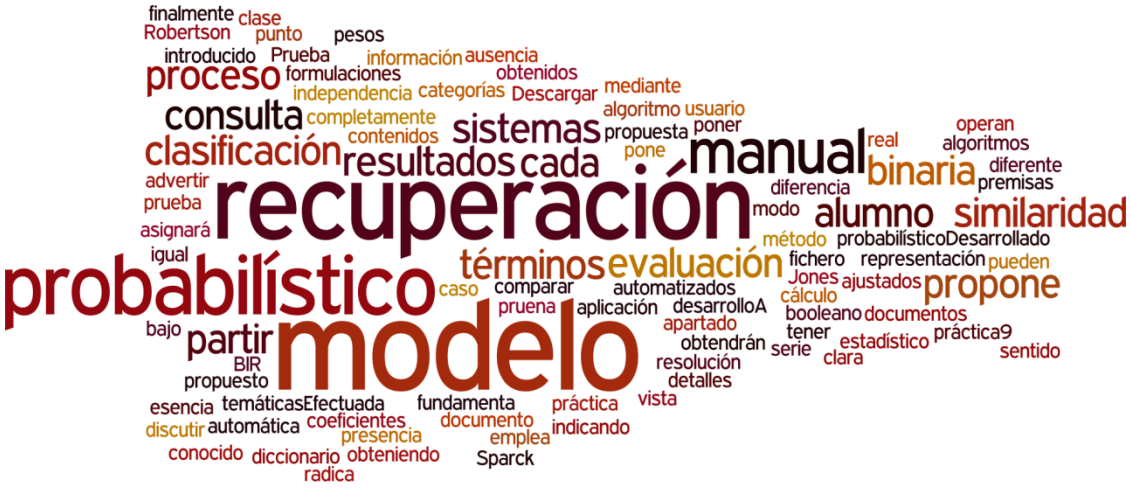
Técnicas avanzadas de recuperación de información

Procesos, técnicas y métodos

Manuel Blázquez Ochando



Monografías electrónicas

mblazquez.es

025.4.036:004 **BLA** **tec**	BLÁZQUEZ OCHANDO, Manuel Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos / Manuel Blázquez Ochando .– Madrid: mblazquez.es, 2013. Xp. ; 21cm.– (Libros y manuales de la Documentación; 4) ISBN 978-84-695-8030-1 1. Biblioteconomía y Documentación 2. Recuperación de Información I. Título II. Series

UNIVERSIDAD COMPLUTENSE DE MADRID
Facultad de Ciencias de la Documentación

1ªed. mayo 2013, Madrid

Publicado por mblazquez.es

ISBN 978-84-695-8030-1

Índice

1. Introducción

¿Qué son técnicas avanzadas de recuperación de información? Son todos aquellos procesos destinados a la recuperación de información, desde la generación de las colecciones, su depuración, indexado, tratamiento textual, clasificación, almacenamiento, recuperación mediante modelos booleanos, vectoriales, probabilísticos, basados en el lenguaje, así como todos aquellos elementos que inciden en cualquier aspecto relacionado como por ejemplo el interfaz de consulta, el comportamiento del usuario, la retroalimentación de las consultas y la representación de la información.

Todos estos aspectos de la recuperación de información serán tratados desde la óptica de la Documentación. Esto implica un enfoque práctico y menos teórico, con el que se pretende enseñar la forma en la que actuan tales componentes, su interrelación, aplicaciones reales a la recuperación en la web, su aplicación en motores de búsqueda, catálogos bibliográficos OPAC, así como otras aplicaciones bibliográfico-documentales e informacionales. En este sentido el presente curso representa una guía eficaz y directa para comprender cómo aprovechar tales técnicas e implantarlas en casos reales.

2. Cadena documental de la recuperación de información

La recuperación de información no puede concebirse como un elemento único e indisoluble como por ejemplo única y exclusivamente un modelo o algoritmo de recuperación. La recuperación de información debe estudiarse como un compendio de procesos altamente interrelacionados que conforman una verdadera cadena documental en recuperación de información. Esta es la visión que se proporciona en la *figura1*.

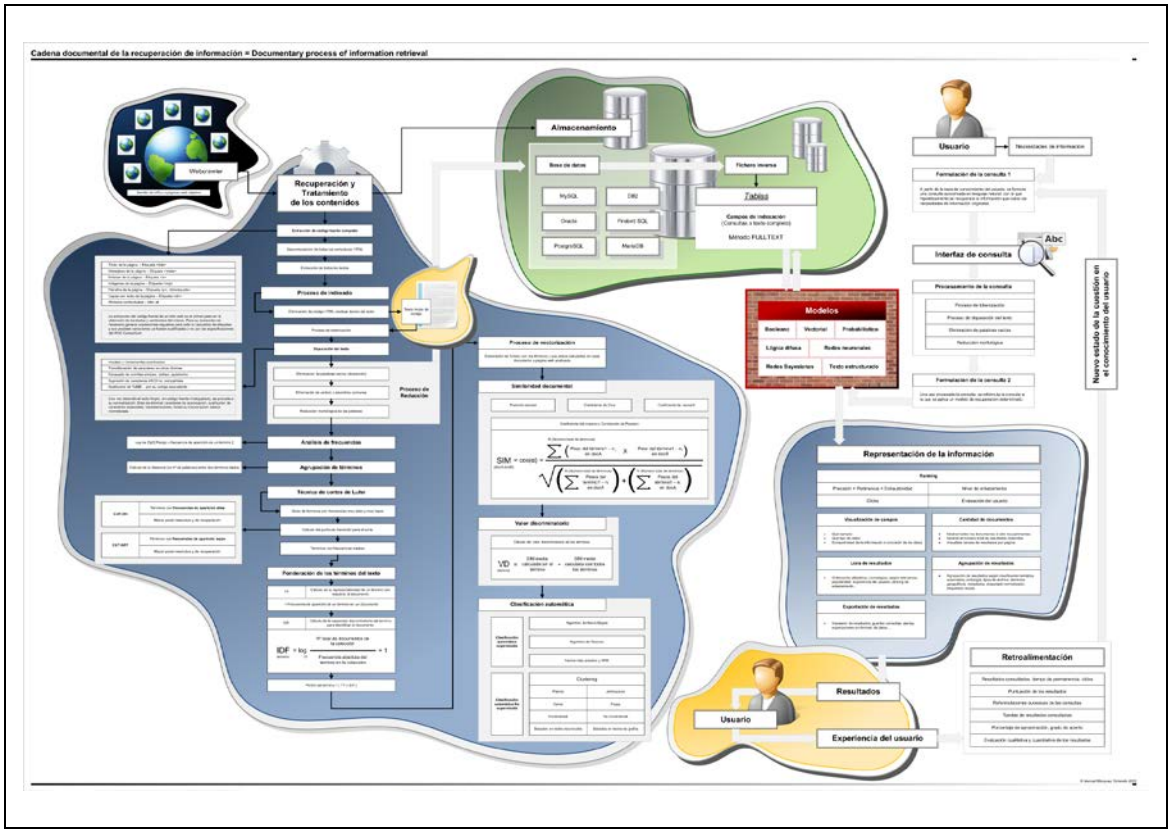


Figura 1. La cadena documental en recuperación de información

Para que la recuperación de información tenga lugar, es necesario comenzar a estudiar los procesos de webcrawling. Un webcrawler es un programa especializado en el análisis y rastreo de sitios, páginas y recursos de la web, de forma automática o semiautomática, recuperando como objeto principal, todos los enlaces o lo que es lo mismo el tejido hipertextual de la web. De esta forma es posible determinar que una página web contiene cinco enlaces de los que se obtienen a su vez diez. Este fenómeno se le conoce como nivel de profundidad del análisis. Cada enlace extraído con el webcrawler es analizado y obtiene a su vez terceros enlaces que son igualmente

analizados, desentrañando una estructura escalar y jerárquica que puede ser almacenada y procesada por el programa. De esta forma, se pueden realizar análisis de sitios web con distintos niveles de profundidad y poder obtener el número total de enlaces para cada nivel, el número de enlaces internos, externos, documentos ofimáticos, etc. De hecho los grandes buscadores emplean miles de programas webcrawler para generar un mapa de la web que es utilizado por los usuarios para recuperar cualquier contenido enlazado y publicado en un servidor web remoto. No obstante, un webcrawler no actúa por sí solo. Requiere de una lista de enlaces de dominios, sitios o páginas web con las que trabajar inicialmente. Esta lista de páginas web objetivo, se denomina *seed* o *semilla*. A partir de los enlaces indicados en la semilla el webcrawler comienza un análisis de cada página uno a uno, recuperando todos sus contenidos. Inicialmente el contenido recuperado de las páginas web es el código fuente completo de la misma. Este código es fundamentalmente HTML. Dentro de la codificación HTML, se encuentra el texto de interés para su indexación y su recuperación efectiva. Para que tenga lugar el proceso de indexado se hace obligada la eliminación del código HTML, a fin de obtener el texto límpio de códigos. Una vez se tiene el texto completo de la página se realiza un proceso denominado *tokenización* que toma su nombre del inglés *token* cuyo significado es fichas o muestras. El proceso consiste en la división del texto en elementos más pequeños, dicho de otra forma en palabras. De hecho para poder tratar adecuadamente el texto de la página web resulta imprescindible tratar palabra por palabra para aplicar cambios sustanciales denominados *normalización y depuración del texto*. Ello consiste en la transformación de vocales y consonantes acentuadas en su equivalente sin acentuar, la transliteración de caracteres en otros idiomas, el escapado de comillas simples, dobles, apóstrofos, la supresión de caracteres ASCII no compatibles, la sustitución de caracteres especiales por su código equivalente, etc. Resulta evidente que el análisis palabra por palabra facilita la ejecución de todos estos cambios, en vez de tenerlos que hacer en bloque, ya que en muchas ocasiones la identificación de tales casos de normalización resulta bastante compleja. Además palabra por palabra también se identifica de forma más sencilla si el término resulta ser una palabra vacía. Las palabras vacías, también denominadas stopwords, son términos que carecen de carga semántica y que no tienen una validez clara a la hora de recuperar información, dada su alta frecuencia de aparición y poca representatividad del contenido. También son eliminados ciertos verbos y adverbios demasiado comunes como para ofrecer una representatividad y precisión de los contenidos de la página web.

Finalmente también se lleva a cabo, en muchos casos, el proceso de *reducción morfológica* o *stemming*. En este caso se consigue comprimir el tamaño de las palabras y por ende de los textos de la página web, ya que la reducción morfológica consiste en la supresión de los prefijos, sufijos, géneros y desinencias de los términos del texto, con el objetivo de mejorar, la exhaustividad de la recuperación de los contenidos. Cuando el texto ha sido normalizado y depurado, se llevan a cabo una serie de procesos de tipo *estadístico-matemático*s. Estos son el *análisis de frecuencias*, la *técnica de cortes de Luhn* y la obtención de *pesos TF-IDF*. El análisis de frecuencias consiste en la contabilización del número de ocurrencias de un término en el texto del documento, denominado frecuencia relativa al documento, y la contabilización del número de ocurrencias del mismo término en todos los documentos de la colección, denominado frecuencia absoluta. Cuando un término tiene una frecuencia absoluta baja y una frecuencia relativa baja, será indicativo de una baja representatividad de la colección pero una alta capacidad discriminatoria, dado que identifica muy bien un determinado documento. Al contrario una frecuencia absoluta alta, no siempre significa una alta representatividad del término para con la colección. En tal caso dependerá de que el término no sea una palabra vacía. Como puede deducirse el análisis de frecuencias resulta clave para saber cómo será la recuperación de los contenidos a través de los términos que utilice el usuario en su consulta. En relación a la técnica de cortes de Luhn, resulta importante su uso para reducir aún más el tamaño del texto de la página web obtenida originalmente a través del webcrawler. Los cortes de Luhn permiten eliminar términos con una altísima frecuencia de aparición (cut-on) y una mínima frecuencia de aparición (cut-off), ya que tales términos no son significativos para recuperar contenidos con un equilibrio entre precisión y exhaustividad. Finalmente la obtención de los pesos TF-IDF, consiste en el cálculo de la importancia de un término para discriminar y representar al documento y a la colección. Se trata de un valor o coeficiente de tipo numérico y decimal que permite traducir al lenguaje matemático estadístico cada término del texto. Tanto las frecuencias como los pesos son ampliamente utilizados en los cálculos de similaridad utilizados por los distintos algoritmos de recuperación de información en la mayoría de los modelos conocidos. Pero no es posible recuperar ninguna información si toda la información y todos los resultados de los pasos dados hasta el momento no son almacenados adecuadamente. Cualquier sistema de recuperación, dispone de un método de almacenamiento o base de datos, que genera un fichero inverso de todos los contenidos de las tablas que alojan la

información de las páginas web recuperadas y su texto depurado. Pero además se generan otro tipo de ficheros en la base de datos, como por ejemplo el fichero diccionario que recoge toda la información de cada término de cada texto de cada página web, así como el cursor de posición, su frecuencia de aparición, sus pesos, etc. A partir de aquí los textos pueden ser indexados correctamente, ya que serán éstos y no otros los textos que conformarán la base de conocimiento sobre la que se recuperarán informaciones y contenidos, aplicando los distintos modelos de recuperación. Los modelos de recuperación más importantes son el booleano, vectorial, probabilístico, de lógica difusa, de redes neuronales, de redes bayesianas y de texto estructurado. No obstante, hay que tener en cuenta, que la recuperación no es posible si no existe un componente imprescindible, el usuario. El usuario formula las consultas en un sistema de recuperación de información, debido a que tiene unas necesidades de información claras. Durante el proceso de formulación, el usuario cuenta con unos conocimientos limitados que no suelen corresponderse con la base de conocimiento. Por ello trata de expresar con su lenguaje natural y de forma aproximada una hipótesis de términos que deberían resolver sus dudas. A esta hipótesis se la denomina *consulta del usuario*. El usuario realiza la consulta en un elemento indispensable que pone en comunicación el algoritmo de recuperación de información de alguno de los modelos anteriormente citados, con la consulta. Este elemento se denomina, *Interfaz de consulta*. Se trata habitualmente de una página web con un formulario más o menos complejo que permite al usuario escribir su consulta y procesarla en el sistema de recuperación. Una vez que es recibida por el sistema, la consulta del usuario que en definitiva consiste en una cadena de caracteres conformada por palabras o frases, recibe el mismo proceso de depuración y normalización que cualquier otro texto. Se aplica la tokenización, la eliminación de palabras vacías, la transformación de caracteres, su transliteración, etc. Como resultado de ello, el sistema genera una consulta interna distinta a la que en origen formuló el usuario, añadiendo o adaptando sus términos e incluso añadiendo otros para optimizar la calidad de los resultados. Por tanto la consulta definitiva es procesada con alguno de los modelos de recuperación y se devuelve una serie de resultados que deben ser representados para su visualización y aprovechamiento por parte del usuario. Este estadio se denomina *fase de representación de la información*. Para ello se ordenan los resultados obtenidos generando un ranking que puede estar establecido por el coeficiente de similaridad de los documentos para con la consulta del usuario, según la precisión obtenida, la pertinencia más la exhaustividad, según el

número de clics recibidos en los resultados, su nivel de enlazamiento o la evaluación recibida por terceros usuarios. Además se tiene en cuenta la visualización de los campos de datos de los resultados. Esto es, qué tipo de información se muestra y con qué nivel de detalle. Por otra parte el número de resultados o documentos a mostrar por página, su agrupación y sus posibilidades para ser exportados. Como es lógico los resultados pueden o no satisfacer al usuario, generando un aspecto vital para mejorar los sistemas de recuperación. Se trata de la *experiencia del usuario*. La experiencia del usuario es el comportamiento que éste demuestra cuando se encuentra con los resultados del sistema, provocando diversas actitudes o acciones para con la información. Por ejemplo el usuario puede mostrar interés por una serie de contenidos y no por otros, puede pasar de la primera página para encontrar la solución a su problema, puede pasar mucho tiempo observando un resultado y muy poco otro similar. Todo ello forma parte de la experiencia que en muchos casos es difícil observar desde una herramienta informática como puede ser un buscador. Para ello existen dos métodos de obtener el comportamiento del usuario. En primer lugar a través de los archivos de consultas, también denominados *querylogs* que registran las consultas y las actividades de los usuarios a través de sus clics y en segundo lugar la retroalimentación que ellos generan cuando puntúan los resultados o reformulan sus consultas originales, ofreciendo pistas sobre qué información pretenden obtener y en qué manera desean que les sea suministrada.

3. Conceptos básicos en recuperación de información

La recuperación de información

- Parte de la informática que estudia la recuperación de la información (no datos) de una colección de documentos escritos. Los documentos recuperados pueden satisfacer una necesidad de información de un usuario expresada normalmente en lenguaje natural. BAEZA YATES, R.; RIBEIRO-NETO, B. 1999. Modern Information Retrieval. Addison Wesley

- La localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta. KORFHAGE, R. 1997. Information storage and retrieval. Wiley Computer

- Un sistema de recuperación de información procesa archivos de registros y peticiones de información, e identifica y recupera de los archivos ciertos registros en respuesta a las peticiones de información. SALTON, G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley

- La recuperación de información se centra en la representación, almacenamiento, organización y acceso a elementos de información. Estos procesos deberían proporcionar al usuario la capacidad de acceder a la información que necesita. Sin embargo existe un problema bastante importante en lo referente a la caracterización de las necesidades de información del usuario, que no suele ser fácil de solucionar. SÁNCHEZ JIMÉNEZ, R. 2011? [Asignatura] Técnicas avanzadas de recuperación de información.

- Es el proceso por el cual las demandas informativas y documentales del usuario son resueltas en un sistema de información, compuesto por un corpus documental de volumen variable, cuyo tratamiento de indexación y almacenamiento hacen posible su estructuración, interrogación y representación,

por medio del empleo de algoritmos matemáticos, estadísticos y semánticos. BLÁZQUEZ, M. 2012 [Asignatura]

La consulta

- **Necesidad de información.** Es la declaración en lenguaje natural de la información que demanda o requiere el usuario para el desempeño de sus actividades y funciones.

- **Formulación del usuario.** Proceso racional del usuario para confeccionar la frase o sucesión de términos con los que efectuará la consulta ó interrogación del sistema.

- **Consulta del usuario.** Es la expresión con la que se configura la demanda informativa del usuario, por regla general, en lenguaje natural, utilizando los términos y palabras que le resultan más aproximados al objeto de recuperación ó cuya previsión e intuición le sugieren un mejor aprovechamiento del sistema en su búsqueda.

- **Formulación del sistema.** Procesamiento y reformulación de la consulta del usuario que implica su descomposición en unidades mínimas (término a término), sustitución de caracteres extraños, procesos de reducción, eliminación de palabras vacías, eliminación de signos diacríticos, sustitución y adición de términos normalizados. Finalmente una vez depurada y adaptada la consulta, se aplican los operadores booleanos y especiales propios del algoritmo de recuperación que se fuere a emplear.

- **Consulta del sistema.** Es el resultado de la formulación del sistema partiendo de la consulta del usuario. Por regla general una sentencia de consulta optimizada para la recuperación en el sistema de información que equivale a la expresada por el usuario en lenguaje natural. Dicho de otra forma, es la traducción de la consulta del usuario a un lenguaje documental ó técnico, propio de la recuperación de información.

- **Expansión de consulta.** Es un proceso de reformulación automática del sistema que permite añadir nuevos términos a la consulta para mejorar el contexto de la consulta original del usuario. Esto se consigue mediante procesos de clustering, que determinan la frecuencia de aparición de un grupo de términos contiguos, relacionados con la consulta del usuario, presentes en documentos clasificados dentro de un mismo ámbito temático (en el caso de análisis del contexto local) y en torno a toda la colección (en el caso de análisis del contexto global).

- **Patrón.** Expresión sintáctica que define una serie de caracteres textuales, alfabéticos, numéricos y especiales, que se ajustarán por coincidencia en una palabra o término de un texto determinado.

- **Expresión regular.** También conocidas como REGEXP y POSIX, son aquellas expresiones sintácticas complejas y normalizadas, compuestas a base de patrones que permiten la definición de consultas de datos en un corpus documental dado, mediante cadenas de caracteres, repeticiones y concatenaciones, establecidas por sus reglas de construcción. (Véase LEVITHAN, S.; GOYVAERTS, J. 2009. Regular Expressions Cookbook. OReilly. Disponible en: http://www.bookf.net/p/3844)

La base de conocimiento

- **Colección.** Es sinónimo de base de conocimiento, fondo, biblioteca de documentos ó corpus documental. El concepto colección hace referencia a un compendio de documentos seleccionados previamente u obtenidos mediante métodos de minería de datos ó webcrawling.

- **Colección de referencia.** Aquella colección utilizada para la experimentación de los modelos de recuperación de información y sus algoritmos. Ello implica la disposición de plantillas de resultados con los documentos relevantes para cada consulta de prueba, de cara a la evaluación del SRI.

- **Documento.** Elemento básico con el que se conforman las colecciones y unidad básica de recuperación. Se considera documento a todo tipo de información independiente, artículos, monografías, sitios y páginas web, resúmenes, textos completos, etc.

- **Documento sustituto.** Símil de un documento de una colección, fiel a sus contenidos mediante sus elementos básicos como título, resumen, frase de contextualización y URL. Se utiliza en las páginas de resultados, en procesos de visualización y representación.

- **TREC.** Una de las colecciones de referencia más importantes a nivel internacional que contiene más de un millón de documentos y que se ha utilizado ampliamente por especialistas en recuperación de la información, en las conferencias TREC (Text REtrieval Conference. Disponible en: http://trec.nist.gov/). La colección TREC ha sido desarrollada por el NIST (National Institute of Standards and Technology. Disponible en: http://www.nist.gov/) y se ha convertido en un estándar para la comparación de modelos y algoritmos de recuperación.

Evaluación y resultados de la recuperación

- **Precisión.** En recuperación de información, precisión es la medida que define cuantitativamente la relación entre los documentos recuperados y su relevancia para satisfacer la consulta del usuario.

- **Exhaustividad.** También denominado Recall es la capacidad del sistema de información para recuperar todos los documentos relevantes con respecto a la totalidad de los existentes en la colección, de acuerdo a los condicionamientos y especificaciones de la consulta del usuario.

- **Pertinencia.** Aquel documento que añade nueva información a la previamente almacenada en la mente del usuario, que le resulta útil en el trabajo que ha propiciado la pregunta. FOSKETT, D.J. 1972. A note on the concept of

relevance. Information Storage and Retrieval, vol.8, nº2. pp 77-78. El conjunto pertinente de documentos recuperados puede definirse como el subconjunto de los documentos almacenados en el sistema que es apropiado para la necesidad de información del usuario. SALTON, G. 1983. Introduction to modern information retrieval. Mc Graw Hill.

