22,2	27,5	22,5	27,9	22,2	26,5	15,4	40,6	13,8

C: consulta. **M**: media **P**: precisión. **C**: cobertura. **DT:** desviación típica