- **Relevancia.** Un mismo documento puede ser considerado relevante, o no relevante, por dos personas distintas en función de los motivos que producen la necesidad de información o del grado de conocimiento que sobre la materia posean ambos. Llegados a un caso extremo, un mismo documento puede parecer relevante o no a la misma persona en momentos diferentes de tiempo. LANCASTER, F.W.; WARNER, A. J. 1993. Information Retrieval Today. Information Resources. Aunque puede usarse otra terminología, la voz relevancia parece la más apropiada para indicar la relación entre un documento y una petición de información efectuada por un usuario, aunque puede resultar erróneo asumir que ese grado de relación es fijo e invariable, siendo mejor decir, que un documento ha sido juzgado como relevante a una específica petición de información. LANCASTER, F.W.; WARNER, A. J. 1993. Information Retrieval Today. Information Resources. Es el grado de importancia y significación que concede el usuario a los resultados obtenidos en un sistema de información. BLÁZQUEZ, M. 2012 [Asignatura]. Técnicas avanzadas de recuperación de información

- **Rendimiento.** Es un factor para la evaluación de un sistema de recuperación de información, que se obtiene evaluando la pertinencia y exhaustividad de los resultados generados por un conjunto de consultas de prueba en la colección de referencia, con respecto a las soluciones propuestas para el mismo por los especialistas.

- **Ranking.** También denominado alineado de los documentos es el proceso de evaluación de los resultados obtenidos, tras aplicar un modelo de recuperación de información, reflejando en un coeficiente ó indicador numérico la relevancia, precisión y exhaustividad de los mismos, para una consulta dada por el usuario.

Sistema de recuperación de información

- **Tarea de recuperación.** Aquellas rutinas algorítmicas ejecutadas por el sistema de información en respuesta a una solicitud del usuario. BAEZA YATES, R.; RIBEIRO-NETO, B. 1999. Modern Information Retrieval. Addison Wesley

- **Algoritmo de recuperación.** Es el conjunto de métodos documentales, rutinas de tratamiento de información y procedimientos automáticos de tipo matemático-estadístico, ya predefinido en el funcionamiento de un programa informático, tales como la depuración, indexación, comparación de consultas, aplicación de modelos de recuperación, representación, evaluación y análisis necesarios para que el sistema de información satisfaga las necesidades de información del usuario. El orden en que se ejecutan, la casuística de la consulta y la experiencia del usuario, son factores que influyen en la ejecución de los algoritmos de recuperación, generando un grado de variabilidad en los resultados obtenidos.

- **Filtrado.** Proceso de refinamiento y perfección de la consulta del usuario por el que se delimita, especifica ó amplía la búsqueda original, una vez que los resultados de la búsqueda satisfacen parcialmente la demanda informativa del usuario.

- **Coincidencia exacta.** Es el mecanismo por el cual sólo los documentos que satisfacen algunos criterios y rasgos bien especificados en la consulta son recuperados y devueltos al usuario como una respuesta unívoca, cumpliéndose al 100% en sus expectativas.

- **Recuperación de datos.** La recuperación de elementos (tuplas, los objetos, páginas Web, documentos) cuyo contenido cumple los requisitos especificados en una consulta de usuario basada en expresión regular ó por coincidencia de patrones. BAEZA YATES, R.; RIBEIRO-NETO, B. 1999. Modern Information Retrieval. Addison Wesley

- **Minería de datos.** La extracción de nuevos datos, documentos ó información parcial de cualquier tipo, mediante métodos de crawling. BAEZA YATES, R.; RIBEIRO-NETO, B. 1999. Modern Information Retrieval. Addison Wesley

- **Clustering.** Es la agrupación de documentos que satisfagan un conjunto de propiedades comunes. El objetivo es aunar aquellos documentos que están relacionados entre sí. El Clustering puede ser utilizado, por ejemplo, para expandir una consulta de usuario con nuevos términos propios del contexto de los documentos recuperados. BAEZA YATES, R.; RIBEIRO-NETO, B. 1999. Modern Information Retrieval. Addison Wesley

4. Génesis del desarrollo de la colección: el proceso de crawling

El proceso de crawling resulta de gran importancia para generar la colección o lo que es lo mismo la base de conocimiento para el sistema de recuperación de información. Ello es debido no sólo a la extracción de los textos obtenidos a través de la web, sino por la capacidad de diferenciar distintos tipos de contenidos en las páginas que se encuentran altamente estructurados. Éstos son los enlaces propiamente dichos, los enlaces a documentos ofimáticos, los enlaces a documentos audiovisuales, los enlaces a documentos gráficos, los metadatos, las meta-etiquetas, los títulos y titulares de la página, los párrafos de la misma, los canales de sindicación, el texto depurado y el código fuente original. Para obtener todos los elementos de una página web y almacenarlos adecuadamente se emplea un esquema similar al que se propone en la *figura2*.

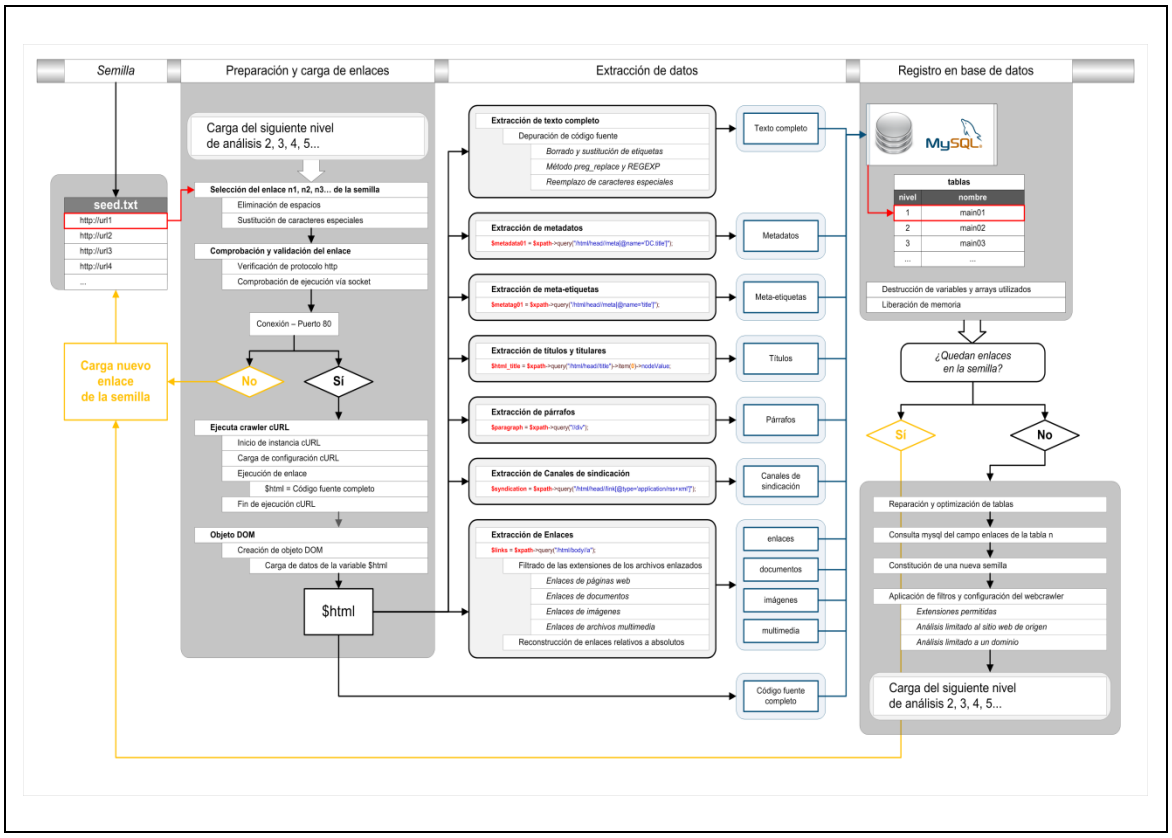


Figura 2. El proceso de crawling

Se muestra el proceso real de un programa webcrawler que toma las direcciones URL especificadas en el archivo de *semilla* o *seed.txt*. Para cada enlace, el programa

comprueba la validez del mismo, verificando que se encuentra activo mediante su ejecución vía socket con el puerto 80. Si el enlace no funciona, está roto o no se encuentra el contenido, el sistema lo deshecha, seleccionando un nuevo enlace de la semilla. En caso contrario, el webcrawler ejecuta una serie de funciones de tipo cURL, que permiten descargar el código fuente de la página del enlace. Dicho código fuente se encuentra en formato HTML y es almacenado en una variable que permita posteriormente la manipulación y extracción de los contenidos depositados dentro de ella. En este sentido seleccionar la información de todo el código fuente puede realizarse de dos formas distintas. En primer lugar, empleando técnicas de expresiones regulares REGEXP que conlleva una mayor complejidad. En segunda lugar usando la técnica DOM (Document Object Model) que permite crear un árbol jerárquico de la estructura de etiquetas HTML de la página web que fue almacenada en la variable. Ello permite, la extracción de metadatos, meta-etiquetas, enlaces, párrafos, titulares y demás elementos a través de un lenguaje de consulta, filtrado y selección denominado XPath. Se observará que XPath permite indicar una ruta de nodos, que en este caso corresponden a las etiquetas y su anidamiento hasta la obtención de los valores contenidos dentro de ellas e incluso en sus atributos. De esta forma, se distinguen todos sus elementos de forma ordenada e independiente, para ser finalmente procesados por la base de datos del sistema. Dado que se trata de un webcrawler, el proceso no acaba hasta que finalizan todos los enlaces de la semilla y posteriormente hasta que no se analizan los enlaces derivados de cada enlace de la semilla, hasta que se alcance el nivel de profundidad tope para el análisis, tal como fuere configurado el programa.

5. Mecanismos de depuración de páginas web para la extracción y procesamiento de textos

Cuando se generan colecciones de documentos basados en páginas web, los programas webcrawler efectúan un proceso de extracción del código fuente, mediante el empleo técnicas cURL y DOM. Al hacerlo no sólo se adquiere el texto sujeto a recuperación, sino todo el conjunto de etiquetas en formato HTML y CSS que lo acompañan. Si tales etiquetados no son eliminados, no se puede iniciar el procesamiento de la información y su correspondiente tratamiento. Por ello, se demuestra la importancia de aplicar mecanismos de depuración del código fuente, que facilite la extracción limpia de los textos, que serán la materia prima con la que se componen las colecciones sobre las que se recupera la información. Para poder comprender la multitud de procesos que se llevan a cabo, se muestra la siguiente *tabla1* en la que puede ver un orden en la consecución de los mismos.

Preparación de la colección de documentos
NOTA: Preparar los documentos implica en esencia los procesos destacados en esta tabla. No obstante, pueden aplicarse otros derivados del procesamiento del lenguaje natural, analizadores sintácticos para el reconocimiento de las oraciones del texto, identificación de la naturaleza de los complementos y sintagmas de las frases, etc. Tales funciones se situarían entre el proceso de Normalización de textos y el de indexación, puesto que antes de proceder a su almacenamiento, se requeriría del análisis semántico-sintagmático correspondiente.
Proceso de depuración
Depuración y supresión de código fuente Identificación de casos especiales
Normalización de textos
Tokenización Conversión a minúsculas, eliminación de signos de puntuación y acentos Eliminación de palabras vacías Transliteración y reemplazo de caracteres especiales
Indexación
Reducción morfológica Almacenamiento Fichero inverso Fichero índice

Tabla 1. Procesos para la preparación de los documentos

Los datos

¿Cómo eliminar todo lo que no sea el texto de un artículo? ¿Cómo extrare los bloques de contenidos? ¿Cómo evitar el problema del anidamiento de etiquetas en HTML? ¿Cómo detectar el inicio y el final del texto? Éstas son algunas de las preguntas más habituales cuando se trata de analizar el contenido textual de un sitio web. Para tratar de enfocar y solucionar estos problemas previamente es necesario considerar los siguientes aspectos:

- Qué tipo de documento web se pretende explotar.
- Qué nivel de anidamiento tiene el código fuente.
- Qué etiquetas contienen la información relevante.
- Que fragmentos de código corresponden al interfaz del sitio web.

Depuración y supresión de códigos

Para efectuar un proceso de depuración y limpieza del contenido, es preciso identificar qué bloque de HTML contiene la información central del documento. En esencia el webcrawler debe ser capaz de determinar qué elementos corresponden al interfaz visual, la navegación hipertextual, los menús, la publicidad, los banners, así como cualquier otro elemento accesorio que no forma parte del contenido. De una forma visual se podría explicar con la siguiente *figura3*.

Figura 3. El contenido útil es el artículo propiamente dicho y no la interfaz de la página

Por regla general, cuando se trata de depurar el texto central de una página web, el webcrawler actúa dentro del dominio del cuerpo del sitio web *<body></body>*, eliminando todas las etiquetas *<table></table>*, *<iframe></iframe>*. A esta primera medida se le suman otrás muchas relativas, a la identificación de contenidos esenciales, como por ejemplo el encabezamiento, el resumen, los párrafos de contenido, cláramente identificados por etiquetas *<h1></h1>*, *<h2></h2>*, *<h3></h3>*, * *, *<blockquote></blockquote>*, *<cite></cite>*. Aún así la ideosincrasia y variedad de las páginas web, obligan a un reconocimiento de casos, estilos. Este es el caso del empleo de la etiqueta * * que corresponde a las capas en HTML y que pueden adquirir cualquier estilo en CSS que por ejemplo imite el resultado de una etiqueta de encabezado *<h1></h1>*. Éste tipo de situaciones son muy comúnes y hacen que el webcrawler contemple el análisis de los estilos con que se editan las informaciones. Además existe el problema añadido de utilizar recursivamente el anidamiento de distintos elementos en HTML, lo que dificulta aún más si cabe, el proceso de depuración, obsérvese la siguiente *tabla2*.

Ejemplo de anidamiento masivo

```
<div style="margin-bottom:1em">
    <div>
        <table style="width: 100%; height: 22px;" border="0">
            <tbody>
                <tr><td>
                <ul>
         <li><div align="left"><a href="…">Segundo ciclo:&amp;nbsp;Licenciatura</a></div></li>
         <li><div align="left"><a href="…" target="_blank">Titulaciones a extinguir y sus pasarelas</a></div></li>
                </ul>
                </td>
                <td> <img src="…"/><iframe src="…" class='framestyle'></iframe></td>
                <td>
                <ul>
          <li><a href="…" target="_blank"><strong>Informaci&amp;oacute;n sobre los nuevos estudios</strong></a></li>
                 </ul>
                </td></tr>
            </tbody>
        </table>
    </div>
</div>
```

Tabla 2. El excesivo anidamiento de una página web puede dificultar los procesos de depuración

Como se puede comprobar, las etiquetas div embeben distintos contenidos, entre ellos tablas, marcos, imágenes, enlaces... en distintos niveles de anidamiento, tal como se demuestra en el sangrado de los elementos. Este caso exige un tratamiento del etiquetado por medio de patrones reconocibles para su detección y extracción. Además existen otras dificultades añadidas. Se trata del etiquetado sin normalización, en desuso,

obsoleto ó especial. Una muestra de ello son algunas de las etiquetas que se muestran en la siguiente *tabla3*.

Muestra de etiquetado especialmente difícil de eliminar. *(Exige recopilar caso a caso)*		
 <di> <table> <tbody> <noscript> <font> <base> </+script> <fb:fan>	<map> <na> <small> <time> <noindex> <nobr> <o:p> <fb:like> <asx>	<basefont> <fieldset> <legend> <base> <sup> <wbr> <area> <layer> <spacer>

Tabla 3. Muestra de etiquetado complejo

Tokenización

Es el proceso que descompone los textos de una colección en sus únidades mínimas, las palabras o términos propiamente dichos. A tales elementos se les denomina "tokens" que conforman una lista de ítems que se utiliza para su análisis PNL (Procesamiento del lenguaje natural), estadístico, lingüístico, almacenamiento y posterior recuperación de información. Para llevar a cabo tal proceso, se utilizan los espacios entre las palabras del texto como divisores de los distintos "tokens". Véase un ejemplo de ello en la siguiente *tabla4*.

Ejemplo de tokenización			
Institución cuya finalidad consiste en la adquisición, conservación, estudio y exposición de libros y documentos. Local donde se tiene considerable número de libros ordenados para la lectura.			
Tokens resultantes			
Institución cuya finalidad consiste en la adquisición,	conservación, estudio y exposición de libros y	documentos. Local donde se tiene considerable número	de libros ordenados para la lectura.
Conversión de tokens en código hexadecimal			
{Institución} 49 6E 73 74 69 74 75 63 69 1 6E	{conservación,} 63 6F 6E 73 65 72 76 61 63 69 1 6E 2C	{documentos.} 64 6F 63 75 6D 65 6E 74 6F 73 2E	{de} 64 65 {libros}

{cuya} 63 75 79 61 {finalidad} 66 69 6E 61 6C 69 64 61 64 {consiste} 63 6F 6E 73 69 73 74 65 {en} 65 6E {la} 6C 61 {adquisición,} 61 64 71 75 69 73 69 63 69 1 6E 2C	{estudio} 65 73 74 75 64 69 6F {y} 79 {exposición} 65 78 70 6F 73 69 63 69 1 6E {de} 64 65 {libros} 6C 69 62 72 6F 73 {y} 79	{Local} 4C 6F 63 61 6C {donde} 64 6F 6E 64 65 {se} 73 65 {tiene} 74 69 65 6E 65 {considerable} 63 6F 6E 73 69 64 65 72 61 62 6C 65 {número} 6E 1 6D 65 72 6F	6C 69 62 72 6F 73 {ordenados} 6F 72 64 65 6E 61 64 6F 73 {para} 70 61 72 61 {la} 6C 61 {lectura.} 6C 65 63 74 75 72 61 2E

Tabla 4. Proceso de tokenización

Los "tokens" a su vez pueden ser identificados mediante una codificación ASCII ó en su defecto Hexadecimal, con el objeto de facilitar la identificación uno a uno cada caracter que compone la palabra. De hecho este proceso permite la identificación de cadenas de caracteres de forma unívoca, de cara a posteriores tratamientos de depuración, eliminación de signos de puntuación ó la reducción morfológica. Obsérvese en la tabla anterior, que algunos términos vienen acompañados de signos de puntuación, que no son de utilidad para el proceso de recuperación.

Conversión a minúsculas, eliminación de signos de puntuación y acentos

Para permitir un proceso de indexación limpio, previamente es necesario convertir el texto a minúsculas y eliminar todos los signos de puntuación del texto. De hecho este punto de la depuración simplifica a posteriori el reconocimiento y proceso de cruzado de la consulta del usuario, permitiendo que independientemente de que escribiera su consulta correctamente, pueda ser recuperada bajo cualquier circunstancia. De esta forma el sistema de recuperación identifica mediante codificación hexadecimal qué caracteres debe reemplazar, modificar o en su defecto eliminar. Es habitual que tales programas integren la referencia completa de caracteres, denominadas también como "tablas de equivalencias entre caracteres". En la siguiente *tabla5* se muestra un sencillo ejemplo del concepto con los caracteres habituales.

Hex	Representación	Hex	Representación	Hex	Representación
20	espacio ( )	40	@	60	`
21	!	41	A	61	a
22	"	42	B	62	b
23	#	43	C	63	c
24	$	44	D	64	d
25	%	45	E	65	e
26	&	46	F	66	f
27	'	47	G	67	g
28	(	48	H	68	h
29	)	49	I	69	i
2A	*	4A	J	6A	j
2B	+	4B	K	6B	k
2C	,	4C	L	6C	l
2D	-	4D	M	6D	m
2E	.	4E	N	6E	n
2F	/	4F	O	6F	o
30	0	50	P	70	p
31	1	51	Q	71	q
32	2	52	R	72	r
33	3	53	S	73	s
34	4	54	T	74	t
35	5	55	U	75	u
36	6	56	V	76	v
37	7	57	W	77	w
38	8	58	X	78	x
39	9	59	Y	79	y
3A	:	5A	Z	7A	z

3B	;		5B	[		7B	{
3C	<		5C	\		7C	|
3D	=		5D	]		7D	}
3E	>		5E	^		7E	~
3F	?		5F	_			

Tabla 5. Tabla de conversión de caracteres básicos

Si bien todos los caracteres representados en la tabla anterior tienen una codificación sencilla de dupla, en el caso de los caracteres acentuados (siempre que se parta de una códificación utf-8 previa por defecto en la página web) se utilizan 2 duplas para identificar que es un caracter acentuado. Véase la siguiente *tabla6*.

Hex	Representación		Hex	Representación		Hex	Representación
C3 A1	á		C3 AD	í		C3 BA	ú
C3 81	Á		C3 8D	Í		C3 9A	Ú
C3 A0	à		C3 AC	ì		C3 B9	ù
C3 80	À		C3 8C	Ì		C3 99	Ù
C3 A2	â		C3 AE	î		C3 BB	û
C3 82	Â		C3 8E	Î		C3 9B	Û
C3 A4	ä		C3 AF	ï		C3 BC	ü
C3 84	Ä		C3 8F	Ï		C3 9C	Ü
C3 A9	é		C3 B3	ó			
C3 89	É		C3 93	Ó			
C3 A8	è		C3 B2	ò			
C3 88	È		C3 92	Ò			
C3 AA	ê		C3 B4	ô			
C3 8A	Ê		C3 94	Ô			
C3 AB	ë		C3 B6	ö			
C3 8B	Ë		C3 96	Ö			

Tabla 6. Tabla de conversión de vocales acentuadas

Transliteración y reemplazo de caracteres especiales

En muchos casos, la conversión de un texto a minúsculas, la supresión de acentos y signos de puntuación no es suficiente. En muchos casos el idioma en el que está escrito el documento o las particularidades del texto implican el uso de caracteres especiales que requieren transliteración o reemplazo por otro caracter más sencillo y equivalente en el teclado estándar. Estos procesos resultan complejos, puesto que en todo momento se debe asegurar la identificación del caracter original. En la *tabla7*, se pueden observar todos los caracteres sometidos a tal consideración de excepcionalidad.

Hex	Representación	Hex	Representación	Hex	Representación
C3 80	À	C4 86	Ć	C5 88	ň
C3 81	Á	C4 87	ć	C5 89	ŉ
C3 82	Â	C4 88	Ĉ	C5 8C	Ō
C3 83	Ã	C4 89	ĉ	C5 8D	ō
C3 84	Ä	C4 8A	Ċ	C5 8E	Ŏ
C3 85	Å	C4 8B	ċ	C5 8F	ŏ
C3 86	Æ	C4 8C	Č	C5 90	Ő
C3 87	Ç	C4 8D	č	C5 91	ő
C3 88	È	C4 8E	Ď	C5 92	Œ
C3 89	É	C4 8F	ď	C5 93	œ
C3 8A	Ê	C4 90	Đ	C5 94	Ŕ
C3 8B	Ë	C4 91	đ	C5 95	ŕ
C3 8C	Ì	C4 92	Ē	C5 96	Ŗ
C3 8D	Í	C4 93	ē	C5 97	ŗ
C3 8E	Î	C4 94	Ĕ	C5 98	Ř
C3 8F	Ï	C4 95	ĕ	C5 99	ř
C3 90	Ð	C4 96	Ė	C5 9A	Ś
C3 91	Ñ	C4 97	ė	C5 9B	ś
C3 92	Ò	C4 98	Ę	C5 9C	Ŝ
C3 93	Ó	C4 99	ę	C5 9D	ŝ

C3 94	Ô		C4 9A	Ě		C5 9E	Ş	
C3 95	Õ		C4 9B	ě		C5 9F	ş	
C3 96	Ö		C4 9C	Ĝ		C5 A0	Š	
C3 98	Ø		C4 9D	ĝ		C5 A1	š	
C3 99	Ù		C4 9E	Ğ		C5 A2	Ţ	
C3 9A	Ú		C4 9F	ğ		C5 A3	ţ	
C3 9B	Û		C4 A0	Ġ		C5 A4	Ť	
C3 9C	Ü		C4 A1	ġ		C5 A5	ť	
C3 9D	Ý		C4 A2	Ģ		C5 A6	Ŧ	
C3 9F	ß		C4 A3	ģ		C5 A7	ŧ	
C3 A0	à		C4 A4	Ĥ		C5 A8	Ũ	
C3 A1	á		C4 A5	ĥ		C5 A9	ũ	
C3 A2	â		C4 A6	Ħ		C5 AA	Ū	
C3 A3	ã		C4 A7	ħ		C5 AB	ū	
C3 A4	ä		C4 A8	Ĩ		C5 AC	Ŭ	
C3 A5	å		C4 A9	ĩ		C5 AD	ŭ	
C3 A6	æ		C4 AA	Ī		C5 AE	Ů	
C3 A7	ç		C4 AB	ī		C5 AF	ů	
C3 A8	è		C4 AC	Ĭ		C5 B0	Ű	
C3 A9	é		C4 AD	ĭ		C5 B1	ű	
C3 AA	ê		C4 AE	Į		C5 B2	Ų	
C3 AB	ë		C4 AF	į		C5 B3	ų	
C3 AC	ì		C4 B0	İ		C5 B4	Ŵ	
C3 AD	í		C4 B1	ı		C5 B5	ŵ	
C3 AE	î		C4 B2	IJ		C5 B6	Ŷ	
C3 AF	ï		C4 B3	ij		C5 B7	ŷ	
C3 B1	ñ		C4 B4	Ĵ		C5 B8	Ÿ	
C3 B2	ò		C4 B5	ĵ		C5 B9	Ź	
C3 B3	ó		C4 B6	Ķ		C5 BA	ź	

C3 B4	ô		C4 B7	ķ		C5 BC	ż	
C3 B5	õ		C4 B9	Ĺ		C5 BD	Ž	
C3 B6	ö		C4 BA	Í		C5 BE	ž	
C3 B8	ø		C4 BC	ļ		C5 BF	ſ	
C3 B9	ù		C4 BD	Ľ		C6 92	ƒ	
C3 BA	ú		C4 BE	ľ		C6 A0	Ơ	
C3 BC	ü		C4 BF	Ŀ		C6 A1	ơ	
C3 BD	ý		C5 80	l·		C6 AF	Ư	
C3 BF	ÿ		C5 81	Ł		C6 B0	ư	
C4 80	Ā		C5 82	ł		C7 BA	Ǻ	
C4 81	ā		C5 83	Ń		C7 BC	Ǽ	
C4 82	Ă		C5 84	ń		C7 BD	ǽ	
C4 83	ă		C5 85	Ņ		C7 BE	Ǿ	
C4 84	Ą		C5 86	ņ		C7 BF	ǿ	
C4 85	ą		C5 87	Ň				

Tabla 7. Tabla de caracteres especiales (En muchos casos se requiere transliteración)

Eliminación de palabras vacías

Las palabras vacías, irrelevantes o "stop words" son aquellas que por si solas carecen de significación y que por su altísima frecuencia de aparición en los textos, generan un ruido innecesario para la recuperación de información. La eliminación de estos términos (preposiciones, artículos determinados, artículos indeterminados, pronombres, conjunciones, contracciones y ciertos verbos y adverbios) mejora la afinación en los modelos de recuperación. Los estudios correspondientes a este fenómeno fueron iniciados por Hans Peter Luhn en 1958 con su investigación sobre el índice KWIC, una técnica de indexación que organizaba las palabras según su consideración como claves para la recuperación o no de la información, teniendo en cuenta el contexto del documento. Este proceso derivó en la acuñación del término "palabra vacía" para referirse a aquellas con un bajo poder discriminatorio y representativo del contenido del documento. Los análisis estadísticos efectuados por Luhn, demostraron que la indexación era un proceso más rápido, cuando se prescinde de tales términos y favoreciendo la economía de espacio requerido para el almacenamiento de la

información. También se demostró, que entre un 30 y un 50% de las palabras de un texto corresponden a tal categoría. De hecho y pese a ser práctica habitual (RIJSBERGEN, C.J. 1979), hasta nuestros días, se siguen utilizando listas de palabras vacías para la depuración de los textos. No obstante, la técnica de eliminación de palabras vacías, se viene suavizando, debido a la introducción de técnicas de PNL (Procesamiento del Lenguaje Natural) que tienen en cuenta la significación de tales palabras cuando están acompañadas de sustantivos, en casos en los que no pueden ser separadas ó eliminadas por conformar una denominación propia, así como por pérdidas en la significación semántica de un sintagma, frase ó palabra, véase la *tabla8*.

Palabras vacías, claves		
Frase	**Palabras vacías**	**Descripción del caso**
Those were the days	those, were, the	La frase corresponde al título de una canción de Boris Fomin
Es así o de esta otra manera	es, así, o, de, esta, otra	Título de un artículo sobre estilo gramatical
De aquí a la eternidad	de, aquí, a, la	Película de 1953 del director Fred Zinnemann
Cómo ha de ser el privado	cómo, ha, de, ser, el	Comedia teatral de Francisco de Quevedo
El otro	el, otro	Obra de teatro de Miguel de Unamuno

Tabla 8. Ejemplo del uso de palabras vacías cuya función identificadora es clave

A continuación se muestran listados reales de palabras vacías utilizados para la depuración de textos en alemán, español, francés, inglés, italiano y portugués.

Palabras vacías del alemán. Disponible en: http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/stop-words_german.txt

ab, bei, da, deshalb, ein, für, haben, hier, ich, ja, kann, machen, muesste, nach, oder, seid, sonst, und, vom, wann, wenn, wie, zu, bin, eines, hat, manche, solches, an, anderm, bis, das, deinem, demselben, dir, doch, einig, er, eurer, hatte, ihnen, ihre, ins, jenen, keinen, manchem, meinen, nichts, seine, soll, unserm, welche, werden, wollte, während, alle, allem, allen, aller, alles, als, also, am, ander, andere, anderem, anderen, anderer, anderes, andern, anders, auch, auf, aus, bist, bsp., daher, damit, dann, dasselbe, dazu, daß, dein, deine, deinen, deiner, deines, dem, den, denn,

denselben, der, derer, derselbe, derselben, des, desselben, dessen, dich, die, dies, diese, dieselbe, dieselben, diesem, diesen, dieser, dieses, dort, du, durch, eine, einem, einen, einer, einige, einigem, einigen, einiger, einiges, einmal, es, etwas, euch, euer, eure, eurem, euren, eures, ganz, ganze, ganzen, ganzer, ganzes, gegen, gemacht, gesagt, gesehen, gewesen, gewollt, hab, habe, hatten, hin, hinter, ihm, ihn, ihr, ihrem, ihren, ihrer, ihres, im, in...

Tabla 9. Muestra de palabras vacías del alemán

Palabras vacías del español. Disponible en: http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/stop-words_spanish.txt

el, la, los, les, las, de, del, a, ante, con, en, para, por, y, o, u, tu, te, ti, le, que, al, ha, un, han, lo, su, una, estas, esto, este, es, tras, suya, a, acá, ahí, ajena, ajenas, ajeno, ajenos, al, algo, algún, alguna, algunas, alguno, algunos, allá, alli, allí, ambos, ampleamos, ante, antes, aquel, aquella, aquellas, aquello, aquellos, aqui, aquí, arriba, asi, atras, aun, aunque, bajo, bastante, bien, cabe, cada, casi, cierta, ciertas, cierto, ciertos, como, cómo, con, conmigo, conseguimos, conseguir, consigo, consigue, consiguen, consigues, contigo, contra, cual, cuales, cualquier, cualquiera, cualquieras, cuancuán, cuando, cuanta, cuánta, cuantas, cuántas, cuanto, cuánto, cuantos, cuántos, de, dejar, del, demás, demas, demasiada, demasiadas, demasiado, demasiados, dentro, desde, donde, dos, el, él, ella, ellas, , ello, ellos, empleais, emplean, emplear, empleas, empleo, en, encima, entonces, entre, era, eramos...

Tabla 10. Muestra de palabras vacías del español

Palabras vacías del francés. Disponible en: http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/stop-words_french.txt

a, adieu, afin, ah, ai, aie, aient, aies, aille, ainsi, ait, all, alla, allais, allait, allant, alle, aller, allerent, allez, allons, alors, apres, aprčs, as, assez, au, au dela, au delŕ, au dessous, au dessus, aucun, aucune, aucunes, aucuns, aupres, auprčs, auquel, aura, aurai, aurais, aurez, auront, aussi, aussitôt, autant, autour, autre, autres, autrui, aux, auxquelles, auxquels, av, avaient, avais, avait, aval, avant, avec, avez, avoir, avons, ayant, ayez, ayons, bah, bas, beaucoup, bien, bonté, bout, but, c, cest a dire, cest ŕ dire, ca, car, ce, ceci, cela, celle, celle ci, celle la, celle lŕ, celles, celles ci, celles la, celles lŕ, celui, celui ci, celui la, celui lŕ, cependant, ces, cet, cette, ceux, ceux ci, ceux la, ceux lŕ, chacun, chacune, chaque, chez, chut, ci, circa, combien, comme, comment, commme, compte, contre, crac, crainte, côtč, d, d, dans, de, deca, dedans, dehors, dela, delŕ, depuis, des, desquelles, desquels, dessous, dessus, devant, deçŕ, dire, divers, diverses, donc, dont, du, duquel, durant, dčs...

Tabla 11. Muestra de palabras vacías del francés

Palabras vacías del inglés. Disponible en: http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/stop-words_english.txt

a, about, above, above, across, after, afterwards, again, against, all, almost, alone, along, already, also, although, always, am, among, amongst, amoungst, amount, an, and, another, any, anyhow, anyone, anything, anyway, anywhere, are, around, as, at, back, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, below, beside, besides, between, beyond, bill, both, bottom, but, by, call, can, cannot, cant, co, con, could, couldnt, cry, de, describe, detail, do, done, down, due, during, each, eg, eight, either, eleven, else, elsewhere, empty, enough, etc, even, ever, every, everyone, everything, everywhere, except, few, fifteen, fify, fill, find, fire, first, five, for, former, formerly, forty, found, four, from, front, full, further, get, give, go, had, has, hasnt, have, he, hence, her, here, hereafter, hereby, herein, hereupon, hers, herself, him, himself, his, how, however, hundred, ie, if...

Tabla 12. Muestra de palabras vacías del inglés

Palabras vacías del italiano. Disponible en: http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/stop-words_italian.txt

ad, al, allo, ai, agli, all, agl, alla, alle, con, col, coi, da, dal, dallo, dai, dagli, dall, dagl, dalla, dalle, di, del, dello, dei, degli, dell, degl, della, delle, in, nel, nello, nei, negli, nell, negl, nella, nelle, su, sul, sullo, sui, sugli, sull, sugl, sulla, sulle, per, tra, contro, io, tu, lui, lei, noi, voi, loro, mio, mia, miei, mie, tuo, tua, tuoi, tue, suo, sua, suoi, sue, nostro, nostra, nostri, nostre, vostro, vostra, vostri, vostre, mi, ti, ci, vi, lo, la, li, le, gli, ne, il, un, uno, una, ma, ed, se, perché, anche, come, dov, dove, che, chi, cui, non, più, quale, quanto, quanti, quanta, quante, quello, quelli, quella, quelle, questo, questi, questa, queste, si, tutto, tutti, a, c, e, i, l, o, ho, hai, ha, abbiamo, avete, hanno, abbia, abbiate, abbiano, avrò, avrai, avrà, avremo, avrete, avranno, avrei, avresti, avrebbe, avremmo, avreste, avrebbero, avevo, avevi, aveva, avevamo, avevate, avevano, ebbi, avesti, ebbe, avemmo, aveste, ebbero, avessi, avesse, avessimo, avessero, avendo, avuto, avuta...

Tabla 13. Muestra de palabras vacías del italiano

Palabras vacías del portugués. Disponible en: http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/stop-words_spanish.txt

de, a, o, que, e, do, da, em, um, para, é, com, não, uma, os, no, se, na, por, mais, as, dos, como, mas, foi, ao, ele, das, tem, à, seu, sua, ou, ser, quando, muito, há, nos, já, está, eu, também, só, pelo, pela, até, isso, ela, entre, era, depois, sem, mesmo, aos, ter, seus, quem, nas, me, esse, eles, estão, você, tinha, foram, essa, num, nem, suas, meu, às, minha, têm, numa, pelos, elas, havia, seja, qual, será, nós, tenho, lhe, deles, essas, esses, pelas, este , fosse, dele, tu, te, vocês, vos, lhes, meus, minhas, teu, tua, teus, tuas,

nosso, nossa, nossos, nossas, dela, delas, esta, estes, estas, aquele, aquela, aqueles, aquelas, isto, aquilo, estou, está, estamos, estão, estive, esteve, estivemos, estiveram, estava, estávamos, estavam, estivera, estivéramos, esteja, estejamos, estejam, estivesse, estivéssemos, estivessem, estiver, estivermos, estiverem, hei, há, havemos, hão, houve, houvemos, houveram, houvera, houvéramos, haja, hajamos, hajam, houvesse, houvéssemos, houvessem, houver, houvermos, houverem, houverei...

Tabla 14. Muestra de palabras vacías del portugués

6. El proceso de indexación

Indexar es la acción de construir un fichero inverso de forma automática ó manual. Este proceso es necesario para localizar y recuperar rápidamente cada uno de los términos del texto de un documento. Esto significa que a cada palabra se le asigna un identificador del documento en el que aparece, un indicador de la posición que ocupa en el texto (párrafo, línea, número de caracter de inicio) y un número de identificación para ese término propiamente dicho (único e irrepetible). De esta forma se conoce la posición exacta de cada término en los documentos de la colección y posibilita el posterior análisis de frecuencias.

Para llevar a cabo la indexación, tal como apunta Baeza Yates, se requiere de una importante capacidad de computación y por ende existe gran dependencia de las prestaciones del hardware para llevarlo a cabo. Si el corpus documental es la propia red WWW, se necesitará un motor de indexación distribuida. Eso significa que existen decenas de equipos informáticos (servidores), trabajando con una agrupación de contenidos determinada, colaborando en común para generar un gran índice que será utilizado a la postre por el motor de búsqueda. La indexación, también deberá ser dinámica. Esta propiedad significa que ante un corpus cambiante como puede ser la web, se necesitan reindexaciones continuas de los contenidos y por ende su modificación y ampliación.

Todo ello justifica que junto al proceso de indexación, se trate de reducir al máximo el tamaño de tales archivos o tablas de la base de datos para conseguir la mejor relación entre tiempo de ejecución de las consultas y exhaustividad del fichero inverso. A esta misión se la denomina "*compresión de la indexación*" y en ella se circunscriben los procesos de depuración que se han mostrado en el artículo anterior, tales como la supresión de palabras vacías, la normalización de palabras, la transliteración de caracteres especiales y la reducción morfológica o "*stemming*" que se abordará en este capítulo.

La compresión de la indexación

- Depuración de los textos de los documentos de la colección
- Supresión de palabras vacías (primer filtro)
- Reducción morfológica
  - Stemming
  - Lematización
- Supresión de palabras vacías (segundo filtro)
- La ley de Zipf y la frecuencia de aparición
- Técnica de cortes de Luhn: Cut-on y Cut-off
- Cálculo del punto de transición

Tabla 15. La compresión de la indexación

Reducción morfológica

La reducción morfológica es el proceso por el cual se depuran todos los términos de un texto, reduciendo su número de caracteres, simplificando su forma original, género, número, desinencia, prefijo, sufijo en una forma de palabra más común o normalizada, debido a que la mayor parte de ellas tienen la misma significación semántica. Este proceso reduce el tamaño de los términos, del diccionario, fichero inverso y mejora el "*recall*" o exhaustividad de los resultados en la recuperación de información

Stemming

Efectúa una reducción de las palabras a sus elementos mínimos con significado, las raíces de las palabra, de hecho "*Stem*" significa tallo ó raíz. De esta forma, los procesos de stemming, acotan las terminaciones de las palabras a su forma más genérica o común, obsérvese la *tabla16*.

Término	Stem		Término	Stem
che	che		consign	consign
checa	chec		consigned	consign
checar	chec		consigning	consign
checo	chec		consignment	consign
checoslovaquia	checoslovaqui		consist	consist
chedraoui	chedraoui		consisted	consist
chefs	chefs		consistency	consist
cheliabinsk	cheliabinsk		consistent	consist
chelo	chel		consistently	consist
chemical	chemical		consisting	consist

chemicalweek	chemicalweek		consists	consist
chemise	chemis		consolation	consol
chepo	chep		consolations	consol
cheque	chequ		consolatory	consolatori
chequeo	cheque		console	consol
cheques	chequ		consoled	consol
cheraw	cheraw		consoles	consol
chesca	chesc		consolidate	consolid
chester	chest		consolidated	consolid
chetumal	chetumal		consolidating	consolid
chetumaleños	chetumaleñ		consoling	consol
chevrolet	chevrolet		consolingly	consol
cheyene	cheyen		consols	consol
cheyenne	cheyenn		consonant	conson
chi	chi		consort	consort
chía	chi		consorted	consort
chiapaneca	chiapanec		consorting	consort
chiapas	chiap		conspicuous	conspicu
chiba	chib		conspicuously	conspicu
chic	chic		conspiracy	conspiraci
chica	chic		conspirator	conspir
chicago	chicag		conspirators	conspir
chicana	chican		conspire	conspir
chicano	chican		conspired	conspir
chicas	chic		conspiring	conspir
chicharrones	chicharron		constable	constabl
chichen	chich		constables	constabl
chichimecas	chichimec		constance	constanc
chicles	chicl		constancy	constanc
chico	chic		constant	constant

Fuente: *Proyecto Snowball. Disponible en: http://snowball.tartarus.org/*

Tabla 16. Ejemplo clásico de stemming

Como se desprende del análisis de los ejemplos de la tabla1, tanto en inglés como en español y en cualquier idioma, un término puede ser reducido a su común denominador, permitiendo la recuperación de todos los documentos cuyas palabras tengan la misma raíz común, por ejemplo (catálogo, catálogos, catalogación, catalogador, catalogar, catalogando, catalogado, catalogándonos).

Todos los términos derivan en tal caso de "*catalog*", haciendo posible que la recuperación sea completa en más de 8 supuestos distintos. No obstante no siempre esta técnica permite funcionar perfectamente todas las consultas que un usuario pueda plantear, es el caso de eliminar prefijos y sufijos cuya raíz puede ser compartida por múltiples palabras, véase *tabla17*.

Término con prefijo	Raíz/Stem	Término con el que causaría confusión
Prevalencia	valenc	Valencia, valencia, valenciano, ambivalencia, polivalencia,
Precatalogar	catalog	Descatalogar, catalogo,

Tabla 17. Ejemplo de conflictos de los procesos de stemming

Uno de los métodos más conocidos para llevar a efecto la reducción morfológica es el algoritmo de Martin Porter, diseñado para eliminar las palabras más comunes del inglés inicialmente y posteriormente aplicado para terceros idiomas, véase bibliografía de Porter y proyecto Snowball.

La ley de Zipf y la frecuencia de aparición

En 1949 el lingüista y científico George Kingsley Zipf, formuló la ley empírica (de tipo potencial) que lleva su nombre y con la que se determina que la frecuencia de cualquier palabra es inversamente proporcional a la posición que ocupa en la tabla de frecuencias. Esto significa que la palabra más frecuente de un texto tiene una frecuencia de aparición que dobla a la de la segunda palabra más frecuente.

La segunda palabra más frecuente tendrá una frecuencia de aparación que dobla a la de la tercera palabra más frecuente. Dicho de otra forma, la frecuencia de aparición de una palabra es inversamente proporcional a su número de orden. Y de esta forma se repetiría el esquema potencial de la ley Zipf con todos los términos del texto, véase la *tabla18* donde se demuestra este concepto.

t	(n)	tf(n)	K = tf(n) x (n) Constante ZIPF	Ley de ZIPF tf(n) = K/(n)
Término, palabra.	Rango o posición de las palabras en el ranking.	Frecuencia de aparición de un término.	La constante de Zipf es igual a la frecuencia de aparición del término por su rango.	La frecuencia de aparición de un término es igual a la constante de Zipf dividida por el rango de la palabra.
de	1	85234	85234	85234 / 1
a	2	42617	85234	85234 / 2
el	3	28411	85233	85234 / 3
un	4	21308	85232	85234 / 4
los	5	17046	85230	85234 / 5
las	6	14205	85230	85234 / 6
que	7	12176	85232	85234 / 7

Tabla 18. Ejemplo de la ley de Zipf

Abundando en lo expresado en la tabla1, se observa que la ley de Zipf determina que la frecuencia de aparición de un término es inversamente proporcional a su número de rango (orden o posición que ocupa entre las palabras más frecuentes) de tal manera que se puede preveer cual será aproximadamente su valor empleando la siguiente formulación, véase *figura4*.

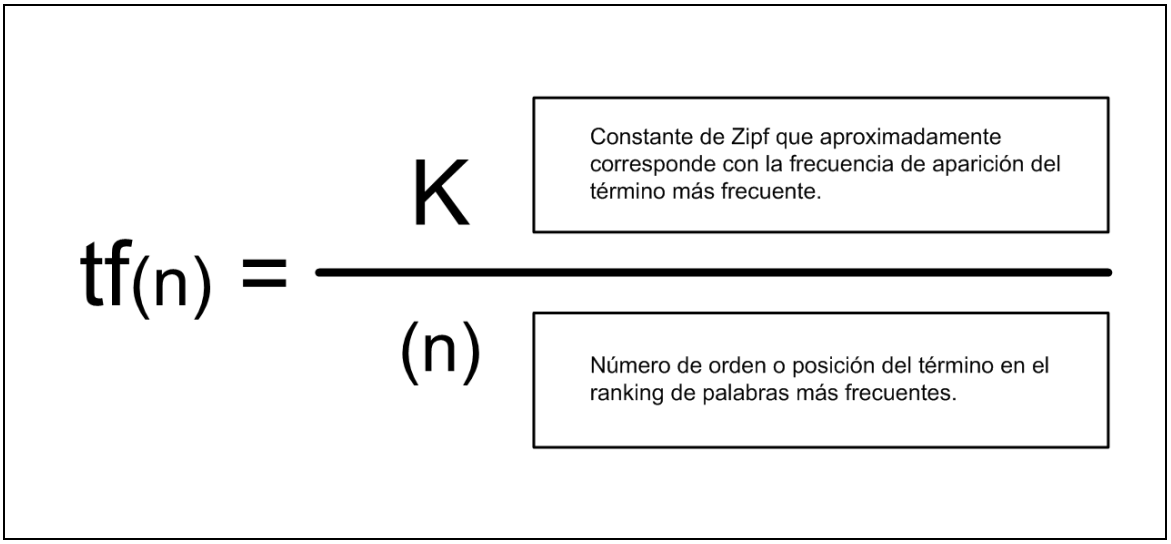


Figura 4. Fórmula correspondiente a la Ley de Zipf

Como se observará, al ser una ley empírica, cuando se calcula la constante *K* para todos los términos del ranking, no siempre el valor es coincidente con la frecuencia de aparición de *tf(1)*. No obstante, este valor es aproximado y permite determinar en qué

medida se cumple la ley de Zipf y cuál es la desviación por error. Basándose en estas formulaciones, Zipf descubre que en efecto las palabras más frecuentes, cerca del 75% del total, correspondían a palabras de mínima representatividad y que aproximadamente entre el 20% y el 30% de las palabras de un documento eran las realmente significativas y susceptibles de recuperación, véase *figura5*.

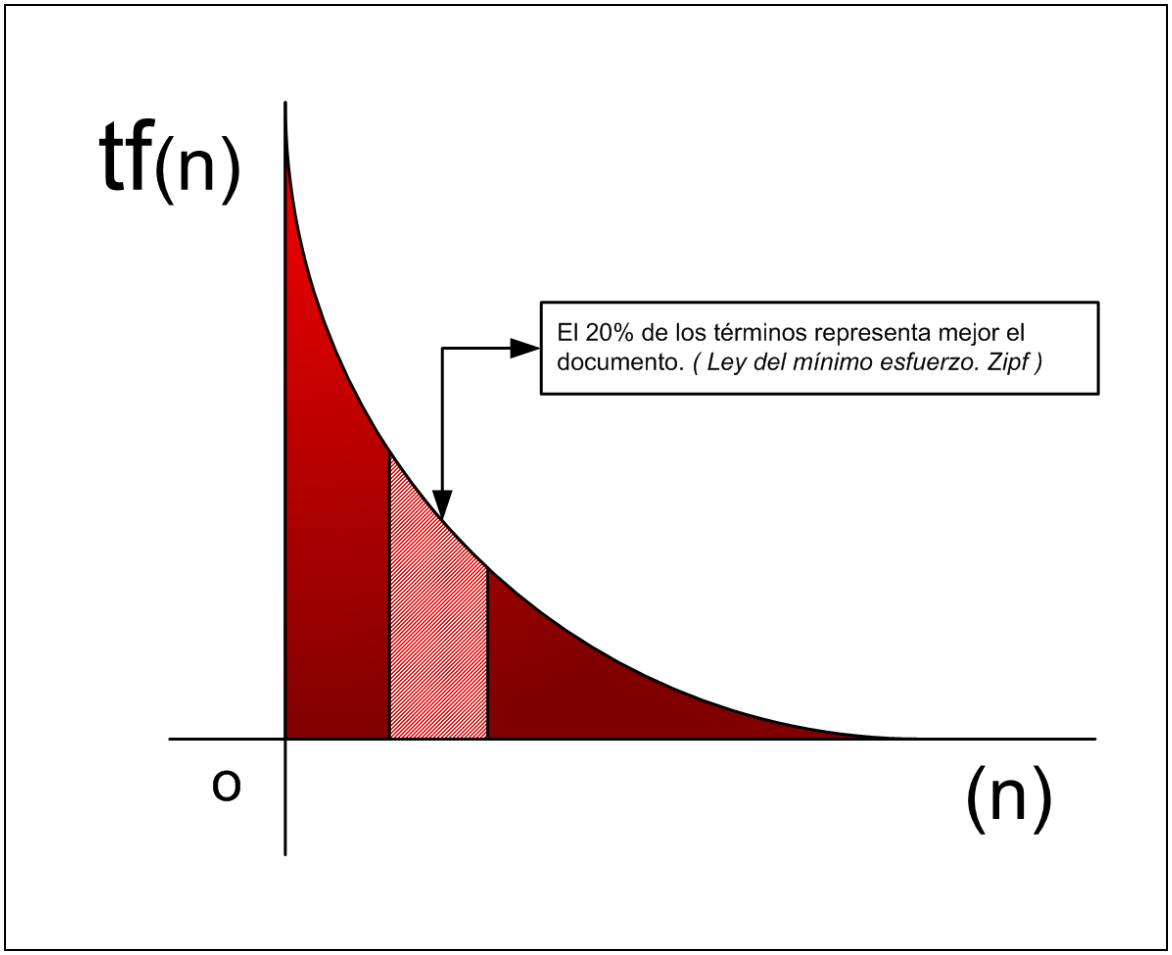


Figura 5. Función logarítmica de la frecuencia de los términos de un documento

Llega a esta conclusión basándose en la ley del mínimo esfuerzo, ya que se supone que el usuario no utilizará términos de búsqueda cuya frecuencia de aparición sea tan baja o tan elevada como para encontrarse en los bordes potenciales del cuadro logarítmico, prefiriendo utilizar por consiguiente un término más común ó habitual con una frecuencia de aparición media.

Técnica de cortes de Luhn: Cut-on y Cut-off

Según lo especificado en la figura2, la expresión logarítmica de los términos de un documento, muestra una curva pronunciada con los términos de altísima frecuencia de aparición y su inverso, aquellos términos de muy baja frecuencia de aparición. Este hecho, ya adelantado por Zipf, dió lugar al empleo de la técnica de cortes que propuso Luhn en 1958, conjetura por la que ya se venía intuyendo que los términos situados en los extremos del eje de abscisas y de ordenadas serán los que menos poder de resolución o representatividad tienen para un determinado documento dentro de la colección.

Consiste en la eliminación de los términos de altísima frecuencia de aparición (Cut-on) y términos de bajísima frecuencia de aparición (Cut-off), entre los que se incluyen los "Hápax", términos cuya frecuencia de aparición es la unidad. Véanse las *figuras 6, 7 y 8* basadas en (SCHULTZ, C.K. 1968).

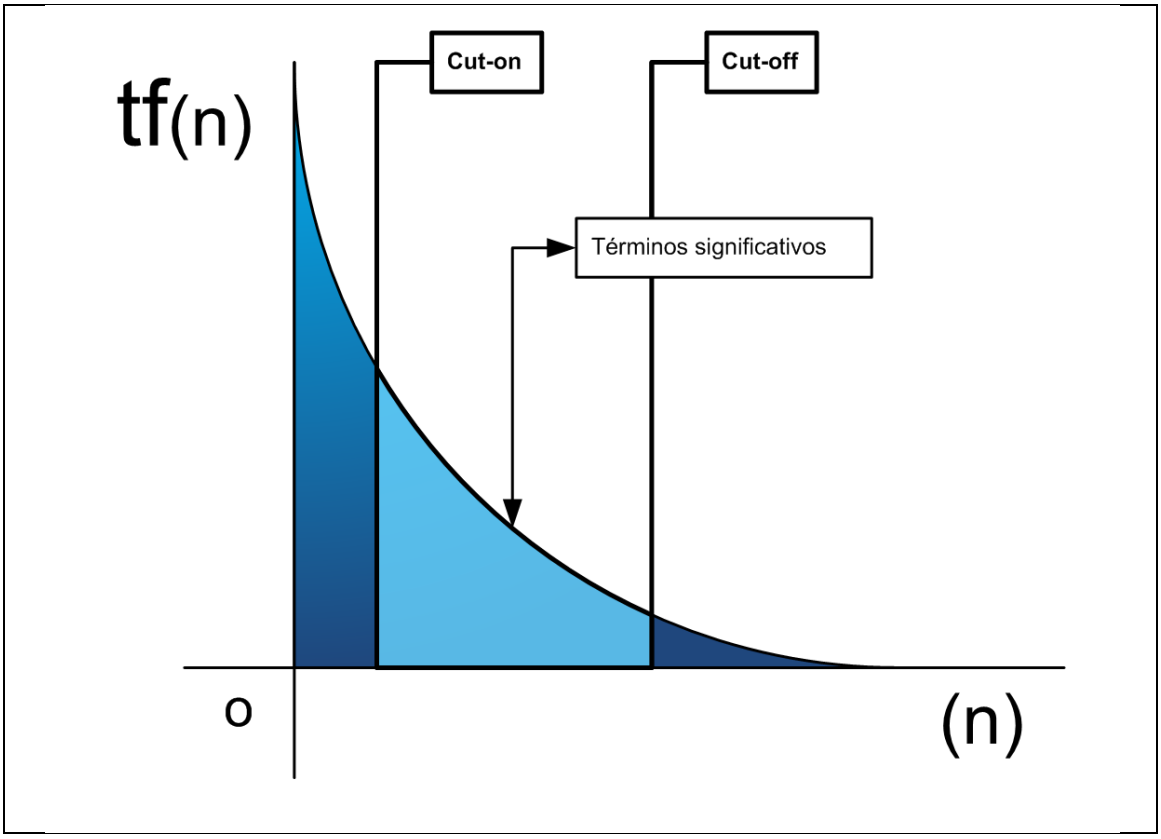


Figura 6. Los cortes de Luhn Cut-on, Cut-off y los términos significativos con frecuencias medias

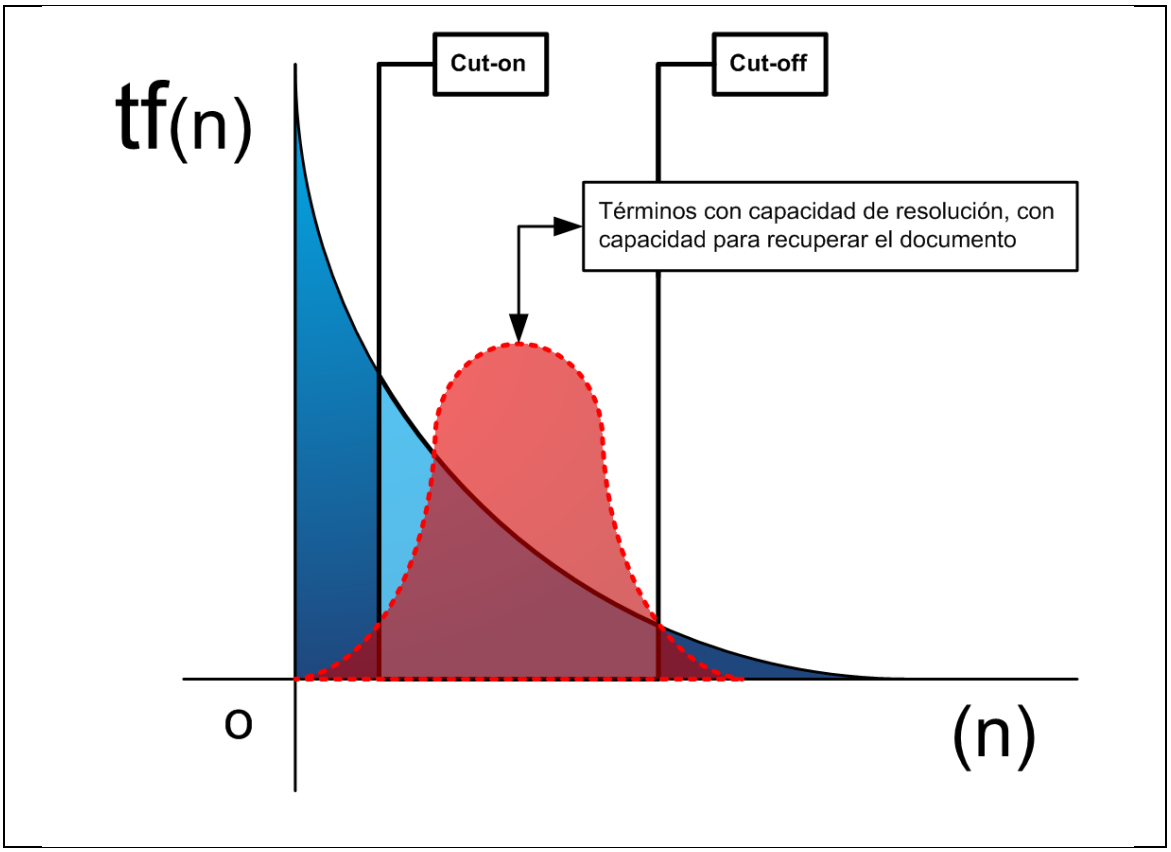


Figura 7. En color rojo, la curva hiperbólica que representa el área de términos con capacidad de resolución y representación del documento

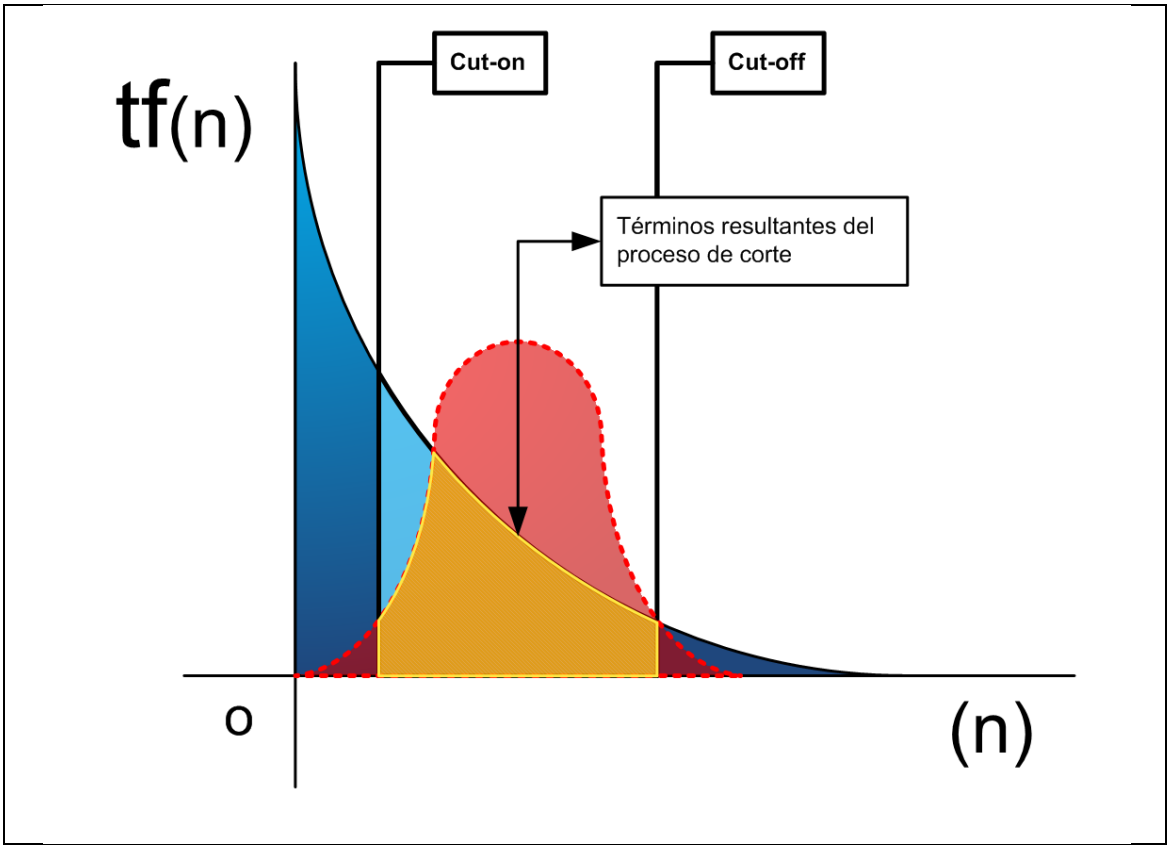


Figura 8. El área amarilla representa los términos significativos y resolutivos resultantes del proceso de corte y de la curva hipérbólica

La cuestión a plantear en este punto sería ¿cómo se aplicaban los cortes? ¿cómo se determinaba que los términos de una frecuencia de aparición determinada eran merecedores o no de supresión y eliminación, previos a la indexación? Inicialmente la técnica de cortes propuesta por Luhn se realizaba de forma arbitraria, eliminando un porcentaje de términos alineados en los extremos del ranking.

Cálculo del punto de transición

Para averiguar de forma científica la estimación del corte superior "*Cut-on*" e inferior "*Cut-off*", una de las soluciones más estudiadas es el cálculo del punto de transición o inflexión entre los términos generales y especializados, cuyas características pueden resumirse en la *tabla19*.

Términos con alta frecuencia de aparición	**Términos con baja frecuencia de aparición**
Generales	Específicos
Tienden a unificar el lenguaje	Tienden a diversificar el lenguaje
Proporcionan nexos de unión en torno al texto	Detallan el contenido del texto
Pulsión	Repulsión
Artículos, preposiciones, conjunciones, contracciones, pronombres, algunos adverbios y verbos	Terminología especializada de un determinado área de conocimiento, vocabulario técnico, científico, Hápax
Rangos cercanos al origen del eje	Rangos situados en el extremo opuesto al origen
Técnica de recorte aplicada: Cut-on	Técnica de recorte aplicada: Cut-off

Tabla 19. Características de los términos según su frecuencia

Teniendo claras estas circunstancias y sabiendo que en los términos con frecuencias medias están los términos representativos; el investigador Andrew Donald Booth (BOOTH, A.D. 1967), perfecciona la técnica del punto de transición de William Goffman (URBIZAGÁSTEGUI ALVARADO, R.; RESTREPO ARANGO, C. 2011), en su trabajo titulado A Law of Occurrences for Words of Low Frequency, donde diseña una formulación refinada, véase *figura9*.

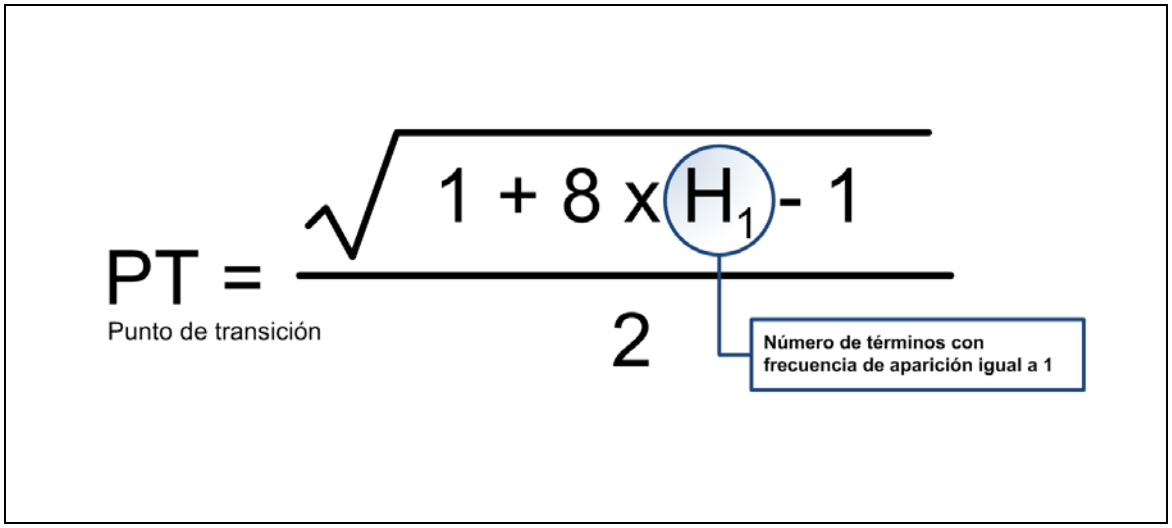


Figura 9. Fórmula de Booth para el cálculo del punto de transición

El punto de transición no es más que una frecuencia de aparición intermedia a partir de la cual, se determina el porcentaje de términos tendientes a la generalidad o a la especificidad (según se separen de la frecuencia de aparición del punto de transición) para determinar dónde efectuar los cortes. Véase *figura10*.

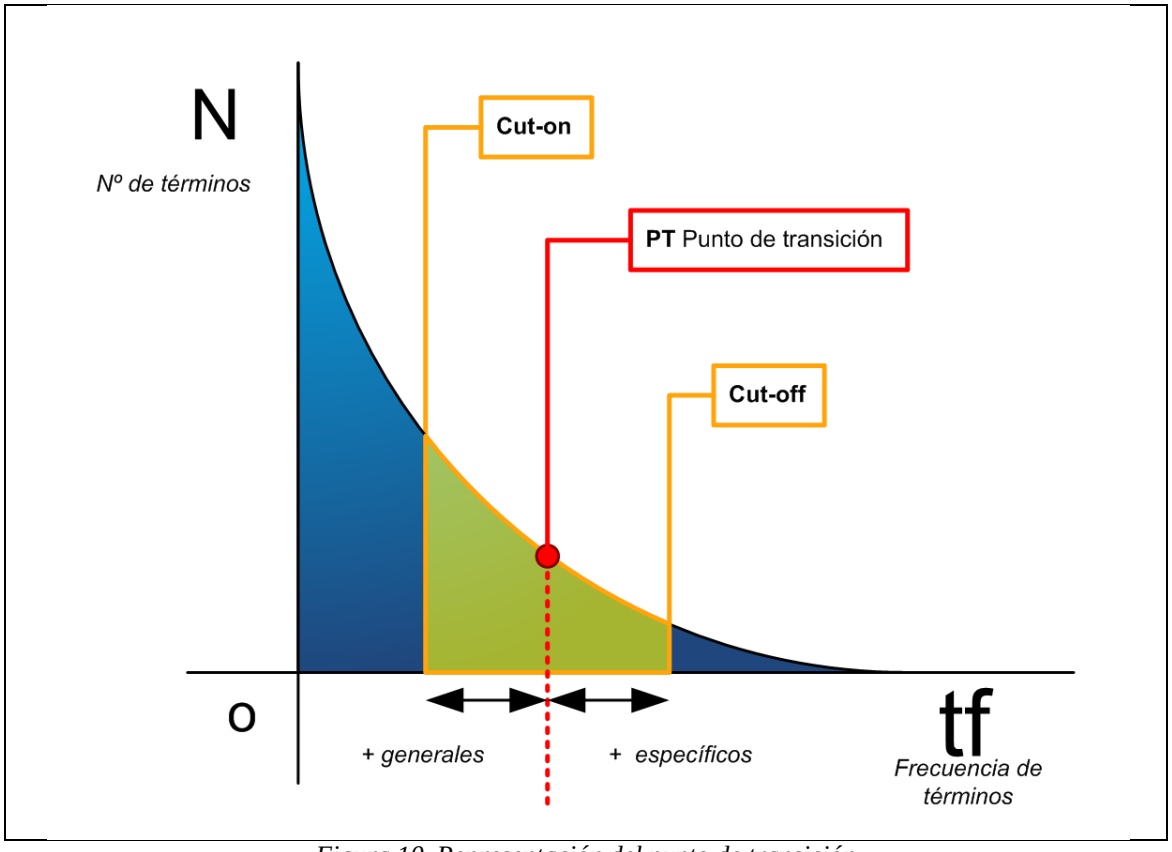


Figura 10. Representación del punto de transición

Pero aún estableciendo un punto de transición, aún faltan por determinar los puntos de corte Cut-on y Cut-off, esto se lleva a cabo mediante 2 rangos definidos o bien mediante porcentajes o por valores aproximados a la frecuencia del punto de corte. Tales rangos no tienen porque resultar simétricos, dependiendo del tipo de recuperación que se pretenda conseguir a la postre. Una recuperación más precisa, deberá incluir más términos con frecuencias de aparición bajas. Si se desea una recuperación más exhaustiva, se deberá primar un margen superior para el Cut-on. Obsérvese en la *figura10*, que tales rangos se definen a partir del valor de la frecuencia de aparición del punto de corte, tanto a derecha como a izquierda, haciendo que el segmento resultante sea el conjunto definitivo a tener en cuenta en la indexación.

7. Modelo Booleano

Es el modelo con más solera en recuperación de información. Se basa en la teoría de conjuntos del álgebra de George Boole, mediante la aplicación de operaciones lógicas AND (Intersección), OR (Unión), NOT (Resta) y XOR (Complemento). Para llevarlo a efecto sólo se necesita un fichero diccionario donde almacenar los términos de los documentos de la colección y la correspondencia de aparición de los mismos, véase *tabla20*.

Id	Término	Id del documento
T1	Archivo	{1, 3, 4, 5,…}
T2	Biblioteca	{2, 3, 4,…}
T3	Museo	{1, 3,…}
T4	Arquitectura	{1,…}
T5	Facultad	{1, 2, 3, 4,…}
T6	Documentación	{1, 2, 3, 4, 5,…}
T7	Investigación	{3, 4, 0…}

Tabla 20. Ejemplo de fichero diccionario con los términos y los identificadores de documento

De esta forma se obtiene que el término *T4* aparezca sólo en el documento 1 y que el término *T3* aparezca en los documentos 1 y 3. Si se analiza esta representación, se observará que el modelo booleano es en efecto un modelo binario por determinar fundamentalmente la presencia o la ausencia de los términos de los textos de la colección. De esta forma también se podría representar en otro ejemplo una matriz que pusiera en relación los términos del diccionario con los documentos, véase *tabla21*.

Diccionario	Doc1	Doc2	Doc3	Doc4	Doc5
Archivo	1	0	1	1	1
Biblioteca	0	1	1	1	0
Museo	1	0	1	0	0
Arquitectura	0	1	0	0	0
Facultad	1	1	1	1	0
Documentación	1	1	1	1	1
Investigación	0	0	1	1	0

Tabla 21. Ejemplo de matriz término-documento, donde se aprecia el modelo binario

Obsérvese que la presencia de un término en un documento se representa con el valor "1" y la ausencia con el valor "0". De esta forma una hipotética consulta booleana para recuperar aquellos documentos en los que aparezcan los términos "archivo" y "biblioteca", se expresará de la siguiente forma:

Q = 10111 AND 01110 = 00110

Tabla 22. Resolución de consulta booleana AND con vectores binarios

Por otro lado, también se debe cuidar la forma de generar el fichero diccionario, si bien en los ejemplos anteriores, no se han mostrado términos coincidentes, la frecuencia de aparición de los términos suele ser superior a la unidad, implicando un proceso de varios pasos para la optimización de los datos recopilados por el sistema, véase la *tabla23*.

Tabla diccionario		**Ordenación alfabética**		**Armonización**		
Término	**ID doc**	**Término**	**ID doc**	**Término**	**TF**	**ID doc**
Lengua	1	*Bilingüismo*	2	*Bilingüismo*	1	2
Cerebro	1	*Capacidad*	1	*Capacidad*	1	1
Idioma	1	*Cerebro*	1	*Cerebro*	2	1, 2
Redes	1	*Cerebro*	2	*Comunicativa*	1	1
Neuronales	1	*Comunicativa*	1	*Español*	1	2
Políglotas	1	*Español*	2	*Funcional*	1	2
Monolingüe	1	*Funcional*	2	*Idioma*	2	1, 2
Persona	1	*Idioma*	1	*Inglés*	1	2
Comunicativa	1	*Idioma*	2	*Lengua*	2	1, 2
Capacidad	1	*Inglés*	2	*Monolingüe*	1	1
Neuropsicología	2	*Lengua*	1	*Neuronales*	1	1
Funcional	2	*Lengua*	2	*Neuropsicología*	1	2
Lengua	2	*Monolingüe*	1	*Persona*	1	1
Bilingüismo	2	*Neuronales*	1	*Políglotas*	1	1
Cerebro	2	*Neuropsicología*	2	*Psicología*	1	2
Idioma	2	*Persona*	1	*Redes*	1	1
Español	2	*Políglotas*	1	*Romance*	1	2
Inglés	2	*Psicología*	2	*Solapamiento*	1	2

Romance	2		*Redes*	1				
Psicología	2		*Romance*	2				
Solapamiento	2		*Solapamiento*	2				

Tabla 23. Ejemplo de transformación del fichero diccionario

Como se puede ver, la tabla diccionario del sistema se compone originalmente de todos los términos de todos los textos del corpus documental, recogiendo el término propiamente dicho y el identificador del documento en el que aparecía. Obsérvese que en este estadio se encuentran términos repetidos en distintos documentos.

El siguiente paso es la ordenación alfabética de los términos de la tabla diccionario, donde se visualiza mejor si cabe la duplicidad de los términos "*cerebro*", "*idioma*" y "*lengua*". Finalmente se lleva a cabo un proceso de armonización en el que se suprimen los términos duplicados hasta que sólo quede uno de ellos. El valor del campo identificador de documentos "*ID doc*" amplía su información con la lista de documentos en los que aparece el término correspondiente. Por otro lado, se añade el campo "*TF*" que indicará la frecuencia de aparición del término para futuros cálculos. Este proceso permite ahorrar memoria y facilita el proceso de recuperación.

Casos fundamentales

El algebra de Boole aplicada a la recuperación de información consta de una serie de casos básicos a los que pueden añadirse múltiples cadenas de resolución booleana. Dicho de otra forma, pueden llevarse a cabo operaciones verdaderamente complejas dependiendo de la cantidad de conjuntos a dirimir.

Operador AND. Q = "A" AND "B"

El operador AND es el encargado de intersecar o especificar que dos condiciones, premisas ó terminos tienen que cumplirse obligatoriamente, simultáneamente ó a la vez. Esto significa que si no se produce de esta forma, el sistema de recuperación no devolverá resultado alguno. Según lo que se muestra en la *figura11*, sólo los documentos que posean el término A y B (zona sombreada) se recuperarán, desechando por lo tanto aquellos términos que o bien sólo contengan A o bien sólo contengan B.

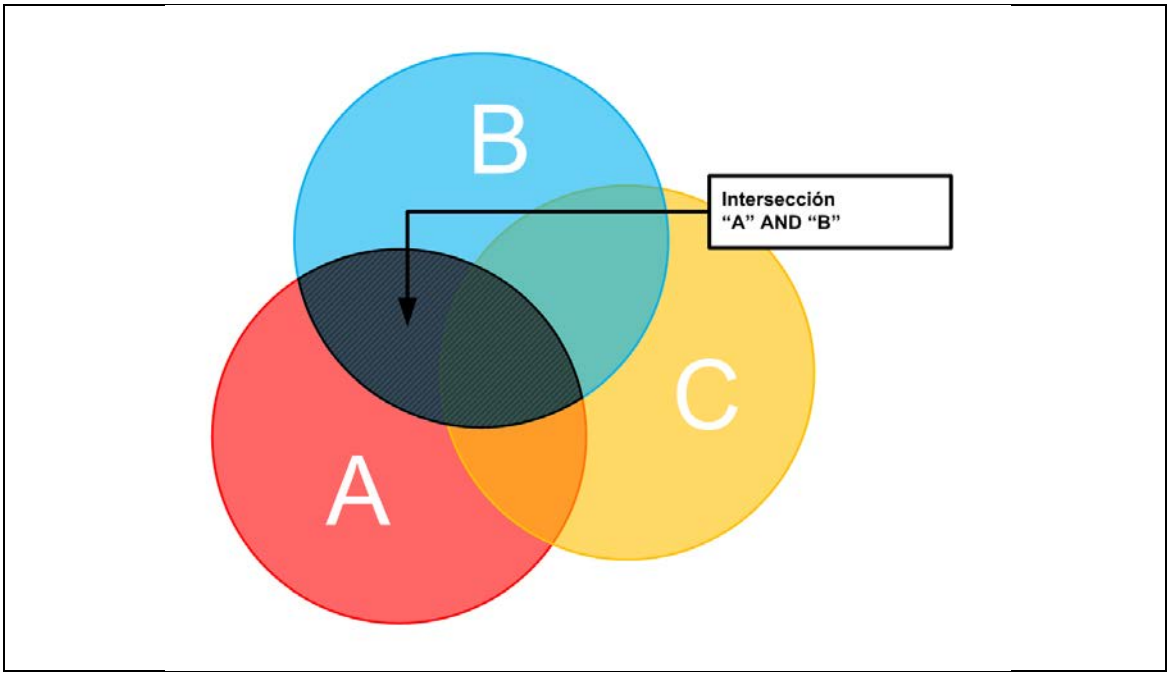


Figura 11. Intersección de documentos con el término A y B

Operador AND. Q = "A" AND "B" AND "C"

En este caso, la consulta propuesta implica la intersección de 3 términos. Por lo tanto, la recuperación sólo se efectuará para aquellos documentos que tengan presentes los términos A, B y C. Si faltase uno de estos términos el documento no se recuperaría, véase *figura12*.

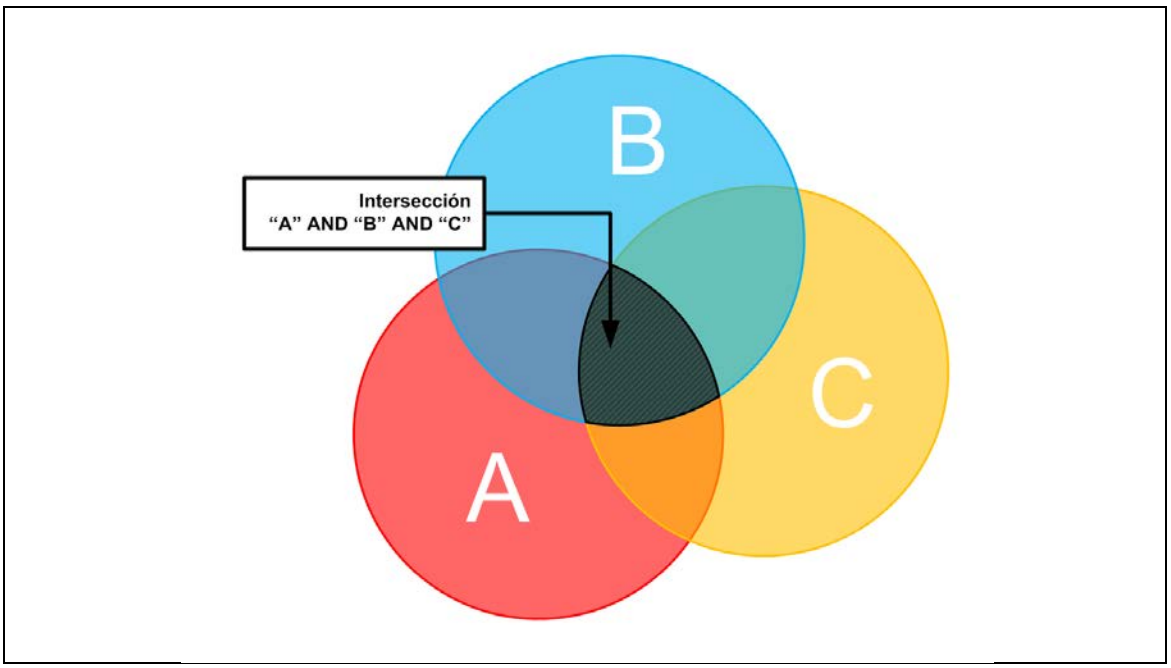


Figura 12. Intersección de documentos con los términos A, B y C

Operador OR. Q = "A" OR "B"

El operador OR implica unión, alternativa y adición. Esto significa que dos conjuntos conectados por el operador OR se sumarán o unirán y si constan de elementos comunes, éstos también se recogerán. En recuperación de información significa que para una consulta de términos A OR B, se recuperarán aquellos documentos que tengan presencia de A, presencia de B y presencia de A y B a la vez. Por ello la consulta AND es más específica y restrictiva que OR, mucho más amplia, véase *figura13*.

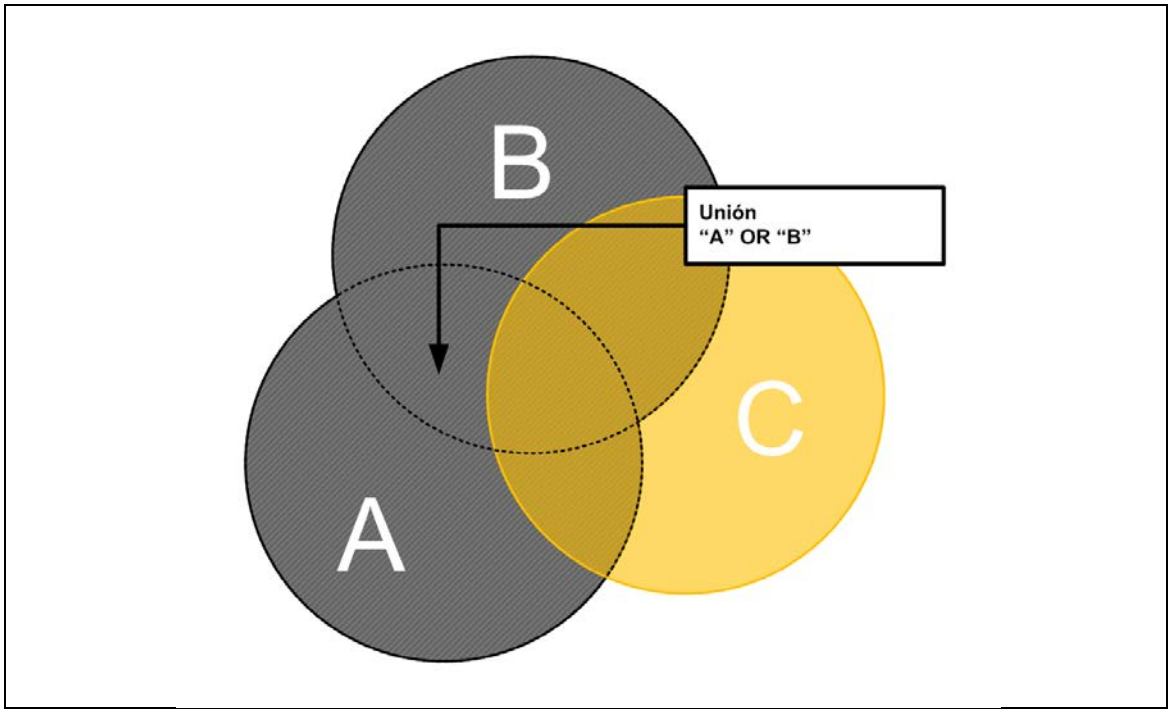


Figura 13. Unión de los documentos con los términos A y B

Operador NOT :: Q = "A" NOT "B"

El operador NOT también conocido como de negación, implica resta, diferencia, reducción o sustracción. Esto es restar a un conjunto de documentos aquellos que contenga el término B. Obsérvese la *figura14*, en la que sólo se recuperan aquellos documentos que contengan los términos A pero no los términos B.

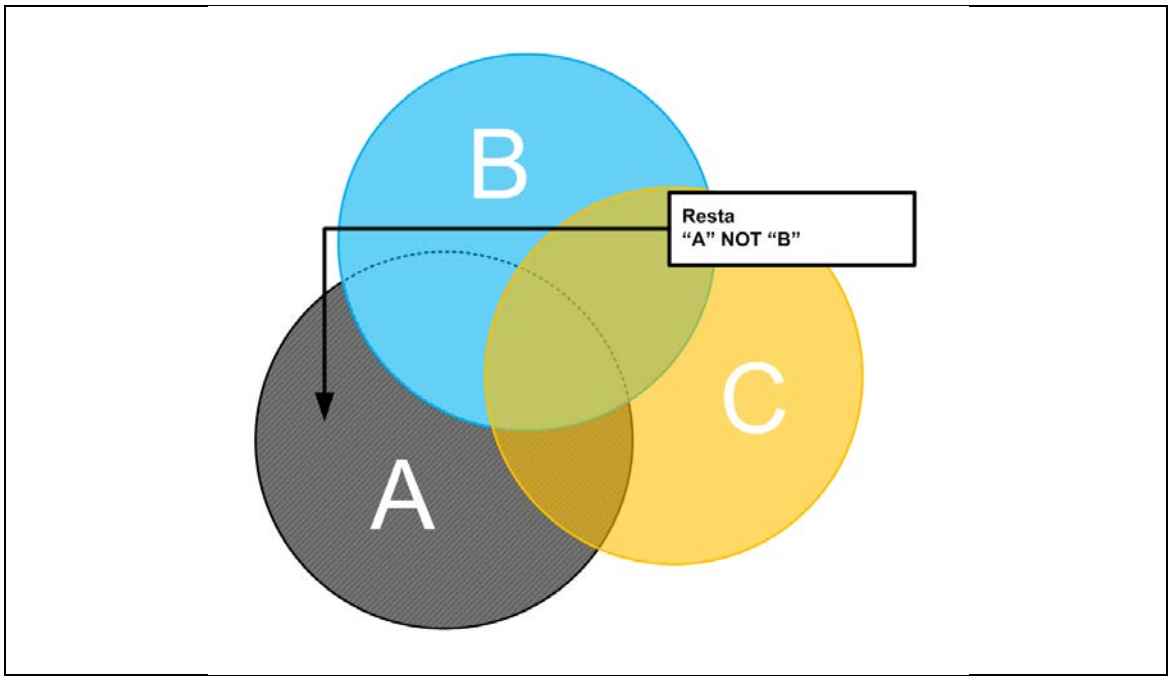


Figura 14. Documentos que contengan el término A pero no B

Operador AND y NOT. Q = "A" AND "B" NOT "C"

La flexibilidad del lenguaje booleano permite combinar distintos operadores para obtener resultados más restringidos. Según se muestra en este caso, véase *figura15,* el operador AND y NOT pueden precisar la distinción de términos basándose en la negación de un tercero. De esta forma la consulta Q recuperará aquellos documentos en los que esté presente el término A y B pero no C.

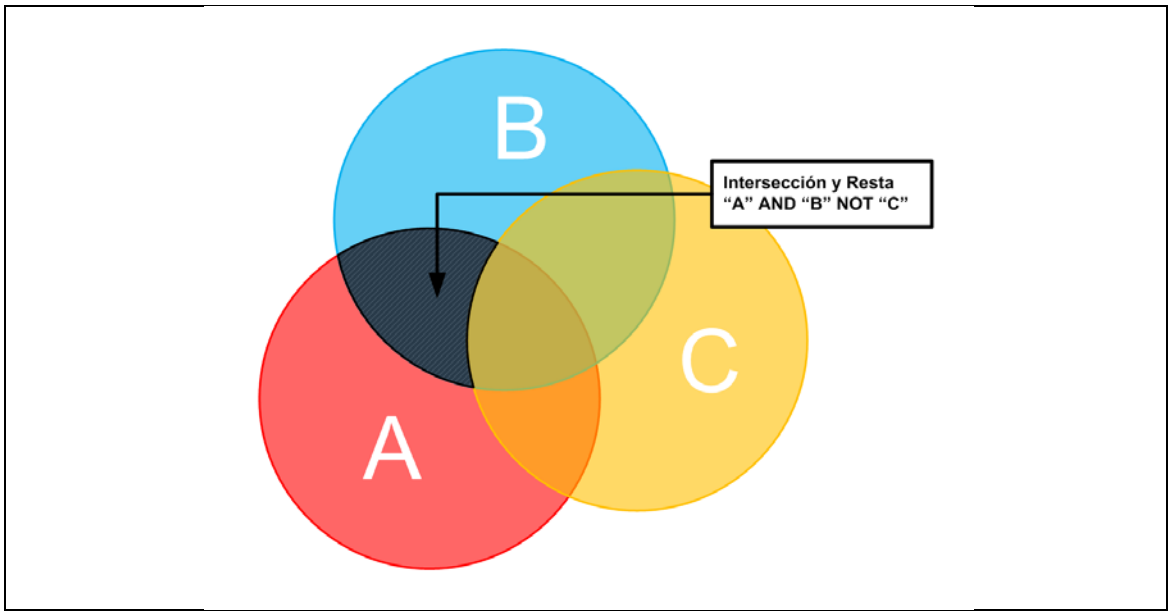


Figura 15. Documentos que contengan los términos A y B pero no C

Operador XOR :: Q = "A" XOR "B" XOR "C"

El operador XOR se utiliza para seleccionar todos los elementos complementarios de los conjuntos. Dicho de otra forma, evita las intersecciones. En la *figura16*, se observa que la zona de documentos que serán recuperados es aquella en la que no se combinan los términos A, B y C. Así por ejemplo, de esta forma la expresión A XOR B es equivalente a (A AND (NOT B)) OR ((NOT A) and B).

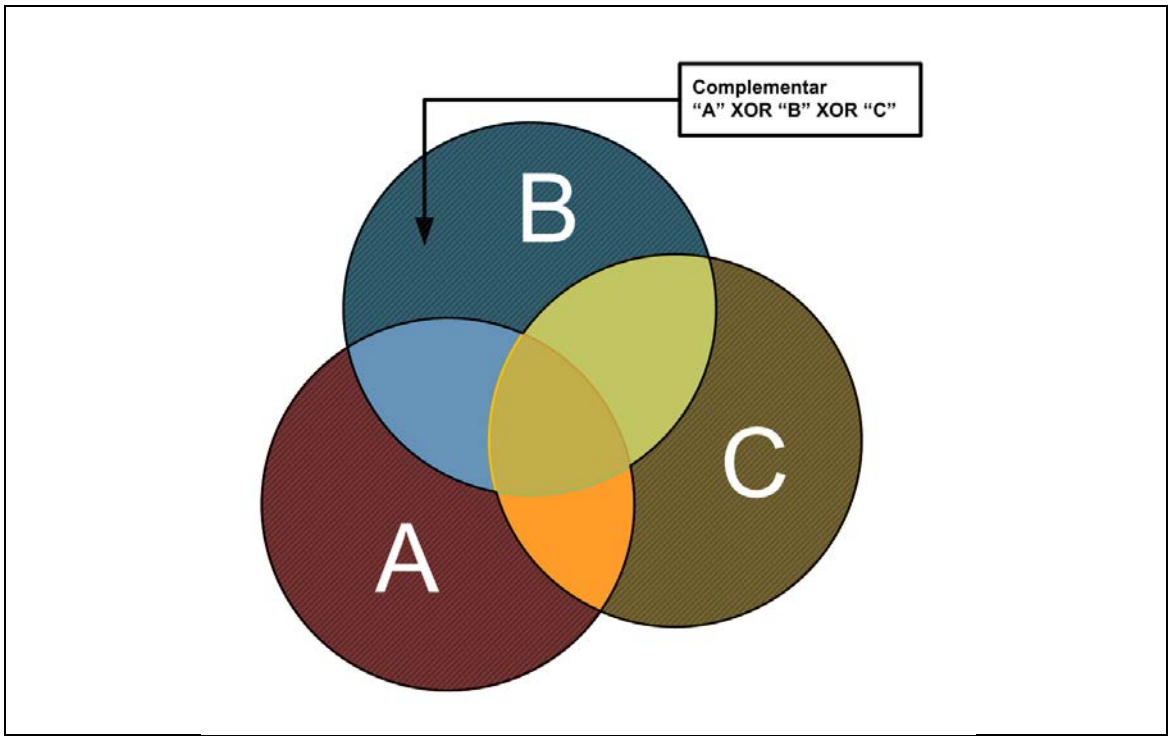


Figura 16. Documentos cuyos términos complementarios sean A, B y C

Ventajas del modelo booleano

- El modelo booleano permite procesar colecciones muy grandes rápidamente. Resulta sistemático y ello supone una gran velocidad de recuperación.
- Es un modelo flexible ya que permite el empleo de distintas conectivas para precisar la consulta del usuario. Permite aproximar bastante las consultas por frase exacta y resulta perfectamente válido para recuperar por medio de vocabulario controlado.
- Entraña ventajas para efectuar una recuperación de información igualada, en el sentido de que el sistema de información presente la mejor respuesta a una necesidad de información expresada por ciertas palabras clave.

Inconvenientes del modelo booelano

- En muchos casos, las necesidades de información son complejas y ello entraña cierta dificultad a la hora de expresar las consultas mediante fórmulas lógicas que pueden incluso llegar a concatenarse.
- A veces el usuario puede imponer una lógica semántica que no se corresponda con la lógica algebraica de Boole, implicando un uso incorrecto de los operadores.
- El volumen de resultados no se puede controlar, ya que la consulta plantea una resolución absoluta para toda la colección en la que se aplica. Esto significa que el resultado puede ser excesivamente grande o pequeño.
- Los resultados obtenidos pueden ser perfectamente relevante o absolutamente irrelevante, no hay gradación o término medio, ya que el funcionamiento del modelo booleano se basa en equiparación exacta. Es decir que no ordena los documentos por orden de relevancia, tal como se llevaría a cabo en un modelo basado en pesos o ponderación de los términos.

Resolución de consultas booleanas ordenadas

El modelo booleano puro no contempla el uso TF (Term Frequency), pero los sistemas de recuperación más modernos suelen utilizar métodos para mejorar la ordenación o el ranking de los resultados mostrados. En este sentido en la *tabla22*, ya se avanzaba esta consideración. El modo de proceder en tal caso es el que se muestra a continuación:

- Input de la consulta booleana
- Ordenar los términos de la tabla diccionario en orden descendiente de frecuencia
- Ejecutar la consulta, resolviéndose por orden de frecuencia de los términos.
- Suma de frecuencias de los términos de consulta para los documentos recuperados.
- Devolución de resultados ordenados.

8. Modelo Vectorial

El modelo de espacio vectorial se basa en el grado de similaridad de una consulta dada por el usuario con respecto a los documentos de la colección cuyos términos fueron ponderados mediante TF-IDF. El modelo vectorial fue presentado por Salton en 1975 y posteriormente asentado en 1983 junto con Mc Gill y se basa en tres principios esenciales (MARTÍNEZ COMECHE, J.A. 2006):

- La equiparación parcial, esto es, la capacidad del sistema para ordenar los resultados de una búsqueda, basado en el grado de similaridad entre cada documento de la colección y la consulta.

- La ponderación de los términos en los documentos, no limitándose a señalar la presencia o ausencia de los mismos, sino adscribiendo a cada término en cada documento un número real que refleje su importancia en el documento.

- La ponderación de los términos en la consulta, de manera que el usuario puede asignar pesos a los términos de la consulta que reflejen la importancia de los mismos en relación a su necesidad informativa.

Si bien en el modelo booleano un documento de la colección puede ser representado por la presencia o ausencia de los términos indexados en el fichero diccionario de la siguiente forma, véase *tabla24*…

Documento1 { 1,0,1,1,1,0,0,1,0,0,0,1,1,0,1,1 }

Tabla 24. Ejemplo de vector binario

…en el modelo de espacio vectorial se emplea el peso de los términos para cada documento, que refleja la relevancia de los términos del documento de cara a su representatividad en la colección, adquiriendo una forma como la que sigue, véase *tabla25*…

Documento1 { 1`452, 0, 2`122, 3`564, 4`123, 0, 0, 2`342, 0, 0, 0, 1`975, 4`543, 0, 6`134, 2`234 }

Tabla 25. Ejemplo de vector de pesos TF-IDF

A este conjunto de números reales, que son los pesos, que representan al documento, se les denomina Vector del documento, permitiendo su representación en el espacio vectorial y en consecuencia, su tratamiento matemático. Por ello la formulación del vector se representa de la siguiente forma, véase *tabla26*.

Id	Término	Documento 1	
		Peso binario	Peso TF-IDF
T1	Clima	1	1,452
T2	Biblioteca	0	0
T3	Universidad	1	2,122
T4	Alcalá	1	3,564
T5	España	1	4,123
T6	Libros	0	0
T7	Geografía	0	0
T8	Población	1	2,342
T9	Electricidad	0	0
T10	Ciencia	0	0
T11	Social	0	0
T12	Luz	1	1,975
T13	Unamuno	1	4,543
T14	Física	0	0
T15	Fluidos	1	6,134
T16	Literatura	1	2,234
Vector del documento1			
Documento1 { $Clima_{(1,452)}$, $Biblioteca_{(0)}$, $Universidad_{(2,122)}$, $Alcalá_{(3,564)}$, $España_{(4,123)}$, $Libros_{(0)}$, $Geografía_{(0)}$, $Población_{(2,342)}$, $Electricidad_{(0)}$, $Ciencia_{(0)}$, $Social_{(0)}$, $Luz_{(1,975)}$, $Unamuno_{(4,543)}$, $Física_{(0)}$, $Fluidos_{(6,134)}$, $Literatura_{(2,234)}$ }			
Fórmula para la representación del vector de un documento			

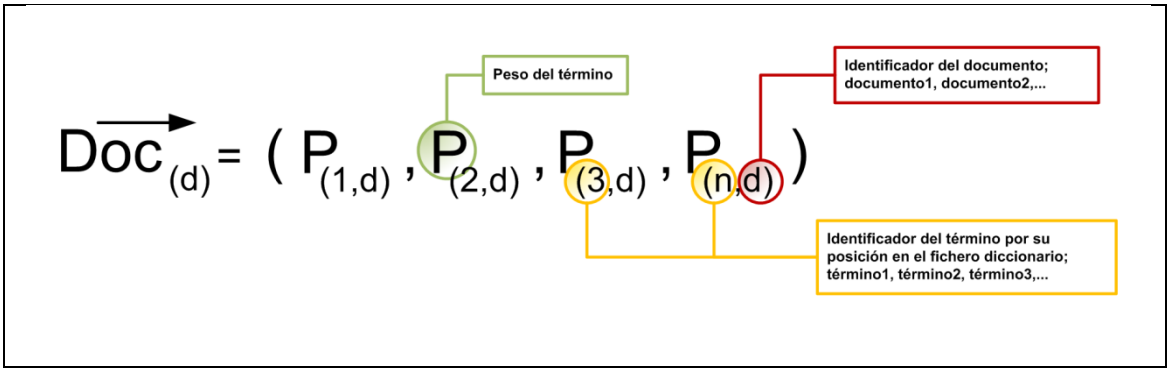


Tabla 26. Representación del vector de un documento

Posteriormente, la colección sigue lo que se denomina un Proceso de Vectorización por el que todos los documentos son representados mediante pesos TF-IDF, la consulta del usuario también requiere de dicho tratamiento. Ello significa que se tiene que ponderar la importancia de los términos de la consulta para poder generar el Vector de la consulta del usuario. Este paso es imprescindible para poder efectuar el Proceso de Equiparación de la consulta con los documentos de la colección y determinar cuáles de ellos son más relevantes, véase *tabla27*.

Cadena de consulta original del usuario				
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá				
Depuración de la consulta del usuario				
Libros Literatura Unamuno Biblioteca Universidad Alcalá				
Fichero diccionario		**Documento1**		**q = pesos de la consulta del usuario**
Id	**Término**	**Peso binario**	**Peso TF-IDF**	
T1	Clima	1	1,452	0
T2	Biblioteca	0	0	1,345
T3	Universidad	1	2,122	1,453
T4	Alcalá	1	3,564	1,987
T5	España	1	4,123	0
T6	Libros	0	0	2,133
T7	Geografía	0	0	0
T8	Población	1	2,342	0
T9	Electricidad	0	0	0
T10	Ciencia	0	0	0
T11	Social	0	0	0
T12	Luz	1	1,975	0
T13	Unamuno	1	4,543	3,452

T14	Física	0	0	0
T15	Fluidos	1	6,134	0
T16	Literatura	1	2,234	4,234

Tabla 27. Obsérvese el documento1 y una consulta q dada por el usuario con sus pesos

Proceso de equiparación mediante el producto escalar

Los procesos de equiparación de los documentos de la colección con respecto a la consulta del usuario, en el modelo booleano, se efectúan mediante cálculos de similaridad. Existen muchas modalidades de comparación o equiparación mediante similaridad, en este caso se presenta una de las más sencillas por su simplicidad y sistematización inmediata. Se trata del producto escalar de los pesos, véase *figura17*.

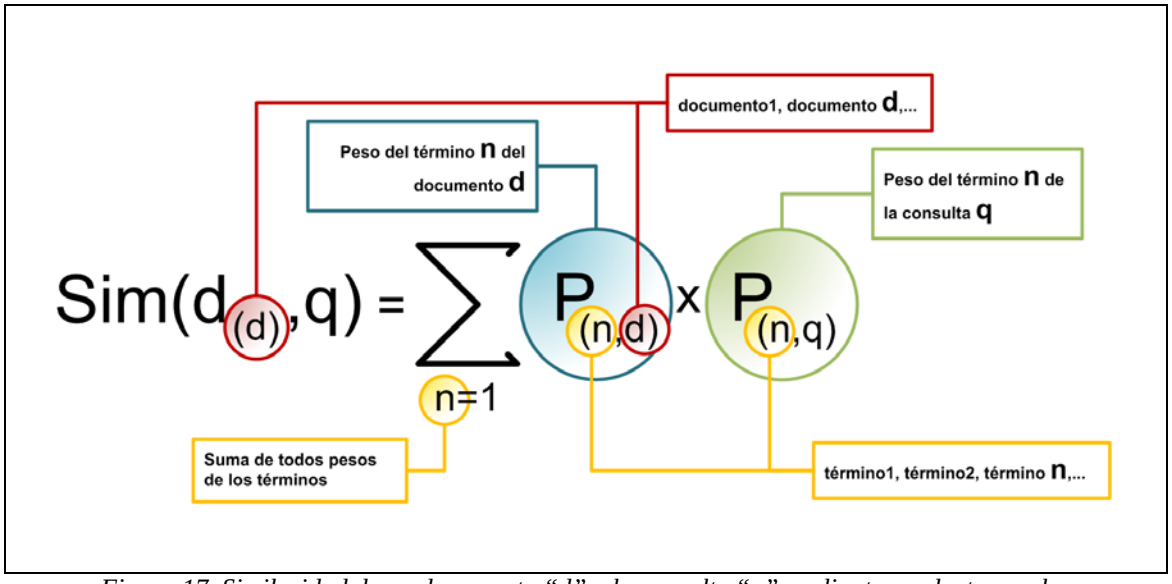


Figura 17. Similaridad de un documento "d" y la consulta "q" mediante producto escalar

De esta forma, la similaridad de un documento y una consulta, es igual a la suma de los productos de sus pesos. (Y no se debe olvidar que cada peso representa a un término). Este método puede aplicarse tanto a pesos binarios como a pesos TF-IDF.

Modalidad de pesos binarios

En el caso de la modalidad binaria, la similaridad de un documento con respecto a la consulta es equivalente a la presencia de los términos de la consulta en el documento, véase *tabla28*. Esto quiere decir que la ausencia de un término de la consulta o del documento implica un producto igual a 0 y por lo tanto no tienen incidencia en el cálculo. Por el contrario la presencia de un término dado tanto en la consulta como en el

documento siempre tendrá valor 1. Por ello sólo basta con contabilizar el número de términos coincidentes de la consulta en el documento y ése será su valor de similaridad.

Cadena de consulta original del usuario			
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá			
Depuración de la consulta del usuario			
Libros Literatura Unamuno Biblioteca Universidad Alcalá			
Fichero diccionario		**Documento1**	**q = pesos binarios de la consulta del usuario**
Id	**Término**	**Peso binario**	
T1	Clima	1	0
T2	Biblioteca	0	1
T3	Universidad	1	1
T4	Alcalá	1	1
T5	España	1	0
T6	Libros	0	1
T7	Geografía	0	0
T8	Población	1	0
T9	Electricidad	0	0
T10	Ciencia	0	0
T11	Social	0	0
T12	Luz	1	0
T13	Unamuno	1	1
T14	Física	0	0
T15	Fluidos	1	0
T16	Literatura	1	1
Proceso de equiparación mediante el producto escalar de pesos binarios			
Sim(doc1,q) = $^{Clima}(1*0) + ^{Biblioteca}(0*1) + ^{Universidad}(1*1) + ^{Alcalá}(1*1) + ^{España}(1*0) + ^{Libros}(0*1) + ^{Geografía}(0*0) + ^{Población}(1*0) + ^{Electricidad}(0*0) + ^{Ciencia}(0*0) + ^{Social}(0*0) + ^{Luz}(1*0) + ^{Unamuno}(1*1) + ^{Física}(0*0) + ^{Fluidos}(1*0) + ^{Literatura}(1*1) =$ **4**			

Tabla 28. Producto escalar de pesos binarios

Como se puede analizar en la *tabla28*, el número de términos coincidentes de la consulta con el documento1 es 4 que corresponde a los términos Universidad, Alcalá, Unamuno y Literatura. Por lo tanto, en una escala de 6 (Por ser todos los términos empleados en la consulta original depurada del usuario), el documento1, tiene un alto grado de coincidencia y por ende tiene más probabilidades de ser relevante.

Modalidad de pesos TF-IDF

En el caso de la modalidad de pesos binarios, las limitaciones en la definición de la representatividad de los términos de cada documento quedan patentes. Resulta por tanto un resultado bastante limitado y parcial. Por ello el método de la similaridad mediante el producto escalar se aplica habitualmente con pesos TF-IDF, mucho más precisos, véase *tabla29.*

Cadena de consulta original del usuario				
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá				
Depuración de la consulta del usuario				
Libros Literatura Unamuno Biblioteca Universidad Alcalá				
Fichero diccionario		**Documento1**	**Documento2**	**q = pesos de la consulta del usuario**
Id	**Término**	**Peso TF-IDF**	**Peso TF-IDF**	
T1	Clima	1,452	0	0
T2	Biblioteca	0	2,093	1,345
T3	Universidad	2,122	0	1,453
T4	Alcalá	3,564	0	1,987
T5	España	4,123	4,245	0
T6	Libros	0	1,234	2,133
T7	Geografía	0	0	0
T8	Población	2,342	0	0
T9	Electricidad	0	0	0
T10	Ciencia	0	0	0
T11	Social	0	2,345	0
T12	Luz	1,975	0	0
T13	Unamuno	4,543	2,135	3,452
T14	Física	0	0	0
T15	Fluidos	6,134	0	0
T16	Literatura	2,234	3,456	4,234
Proceso de equiparación mediante el producto escalar de pesos TF-IDF				
Sim(doc1,q) = ^{Clima}(1,452*0) + ^{Biblioteca}(0*1,345) + ^{Universidad}(2,122*1,453) + ^{Alcalá}(3,564*1,987) + ^{España}(4,123*0) + ^{Libros}(0*2,133) + ^{Geografía}(0*0) + ^{Población}(2,342*0) + ^{Electricidad}(0*0) + ^{Ciencia}(0*0) + ^{Social}(0*0) + ^{Luz}(1,975*0) + ^{Unamuno}(4,543*3,452) + ^{Física}(0*0) + ^{Fluidos}(6,134*0) + ^{Literatura}(2,234*4,234) = 3,083 + 7,082 + 15,682 + 9,459 = **35,306**				

Sim(doc2,q) = ^{Clima}(0*0) + ^{Biblioteca}(2,093*1,345) + ^{Universidad}(0*1,453) + ^{Alcalá}(0*1,987) + ^{España}(4,245*0) + ^{Libros}(1,234*2,133) + ^{Geografía}(0*0) + ^{Población}(0*0) + ^{Electricidad}(0*0) + ^{Ciencia}(0*0) + ^{Social}(2,345*0) + ^{Luz}(0*0) + ^{Unamuno}(2,135*3,452) + ^{Física}(0*0) + ^{Fluidos}(0*0) + ^{Literatura}(3,456*4,234) = 2,815 + 2,632 + 7,370 + 14,633 = **27,450**

Tabla 29. Producto escalar de pesos TF-IDF

El cálculo de la similaridad se aplica a cada uno de los documentos de la colección siguiendo el patrón expuesto en la *tabla29*. Para el documento1 la similaridad con respecto a la consulta del usuario q, será diferente que para el documento2. Obsérvese que al igual que ocurria con los pesos binarios, sólo tienen incidencia aquellos términos presentes tanto en la consulta como en el documento, pues sus pesos se multiplican y se suman sucesivamente al resto. En este caso, la similaridad del documento1 (35,306) es superior a la del documento2 (27,450), siendo éstas unas cifras mucho más precisas que un simple número entero.

Proceso de equiparación mediante la fórmula del coseno

Tal como se ha explicado en la fórmula del producto escalar, el proceso de equiparación es posible cuando en el vector de la consulta y en el del documento existen términos coincidentes. Pero este enfoque no supone la representación del vector de la consulta y del documento. De hecho una de las claves del modelo de espacio vectorial es precisamente la posibilidad de determinar el ángulo que forman los vectores del documento y de la consulta que se está comparando, véase *figura18*.

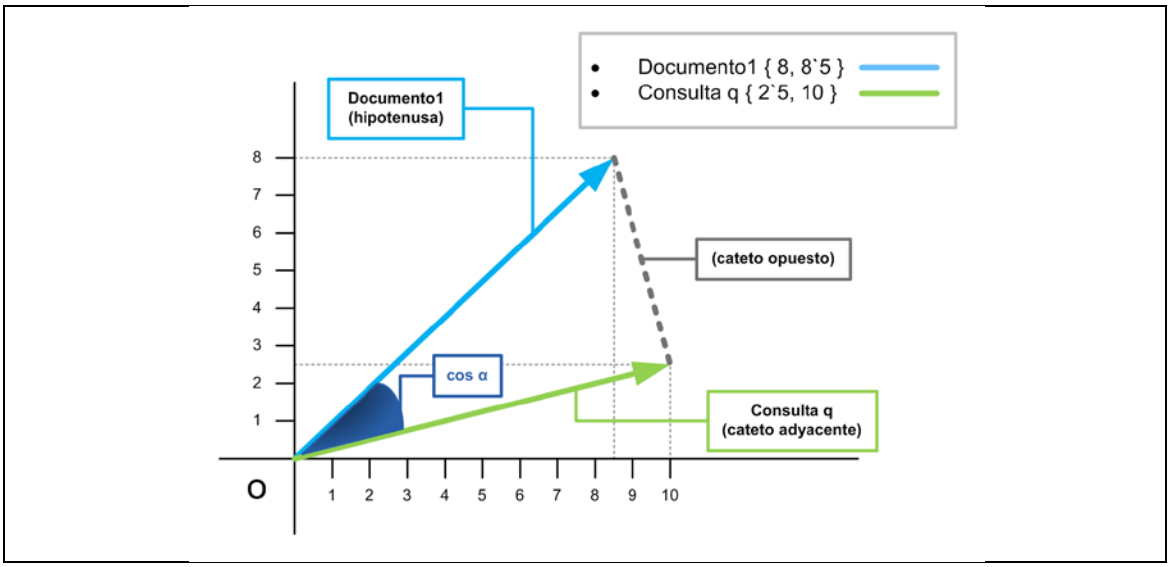


Figura 18. El ángulo del coseno

Es posible medir cuál es la desviación de un documento con respecto a una consulta, por el número de grados del ángulo que forman. Esto es posible porque crean una estructura triangular a la que se aplica el cálculo del ángulo que forma la hipotenusa (en este caso el vector del documento1) y el adyacente (el vector q de la consulta dada por el usuario) que resulta ser el coseno del triángulo. En el caso de la *figura19*, se comprueba visualmente, cierta distancia del vector de la consulta con respecto al documento1; cuando ambos vectores se muestran tan próximos como para superponerse, implicará que el ángulo que forman será menor y que su nivel de coincidencia será superior. De hecho, un coseno de 0º implicaría una similaridad máxima.

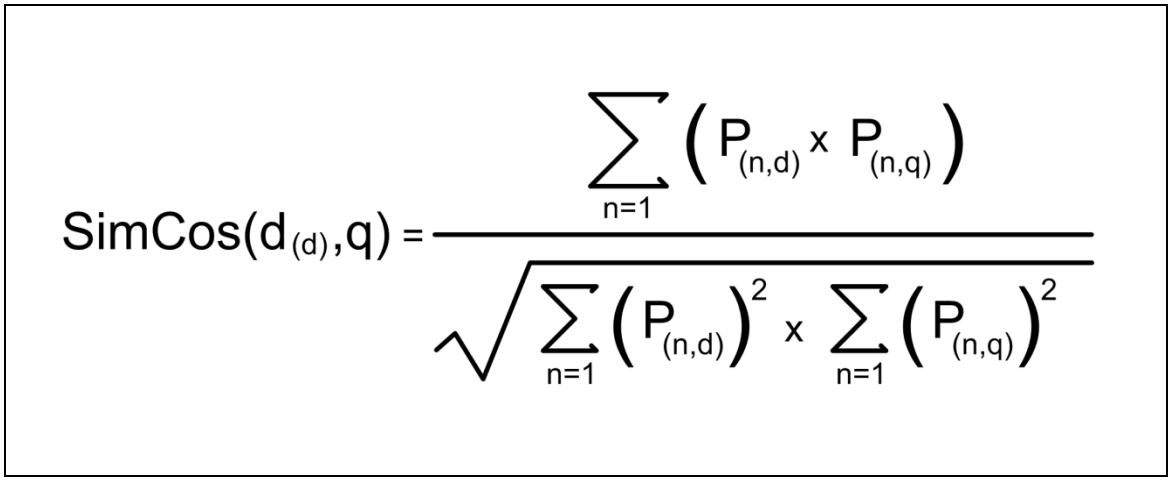


Figura 19. Fórmula para el cálculo de la similaridad del coseno

Por lo tanto, la fórmula aplicada para calcular el coeficiente de similaridad del coseno entre un documento y una consulta es aquella que permite poner en relación los vectores de la consulta y del documento. De hecho el coseno de alfa de un triángulo cualquiera siempre es igual al cateto adyacente entre la hipotenusa. Tomando como clave esa idea, la *figura19* muestra la misma relación pero esta vez con los pesos que forman los vectores del documento y la consulta. De hecho el numerador no deja de ser un producto escalar entre los pesos del documento y la consulta; y el denominador la raíz cuadrada del producto del sumatorio de los pesos del documento y la consulta al cuadrado. La formulación del denominador con raíz cuadrada y cálculo de cuadrados, se diseñó para conseguir un resultado final de la división, inferior a 1, de tal manera que el coeficiente fuera de fácil manejo y lectura. La similaridad del coseno aplicada al ejemplo que se viene utilizando, tendría la forma que sigue a continuación en la *tabla30*.

Cadena de consulta original del usuario				
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá				
Depuración de la consulta del usuario				
Libros Literatura Unamuno Biblioteca Universidad Alcalá				
Fichero diccionario		**Documento1**	**Documento2**	**q = pesos de la consulta del usuario**
Id	**Término**	**Peso TF-IDF**	**Peso TF-IDF**	
T1	Clima	1,452	0	0
T2	Biblioteca	0	2,093	1,345
T3	Universidad	2,122	0	1,453
T4	Alcalá	3,564	0	1,987
T5	España	4,123	4,245	0
T6	Libros	0	1,234	2,133
T7	Geografía	0	0	0
T8	Población	2,342	0	0
T9	Electricidad	0	0	0
T10	Ciencia	0	0	0
T11	Social	0	2,345	0
T12	Luz	1,975	0	0
T13	Unamuno	4,543	2,135	3,452
T14	Física	0	0	0
T15	Fluidos	6,134	0	0
T16	Literatura	2,234	3,456	4,234

Proceso de equiparación mediante el producto escalar de pesos TF-IDF

SimCos(doc1,q)

$$= \frac{35{,}306}{\sqrt{(1{,}452)^2+(2{,}122)^2+(3{,}564)^2+(4{,}123)^2+(2{,}342)^2+(1{,}975)^2+(4{,}543)^2+(6{,}134)^2+(2{,}234)^2} \times \sqrt{(1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2}}$$

$$= \frac{35{,}306}{\sqrt{(2{,}108)+(4{,}503)+(12{,}702)+(16{,}999)+(5{,}485)+(3{,}901)+(20{,}639)+(37{,}656)+(4{,}991)} \times \sqrt{(1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927)}}$$

$$= \frac{35{,}306}{\sqrt{108{,}984} \times \sqrt{42{,}261}} = \frac{35{,}306}{10{,}440 \times 6{,}501} = \frac{35{,}306}{67{,}870} = 0{,}520$$

SimCos(doc2,q)

$$= \frac{27{,}450}{\sqrt{(2{,}093)^2+(4{,}245)^2+(1{,}234)^2+(2{,}345)^2+(2{,}135)^2+(3{,}456)^2} \times \sqrt{(1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2}}$$

$$= \frac{27{,}450}{\sqrt{(4{,}381)+(18{,}020)+(1{,}523)+(5{,}499)+(4{,}558)+(11{,}944)} \times \sqrt{(1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927)}}$$

$$= \frac{27{,}450}{\sqrt{45{,}925} \times \sqrt{42{,}261}} = \frac{27{,}450}{6{,}777 \times 6{,}501} = \frac{27{,}450}{44{,}057} = 0{,}623$$

Tabla 30. Cálculo del coeficiente de similaridad del coseno

Como se puede observar en los resultados del coeficiente de similaridad del coseno para el documento1 y 2 en la *tabla30*, son diametralmente distintos a los obtenidos en la *tabla29*. Esto significa que los pesos de los términos del documento2, lo convierten en más representativo y probablemente más relevante que el documento1, dando por lo tanto una mayor precisión que el cálculo del producto escalar. El máximo valor del coeficiente de similaridad del coseno es 1, que equivaldría a un ángulo de 0º entre los vectores del documento y la consulta.

Proceso de equiparación mediante el coeficiente de Dice

El cálculo del coeficiente de similaridad según Lee Raymond Dice es una adaptación del cálculo del coeficiente del coseno. La diferencia en la formulación estriba en que la cardinalidad del numerador es 2 veces la información compartida y el denominador la suma de los pesos al cuadrado del documento y su consulta. Véase *figura20* y *tabla31*.

$$\text{SimDice}(d_{(d)},q) = \frac{2 \times \sum_{n=1} \left(P_{(n,d)} \times P_{(n,q)} \right)}{\sqrt{\sum_{n=1} \left(P_{(n,d)} \right)^2 + \sum_{n=1} \left(P_{(n,q)} \right)^2}}$$

Figura 20. Fórmula para el cálculo del coeficiente de similaridad de Dice

Proceso de equiparación mediante el coeficiente de Dice

SimDice(doc1,q)

$$= \frac{2 \times 35{,}306}{\sqrt{(1{,}452)^2+(2{,}122)^2+(3{,}564)^2+(4{,}123)^2+(2{,}342)^2+(1{,}975)^2+(4{,}543)^2+(6{,}134)^2+(2{,}234)^2} + \sqrt{(1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2}}$$

$$= \frac{70{,}612}{\sqrt{(2{,}108)+(4{,}503)+(12{,}702)+(16{,}999)+(5{,}485)+(3{,}901)+(20{,}639)+(37{,}656)+(4{,}991)} + \sqrt{(1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927)}}$$

$$= \frac{70{,}612}{\sqrt{108{,}984} + \sqrt{42{,}261}} = \frac{70{,}612}{10{,}440 + 6{,}501} = \frac{70{,}612}{16{,}941} = 4{,}168$$

SimDice(doc2,q)

$$= \frac{2 \times 27{,}450}{\sqrt{(2{,}093)^2+(4{,}245)^2+(1{,}234)^2+(2{,}345)^2+(2{,}135)^2+(3{,}456)^2} + \sqrt{(1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2}}$$

$$= \frac{54{,}900}{\sqrt{(4{,}381)+(18{,}020)+(1{,}523)+(5{,}499)+(4{,}558)+(11{,}944)} + \sqrt{(1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927)}}$$

$$= \frac{54{,}900}{\sqrt{45{,}925} + \sqrt{42{,}261}} = \frac{54{,}900}{6{,}777 + 6{,}501} = \frac{54{,}900}{13{,}278} = 4{,}135$$

Tabla 31. Cálculo del coeficiente de similaridad de Dice

Proceso de equiparación mediante el coeficiente de Jaccard

El cálculo del coeficiente de similaridad de Jaccard* al igual que el de Dice, resultan deudores del coeficiente de similaridad del coseno. Su aplicación, centrada en usos estadísticos, también se aplica a recuperación de información y mide la similitud entre conjuntos. Se puede definir como el tamaño de la intersección (numerador) dividido por el tamaño de la unión de la muestra, en este caso la suma de los pesos al cuadrado del documento y la consulta menos la intersección, véase *figura21* y *tabla32*.

$$SimJacc(d_{(d)},q) = \frac{\sum_{n=1}\left(P_{(n,d)} \times P_{(n,q)}\right)}{\sum_{n=1}\left(P_{(n,d)}\right)^2 + \sum_{n=1}\left(P_{(n,q)}\right)^2 - \sum_{n=1}\left(P_{(n,d)} \times P_{(n,q)}\right)}$$

Figura 21. Fórmula para el cálculo del coeficiente de similaridad de Jaccard

Proceso de equiparación mediante el coeficiente de Jaccard (Tanimoto)

SimJaccard(doc1,q)

$$= \frac{35{,}306}{(1{,}452)^2+(2{,}122)^2+(3{,}564)^2+(4{,}123)^2+(2{,}342)^2+(1{,}975)^2+(4{,}543)^2+(6{,}134)^2+(2{,}234)^2 + (1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2 - 35{,}306}$$

$$= \frac{35{,}306}{(2{,}108)+(4{,}503)+(12{,}702)+(16{,}999)+(5{,}485)+(3{,}901)+(20{,}639)+(37{,}656)+(4{,}991) + (1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927) - 35{,}306}$$

$$= \frac{35{,}306}{108{,}984 + 42{,}261 - 35{,}306} = \frac{35{,}306}{115{,}939} = 0{,}305$$

SimJaccard(doc2,q)

$$= \frac{27{,}450}{(2{,}093)^2+(4{,}245)^2+(1{,}234)^2+(2{,}345)^2+(2{,}135)^2+(3{,}456)^2 + (1{,}345)^2+(1{,}453)^2+(1{,}987)^2+(2{,}133)^2+(3{,}452)^2+(4{,}234)^2 - 27{,}450}$$

$$= \frac{27{,}450}{(4{,}381)+(18{,}020)+(1{,}523)+(5{,}499)+(4{,}558)+(11{,}944) + (1{,}809)+(2{,}111)+(3{,}948)+(4{,}550)+(11{,}916)+(17{,}927) - 27{,}450}$$

$$= \frac{27{,}450}{45{,}925 + 42{,}261 - 27{,}450} = \frac{27{,}450}{60{,}736} = 0{,}452$$

Tabla 32. Cálculo del coeficiente de similaridad de Jaccard

Ventajas del modelo vectorial

- El modelo vectorial es muy versátil y eficiente a la hora de generar rankings de precisión en colecciones de gran tamaño, lo que le hace idóneo para determinar la equiparación parcial de los documentos.
- Tiene en cuenta los pesos TF-IDF para determinar la representatividad de los documentos de la colección.

Inconvenientes del modelo vectorial

- El modelo vectorial por producto escalar tiene la desventaja de que sólo tiene en cuenta la intersección de los términos del documento con respecto a la consulta, por lo que la gradación de los resultados no es tan precisa como en el caso del cálculo del coseno.
- Necesita de la intersección de los términos de la consulta con los documentos, en caso contrario no se produce la recuperación de información.
- Al ser un modelo estadístico-matemático, no tiene en cuenta la estructura sintáctico-semántica del lenguaje natural.

9. Modelo probabilístico

Desarrollado por Robertson y Sparck Jones, fue introducido entre 1977 y 1979 y es conocido como modelo probabilístico ó de independencia binaria (BIR). Se fundamenta en la representación binaria de los documentos, al igual que en el modelo de recuperación booleano, indicando presencia o ausencia de términos mediante 0 y 1. Su diferencia radica en el método estadístico y en las premisas bajo las que se constituye su funcionamiento estableciendo las siguientes aseveraciones:

- Según la consulta planteada por el usuario, los documentos de la colección se clasifican en dos grupos; 1) Conjunto de Documentos Relevantes y 2) Conjunto de Documentos Irrelevantes.

- Existe una respuesta ideal del sistema, constituida por el conjunto de documentos relevantes, a la que se denomina Conjunto de Respuesta Ideal.

- Existe una Consulta Ideal, que es aquella que proporciona un Conjunto de Respuesta Ideal o lo que es lo mismo el conjunto de documentos relevantes para el usuario.

- Aunque a priori se desconoce cuál es la Consulta Ideal (el usuario no tiene porqué conocerla), sí se sabe que es una combinación de 0 y 1 por ser un modelo binario de recuperación. Se desconocen por tanto los términos que se deberían introducir para obtener el Conjunto de Respuesta Ideal.

Ponderación

El objetivo del modelo probabilístico es tomar la consulta del usuario para ser refinada sucesivamente hasta obtener el conjunto de respuesta ideal, mediante la reformulación sucesiva de los términos de su consulta, empleando para ello la ponderación de los términos. Esto significa que se modifican los valores 1 (presencia) por un número (peso) que permita acercar la consulta imperfecta a una consulta ideal. El proceso de ponderación de los términos de la consulta es el cálculo de probabilidad de que exista

dicho término en el conjunto de los documentos relevantes y la probabilidad de que se encuentre presente en el conjunto de los documentos irrelevantes. Véase *figura22*.

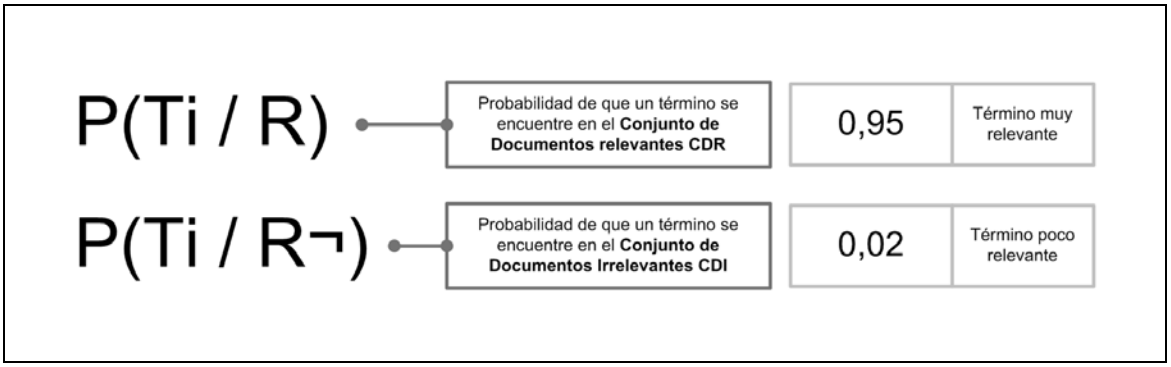


Figura 22. El cálculo de probabilidades como base para la ponderación de los términos

El método por defecto para el cálculo de pesos de los términos de la consulta se puede llevar a cabo mediante la razón de Odds. Es decir, la probabilidad de que el término aparezca en el conjunto de documentos relevantes entre la probabilidad de que el término aparezca en el conjunto de términos irrelevantes, véase *figura23*.

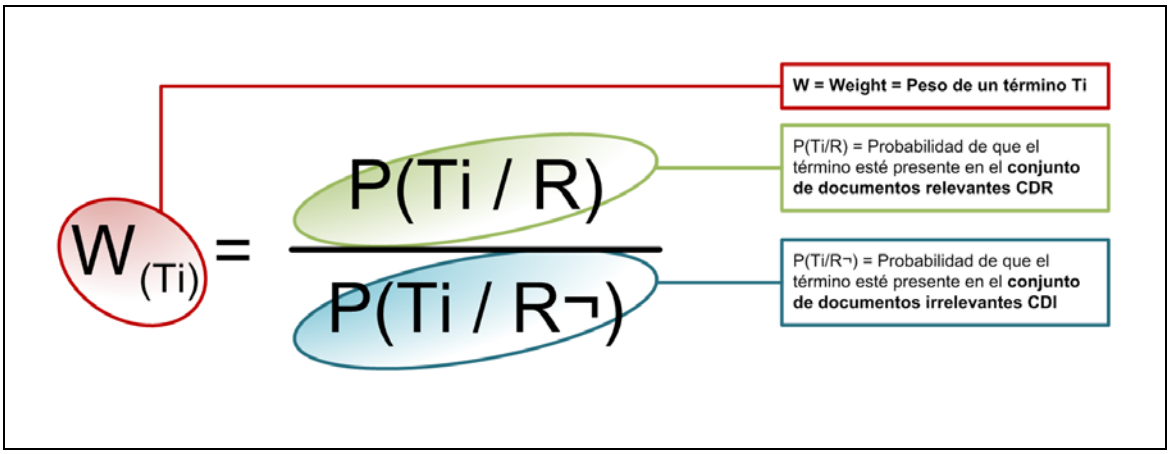


Figura 23. Ratio Odds aplicado al cálculo de pesos de los términos de la consulta del usuario

Esta formulación requiere de un mayor control de precisión, debe observarse que inicialmente se desconoce cuál es el número de documentos relevantes e irrelevantes que conforman la colección. Esta situación, particularmente compleja de averiguar a priori, se resuelve, concediendo unos valores iniciales por defecto, denominados de "*Máxima incertidumbre*". Para la probabilidad de P(Ti/R) se le asigna el valor 0,5 que es intermedio entre 0 y 1 para indicar que la probabilidad de que el término se encuentre entre los documentos relevantes e irrelevantes es la misma, por ello se denomina de máxima incertidumbre. Para la probabilidad de P(Ti/R¬) se asigna el cociente de dividir

la frecuencia de aparición del término en los documentos de la colección, entre el número total de documentos de la colección, véase *figura24*.

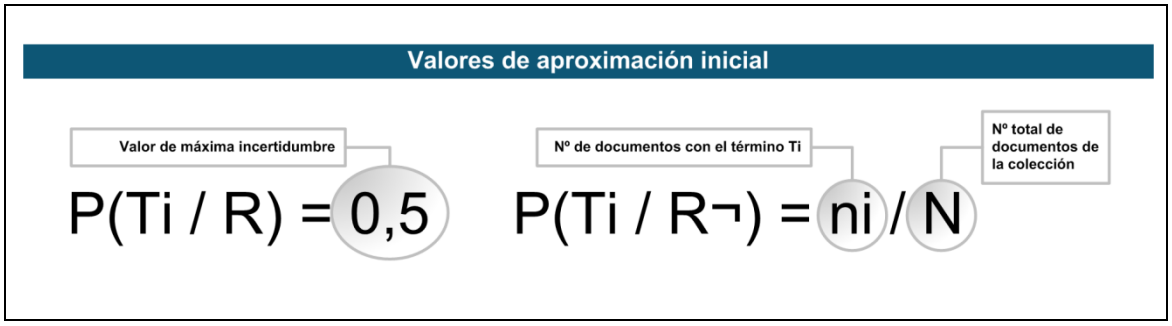


Figura 24. Asignación de valores de aproximación inicial

Pero existen más factores que pueden afectar al peso final del término de la consulta del usuario, por ejemplo, cuando se tiene en cuenta que la probabilidad de la relevancia se basa tanto en la presencia como en la ausencia de los términos de la consulta y en la independencia de la distribución de los términos dentro del conjunto de documentos relevantes. En tal caso, se utiliza una formulación derivada para el cálculo de los pesos, que pone en relación el factor independencia de las distribuciones de términos en documentos relevantes, de presencia por relevancia e irrelevancia de una forma mucho más precisa, véase *figura25*.

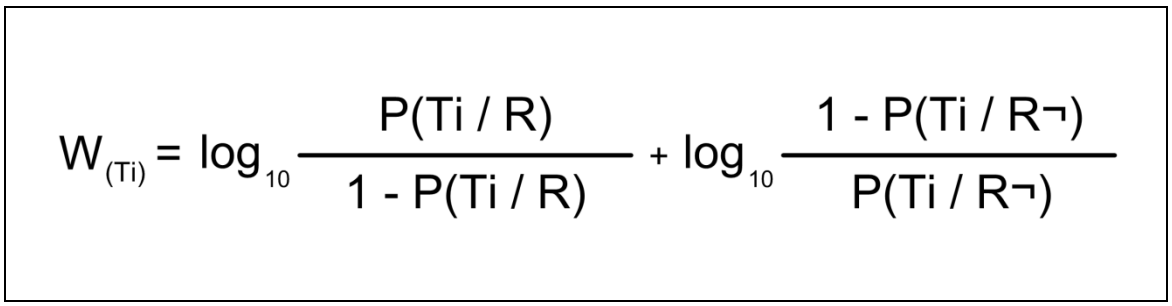


Figura 25. Método estándar para el cálculo de pesos de los términos de la consulta en el modelo probabilístico de independencia binaria

El cálculo del peso para el término de la consulta "*Ti*" de la figura4, incluye la suma de logaritmos de las probabilidades de presencia y ausencia en los conjuntos de documentos relevantes CDR (primera parte de la ecuación) y las probabilidades de presencia y ausencia en los conjuntos de documentos irrelevantes CDI (segunda parte de la ecuación). Aplicando los valores de aproximación inicial propuestos anteriormente, su formulación se asemejaría a la que se muestra en la *figura26*.

$$W_{(Ti)} = \log_{10} \frac{0,5}{1 - 0,5} + \log_{10} \frac{1 - (ni / N)}{(ni / N)}$$

Figura 26. Asignación de valores de aproximación al método estándar

El cálculo de la similaridad

Para cuantificar la similaridad de los documentos de la colección con la consulta expresada por el usuario se emplea la siguiente formulación, véase *figura27*, que pone en relación el peso de los términos de la consulta del usuario con los del documento. Se trata de una variante del cálculo de similaridad mediante el producto escalar, en la que el único elemento variable es el peso de la consulta.

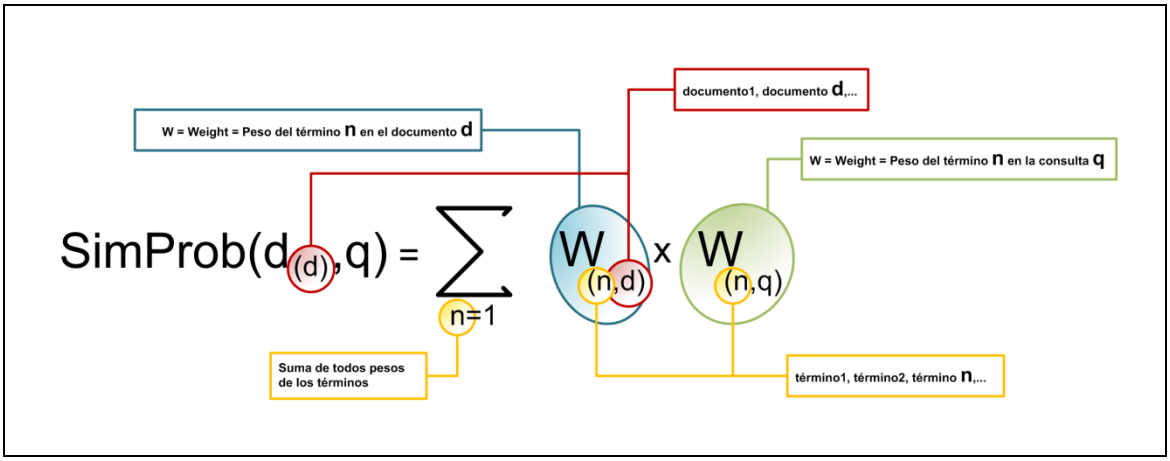


Figura 27. Cálculo de similaridad del modelo probabilístico

Una vez calculada la similaridad entre la consulta y los documentos de la colección, el sistema 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. Dicho de otra forma, se mostrará en primer lugar el documento cuya probabilidad de relevancia sea más alta. El modelo probabilístico, amplía su mecanismo de funcionamiento una vez ofrecidos los resultados al usuario, pidiendo su intervención para que señale la relevancia de los documentos. De esta forma el sistema ajusta mejor el CDR y el CDI, anteriormente mencionados, efectuando una nueva consulta que mejora y adapta el cálculo de los pesos de la consulta. Esta reformulación para el

cálculo de los pesos consiste en asignar a la probabilidad de P(Ti/R) el cociente del número de documentos relevantes en los que se encuentra el término de consulta entre el número de documentos relevantes señalados por el usuario. A la probabilidad de P(Ti/R¬) se le asigna el cociente del número total de documento que tiene el término de consulta menos el número de documentos relevantes en los que se encuentra el término de consulta, entre el número total de documentos irrelevantes menos el número de documentos relevantes señalados por el usuario. Véase *figura28*.

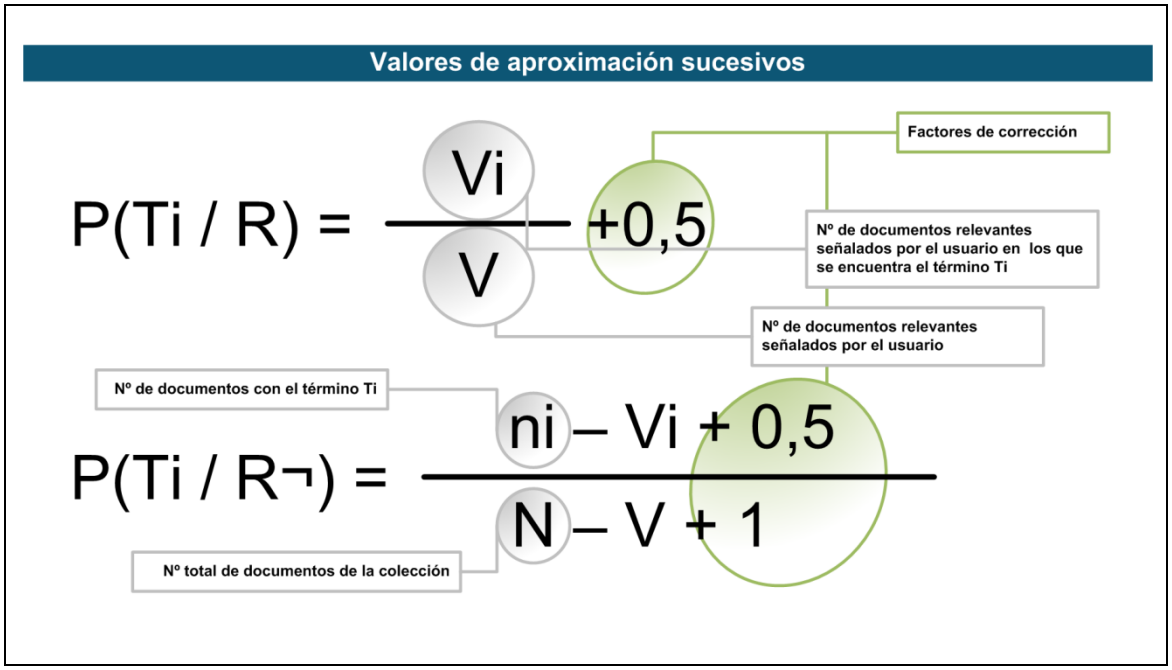


Figura 28. Asignación de valores de aproximación sucesivos

En el momento en el que el sistema asigna valores de aproximación sucesivos, se considera que se está retroalimentando con la información proporcionada por el usuario, este fenómeno se denomina, "retroalimentación por relevancia", lo que le permite calcular nuevos pesos para los términos de la consulta y aproximarse al conjunto de respuesta ideal. Este proceso de retroalimentación puede repetirse consecutivamente tantas veces como se configure en el sistema, aunque resulta habitual la repetición de 1 a 2 ciclos.

Ejemplo de aplicación

En la tabla32, se muestra la colección de prueba anteriormente utilizada para explicar otros modelos, como objeto de análisis probabilístico. Para ello obsérvese que se necesita conocer el número total de documentos que compone la colección y el número de documentos en los que aparecen los términos de la consulta del usuario. Ambos factores son esenciales para calcular los pesos de la consulta en una primera aproximación, aplicando el principio de máxima incertidumbre.

Cadena de consulta original del usuario					
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá					
Depuración de la consulta del usuario					
Libros Literatura Unamuno Biblioteca Universidad Alcalá					
Número total de documentos de la colección: 5235					
Fichero diccionario		**Documento1**	**Documento2**	**q = pesos de la consulta del usuario (aprox. inicial)**	**Frecuencia de aparición de q en la colección (Nº de docs)**
Id	**Término**	**Peso Binario**	**Peso Binario**		
T1	Clima	1	0	0	-
T2	Biblioteca	0	1	1,54	149
T3	Universidad	1	0	1,337	232
T4	Alcalá	1	0	1,954	55
T5	España	1	1	0	-
T6	Libros	0	1	0,508	1241
T7	Geografía	0	0	0	-
T8	Población	1	0	0	-
T9	Electricidad	0	0	0	-
T10	Ciencia	0	0	0	-
T11	Social	0	1	0	-
T12	Luz	1	0	0	-
T13	Unamuno	1	1	2,219	34
T14	Física	0	0	0	-
T15	Fluidos	1	0	0	-
T16	Literatura	1	1	0,378	1543
Cálculo de valores de aproximación inicial					

$$W_{T2\ (Biblioteca)} = \log 10 \frac{0,5}{1-0,5} + \log 10 \frac{1-\left(\frac{149}{5235}\right)}{\left(\frac{149}{5235}\right)} = 0 + \log 10 \left(\frac{0,972}{0,028}\right) = \log 10(34,714) = 1,54$$

$$W_{T3\ (Universidad)} = \log 10 \frac{0,5}{1-0,5} + \log 10 \frac{1-\left(\frac{232}{5235}\right)}{\left(\frac{232}{5235}\right)} = 0 + \log 10 \left(\frac{0,956}{0,044}\right) = \log 10(21,727) = 1,337$$

$$W_{T4\ (Alcalá)} = \log 10 \frac{0,5}{1-0,5} + \log 10 \frac{1-\left(\frac{55}{5235}\right)}{\left(\frac{55}{5235}\right)} = 0 + \log 10 \left(\frac{0,99}{0,011}\right) = \log 10(90) = 1,954$$

$$W_{T6\ (Libros)} = \log 10 \frac{0,5}{1-0,5} + \log 10 \frac{1-\left(\frac{1241}{5235}\right)}{\left(\frac{1241}{5235}\right)} = 0 + \log 10 \left(\frac{0,763}{0,237}\right) = \log 10(3,219) = 0,508$$

$$W_{T13\ (Unamuno)} = \log 10 \frac{0,5}{1-0,5} + \log 10 \frac{1-\left(\frac{34}{5235}\right)}{\left(\frac{34}{5235}\right)} = 0 + \log 10 \left(\frac{0,994}{0,006}\right) = \log 10(165,667) = 2,219$$

$$W_{T16\ (Literatura)} = \log 10 \frac{0,5}{1-0,5} + \log 10 \frac{1-\left(\frac{1543}{5235}\right)}{\left(\frac{1543}{5235}\right)} = 0 + \log 10 \left(\frac{0,705}{0,295}\right) = \log 10(2,39) = 0,378$$

Cálculo de similaridad (Primera vuelta)
Sim(doc1,q) = $^{\text{Universidad}}$(1*1,337) + $^{\text{Alcalá}}$(1*1,954) + $^{\text{Unamuno}}$(1*2,219) + $^{\text{Literatura}}$(1*0,378) = **5,888**
Sim(doc2,q) = $^{\text{Biblioteca}}$(1*1,54) + $^{\text{Libros}}$(1*0,508) + $^{\text{Unamuno}}$(1*2,219) + $^{\text{Literatura}}$(1*0,378) = **4,645**

Tabla 33. Cálculo de similaridad según el método probabilístico

Como se observa los documentos de la colección pueden ser representados mediante un vector binario de 0 y 1, lo que implica un cálculo más sencillo de similaridad que se limitará a la suma de los pesos de los términos de las consultas presentes en cada documento de la colección. Aún así, la precisión es muy buena, si se compara con los resultados obtenidos con la formula de producto escalar en el modelo vectorial. Al igual que en aquel caso, también se pueden utilizar vectores de los documentos, conformados por pesos TF-IDF, obteniendo cifras superiores a las mostradas en la *tabla33*.

En la *tabla34*, se muestra un ejemplo del proceso de retroalimentación, en la que después de la primera vuelta de ejecución del sistema se ofrecen unos resultados que deberán ser evaluados por el usuario. Entre todos los resultados inspeccionados el usuario marca como relevantes 15 de ellos, de entre los cuales puede estar presente o no el término de la consulta, inicialmente utilizado. Por ejemplo el término "Biblioteca" aparece en 5 de los 15 resultados marcados como relevantes para el usuario. El sistema se vale de esta información, para refinar la consulta, asignando nuevos pesos, mediante la fórmula expresada en las *figuras24 y 28*. Como resultado de la precisión del usuario, los coeficientes y en definitiva el ordenamiento de los documentos de la colección resulta más exacto ordenando en sentido decreciente los resultados cuya probabilidad de relevancia sea mayor.

Cadena de consulta original del usuario							
Los libros y la literatura de Unamuno en la biblioteca de la Universidad de Alcalá							
Depuración de la consulta del usuario							
Libros Literatura Unamuno Biblioteca Universidad Alcalá							
Número total de documentos de la colección (N): 5235 // **Documentos relevantes para el usuario (V):** 15							
Fichero diccionario		**Doc1**	**Doc2**	**q = pesos de la consulta del usuario (aprox. inicial)**	**q = pesos refinados (2ª vuelta)**	**Frecuencia de aparición de q en la colección (ni)**	**Documentos relevantes para el usuario con presencia del término de consulta (Vi)**
Id	**Término**	**Peso Binario**	**Peso Binario**				
T1	Clima	1	0	0	0	-	-
T2	Biblioteca	0	1	1,54	2,255	149	5
T3	Universidad	1	0	1,337	2,035	232	5
T4	Alcalá	1	0	1,954	2,995	55	6
T5	España	1	1	0	0	-	-
T6	Libros	0	1	0,508	1,462	1241	6
T7	Geografía	0	0	0	0	-	-
T8	Población	1	0	0	0	-	-
T9	Electricidad	0	0	0	0	-	-
T10	Ciencia	0	0	0	0	-	-
T11	Social	0	1	0	0	-	-
T12	Luz	1	0	0	0	-	-
T13	Unamuno	1	1	2,219	2,917	34	5
T14	Física	0	0	0	0	-	-
T15	Fluidos	1	0	0	0	-	-
T16	Literatura	1	1	0,378	0,746	1543	3

Cálculo de valores sucesivos de aproximación

$$W_{T2\ (Biblioteca)} = \log 10 \frac{\left(\frac{5}{15} + 0{,}5\right)}{1 - \left(\frac{5}{15} + 0{,}5\right)} + \log 10 \frac{1 - \left(\frac{149 - 5 + 0{,}5}{5235 - 15 + 1}\right)}{\left(\frac{149 - 5 + 0{,}5}{5235 - 15 + 1}\right)} = \log 10 \left(\frac{0{,}833}{0{,}167}\right) + \log 10 \left(\frac{0{,}973}{0{,}027}\right) = 2{,}255$$

$$W_{T3\ (Universidad)} = \log 10 \frac{\left(\frac{5}{15} + 0{,}5\right)}{1 - \left(\frac{5}{15} + 0{,}5\right)} + \log 10 \frac{1 - \left(\frac{232 - 5 + 0{,}5}{5235 - 15 + 1}\right)}{\left(\frac{232 - 5 + 0{,}5}{5235 - 15 + 1}\right)} = \log 10 \left(\frac{0{,}833}{0{,}167}\right) + \log 10 \left(\frac{0{,}956}{0{,}044}\right) = 2{,}035$$

$$W_{T4\ (Alcalá)} = \log 10 \frac{\left(\frac{6}{15} + 0{,}5\right)}{1 - \left(\frac{6}{15} + 0{,}5\right)} + \log 10 \frac{1 - \left(\frac{55 - 6 + 0{,}5}{5235 - 15 + 1}\right)}{\left(\frac{55 - 6 + 0{,}5}{5235 - 15 + 1}\right)} = \log 10 \left(\frac{0{,}9}{0{,}1}\right) + \log 10 \left(\frac{0{,}991}{0{,}009}\right) = 2{,}995$$

$$W_{T6\ (Libros)} = \log 10 \frac{\left(\frac{6}{15} + 0{,}5\right)}{1 - \left(\frac{6}{15} + 0{,}5\right)} + \log 10 \frac{1 - \left(\frac{1241 - 6 + 0{,}5}{5235 - 15 + 1}\right)}{\left(\frac{1241 - 6 + 0{,}5}{5235 - 15 + 1}\right)} = \log 10 \left(\frac{0{,}9}{0{,}1}\right) + \log 10 \left(\frac{0{,}763}{0{,}237}\right) = 1{,}462$$

$$W_{T13\ (Unamuno)} = \log 10 \frac{\left(\frac{5}{15} + 0{,}5\right)}{1 - \left(\frac{5}{15} + 0{,}5\right)} + \log 10 \frac{1 - \left(\frac{34 - 5 + 0{,}5}{5235 - 15 + 1}\right)}{\left(\frac{34 - 5 + 0{,}5}{5235 - 15 + 1}\right)} = \log 10 \left(\frac{0{,}833}{0{,}167}\right) + \log 10 \left(\frac{0{,}994}{0{,}006}\right) = 2{,}917$$

$$W_{T16\ (Literatura)} = \log 10 \frac{\left(\frac{3}{15} + 0{,}5\right)}{1 - \left(\frac{3}{15} + 0{,}5\right)} + \log 10 \frac{1 - \left(\frac{1543 - 3 + 0{,}5}{5235 - 15 + 1}\right)}{\left(\frac{1543 - 3 + 0{,}5}{5235 - 15 + 1}\right)} = \log 10 \left(\frac{0{,}7}{0{,}3}\right) + \log 10 \left(\frac{0{,}705}{0{,}295}\right) = 0{,}746$$

Cálculo de similaridad (Segunda vuelta)

Sim(doc1,q) = $^{\text{Universidad}}(1*2{,}035) + {}^{\text{Alcalá}}(1*2{,}995) + {}^{\text{Unamuno}}(1*2{,}917) + {}^{\text{Literatura}}(1*0{,}746) =$ **8,693**

Sim(doc2,q) = $^{\text{Biblioteca}}(1*2{,}255) + {}^{\text{Libros}}(1*1{,}462) + {}^{\text{Unamuno}}(1*2{,}917) + {}^{\text{Literatura}}(1*0{,}746) =$ **7,38**

Tabla 34. Ejemplo de retroalimentación por relevancia

Ventajas del modelo probabilístico

- Retroalimentación por relevancia, acepta feedback.
- Asume la independencia de los términos de la consulta.
- Asigna pesos a los términos, permitiendo recuperar los documentos que probablemente sean relevantes.

- Es considerado uno de los mejores modelos dados sus buenos resultados con colecciones reales y corpus de entrenamiento.
- Su método de recuperación es mediante equiparación parcial, superando al método de equiparación exacta del modelo booleano.

Inconvenientes del modelo probabilístico

- Mantiene el modelo binario de recuperación de información, no teniendo en cuenta todos los términos del documento como ocurriría en el modelo vectorial.
- Asigna pesos a los términos, permitiendo recuperar los documentos que probablemente sean irrelevante.
- Requiere alta capacidad de computación, resultando complejo de implementar.
- Necesita efectuar una hipótesis inicial que no siempre resulta acertada.
- No tiene en cuenta la frecuencia de aparición de cada término en el documento, tal como lo haría un modelo vectorial.

10. Ejercicios prácticos

- Práctica1. Preparación de la semilla
- Práctica2. Generando la colección
- Práctica3. Depuración de la colección
- Práctica4. Indexación y recuperación con Lemur
- Práctica5. Calculando pesos TF-IDF
- Práctica6. Probando el modelo booleano
- Práctica7. Prueba manual del modelo vectorial
- Práctica8. Prueba automática del modelo vectorial
- Práctica9. Prueba manual del modelo probabilístico
- Práctica10. Prueba automática del modelo probabilístico
- Práctica11. Método de evaluación de un sistema de recuperación de información

Práctica1. Preparación de la semilla

Todo el proceso comienza con la preparación de un archivo de semilla (también denominado seed file) formado por un listado de direcciones URL, correspondientes a un dominio, ámbito ó área de conocimiento. Dependiendo de la cobertura temático-geográfica así se obtendrá una colección más o menos heterogénea. En este caso, se propone la acotación temática para diseñar un buscador especializado en áreas de conocimiento científicas.

1. Seleccionar un paquete de áreas de conocimiento, de los que se muestran a continuación.

Selección	Paquete	Áreas de conocimiento
	Pack1. Technology and Engineering	Computer Science (317)
	Pack2. Technology and Engineering	Chemical Technology (32)
		Construction (15)
		Electrical and Nuclear Engineering (62)

	Pack3. Technology and Engineering	General and Civil Engineering (148)
	Pack4. Technology and Engineering	Environmental Engineering (9)
	Pack5. Technology and Engineering	Environmental Technology (9)
		Hydraulic Engineering (4)
		Industrial Engineering (18)
		Manufactures (10)
		Materials (37)
		Mechanical Engineering (40)
	Pack6. Technology and Engineering	Military Science (10)
		Mining and Metallurgy (14)
		Technology General (90)
		Transportation (29)
	Pack7. Science General	Science General (119)
		Information Theory (1)
	Pack8. Physics and Astronomy	Astronomy (19)
		Physics (80)
		Mathematics (203)
		Statistics (44)
		Analytical Chemistry (14)
		Chemical Engineering (16)
		Chemistry General (103)
		Inorganic Chemistry (5)
		Organic Chemistry (14)
		Biology (246)
		Biochemistry (49)

		Biotechnology (44)

2. Cada paquete y áreas de conocimientos corresponden con la clasificación de materias del DOAJ (Directory of Open Access Journals), disponible en: http://www.doaj.org/doaj?func=browse&uiLanguage=en

3. La elaboración de semillas es un proceso consistente en la confección de una lista de URLs, correspondientes al ámbito, área temática ó dominio que se pretende rastrear ó explorar posteriormente con un sistema webcrawler.

4. Por cada área de conocimiento del paquete elegido, se deberá generar una lista de URLs en documento ".txt" con sus correspondientes revistas del recurso DOAJ.

5. Comprimir en archivo ".rar" el presente documento junto con las listas .txt para su correspondiente envío.

Práctica2. Generando la colección

Generar una colección de prueba, requiere del uso de programas de tipo webcrawler, especializados en el rastreo y recopilación de sitios web. Para simular este proceso, se propone la instalación de dos programas:

- HUNFELD, U. 2011. PHPCrawl. Disponible en: http://phpcrawl.cuab.de/
  Descarga oficial:
  http://sourceforge.net/projects/phpcrawl/files/latest/download?source=files
  Descargar PHPCrawl modificado:
  http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/phpcrawler.zip

- SAABAS, A. 2009. Sphider. Disponible en: http://www.sphider.eu/
  Descarga oficial:
  http://www.sphider.eu/dl.php?file=sphider-1.3.5.zip
  Descargar Sphider modificado:
  http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/sphider.zip

Los requisitos de instalación para ambos programas son el servidor Apache HTTP, base de datos MySQL y módulo PHP. Se recomienda el uso de una distribución compacta portable Server2Go, véanse las instrucciones de instalación y configuración en: http://ccdoc-automatizacion.blogspot.com/2008/02/04-fundamentos-tecnolgicos-de-la.html

Instalación de PHPCrawl

- Descargar la versión modificada de PHPCrawl.

- Descomprimir los contenidos del programa en la carpeta "phpcrawler". (nombre del directorio en mínusculas, sin subcarpetas que aniden varios niveles el acceso a los contenidos del programa)
- Copiar la carpeta "phpcrawler" y pegar en el directorio "USB:\server2go\htdocs\phpcrawler".
- Instalación completada.

Instalación de Sphider

- Descargar la versión modificada de Sphider.
- Descomprimir los contenidos del programa en la carpeta "sphider". (nombre del directorio en mínusculas, sin subcarpetas que aniden varios niveles el acceso a los contenidos del programa)
- Copiar la carpeta "sphider" y pegar en el directorio "USB:\server2go\htdocs\sphider".
- Iniciar el servidor Server2Go y acceder al programa PhpMyAdmin, mediante URL http://127.0.0.1:4001/phpmyadmin/
- Crear base de datos vacía "sphider".
- Acceder desde el mismo navegador al programa Sphider "http://127.0.0.1:4001/sphider/". Mostrará una página de inicio con diversas opciones.
- Seleccionar la opción "Instalar Sphider" y el programa creará todas las tablas necesarias para su funcionamiento en la base de datos creada anteriormente.
- Copiar los archivos de semilla, elaborados en la práctica anterior y emplazarlos en la siguiente ubicación "USB:\server2go\htdocs\sphider\admin\seedtxt\".
- Desde la página de inicio de Sphider, seleccionar la opción "Preparar semilla" para convertir los archivos de semilla en formato txt, en sentencias sql, para su importación posterior en Sphider. Se le requerirá que seleccione un archivo para la conversión. Marque el primero y efectúe el proceso de conversión, hasta su finalización. Si existen más archivos de semilla disponibles, repita el proceso con el siguiente archivo.
- El resultado de este proceso es la creación de archivos de semilla con extensión ".sql" que serán utilizados para su importación desde el gestor de bases de datos

PhpMyAdmin. Se encuentran disponibles en la siguiente ruta "USB:\server2go\htdocs\sphider\admin\seedsql\".

- Acceder al programa PhpMyAdmin, mediante URL http://127.0.0.1:4001/phpmyadmin/ . Seleccionar la base de datos "sphider". A continuación seleccionar la tabla "sites" en la que se deberan cargar las direcciones URL de las semilla con extensión ".sql" generadas en el paso anterior.
- Obsérvese que PhpMyAdmin mostrará una pestaña con la opción "Importar". Al hacer click deberá examinar la localización del archivo a importar, que se encuentra en la ruta "USB:\server2go\htdocs\sphider\admin\seedsql\". No hace falta efectuar ningún tipo de configuración añadida. Seguidamente haga click en el botón "Continuar" y se transferirán todos los datos de la semilla.
- Instalación completada.

Instalación múltiple de Sphider

- Por motivos de evaluación científica, la práctica2 requiere de tántas instalaciones de Sphider, como semillas se generaron en la práctica1. En tal caso se repetirá el proceso enunciado en este apartado, variando el nombre de los directorios de instalación "sphider1, sphider2, sphider3,... sphiderN", así como creando las bases de datos en blanco "sphider1, sphider2,... sphiderN".
- Se deberá, modificar el archivo de configuración "USB:\server2go\htdocs\sphider\settings\database.php" para modificar el nombre de la base de datos para cada instalación.
  - $database="sphider1, sphider2, sphider3...";
  - $mysql_user = "root";
  - $mysql_password = "root";
  - $mysql_host = "localhost";

Práctica3. Depuración de la colección

Estudiados los diversos procesos de depuración de la colección, se propone la puesta en práctica de los mismos; la supresión del código fuente, la tokenización, la normalización de caracteres y la eliminación de palabras vacías. Para ello se han creado programas de simulación que permiten llevar a cabo pruebas con multitud de variantes, textos y códigos, permitiendo generar tantas experiencias como casos se deseen.

- Ejercicio de supresión del código fuente
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/depuracion01.php
- Ejercicio de tokenización
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/depuracion02.php
- Ejercicio de normalización de caracteres
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/depuracion03.php
- Ejercicio de eliminación de palabras vacías
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/depuracion04.php

Ejercicio de supresión del código fuente

1. Introduce el código fuente de la página de portada de la Universidad de Harvard.

¿Qué resultado se obtiene?
¿Depura todos los códigos?
En caso de que no depure todo el código, ¿a qué corresponde éste?
Qué técnica habría que utilizar para eliminar dicho código

2. Introduce el código fuente de la página de especificaciones de HTML 4.0 del W3C Consortium

¿Qué resultado se obtiene?
¿Depura todos los códigos?
En caso de que no depure todo el código, ¿a qué corresponde éste?
¿Resulta más limpio el resultado que el caso anterior? Razona tu respuesta

3. Elige una página web científica, especializada en documentación, con contenido textual para probar nuevamente el programa de supresión del código fuente.

¿Qué URL has elegido?
¿Qué resultado se obtiene al aplicar la depuración del código fuente?
¿Comenta lo más llamativo de los resultados?

Ejercicio de tokenización

4. Prueba los siguientes textos en el ejercicio de tokenización y responde a las preguntas que se formulan.

Texto1
What should a 21st Century library look like? To ask that question is to conjure futuristic visions--of libraries that resemble sleek Apple stores; of librarians who stroll around their branches with computer tablets, and of robots that stack books in shelves, provided, of course, there still are books. Such issues are no longer academic, not with a new library commissioner heading to Chicago, especially one from digitally-savvy San Francisco. The debut of new library commissioner Brian Bannon (above, left, with Mayor Rahm Emanuel and outgoing library commissioner Mary Dempsey), who is expected to start this month, gives Chicago a chance to think afresh about its libraries--and how good design can uplift the experience of the millions of people who use them. Bannon, it turns out, is no stranger to architecture. Before he became chief information officer for the San Francisco Public Library, he was the system's chief of branches. In that role, he managed a $200 million Branch Library Improvement Program (BLIP, for short) that has so far renovated 16 libraries and built six new ones. The upgrades sparked increases in visits and checked-out materials. Many of the libraries won LEED (Leadership in Energy & Environmental Design) certification, which suggests that Bannon should have no trouble adapting to Chicago's emphasis on energy-saving green design. He also appears in sync with Chicago's philosophy of turning libraries from imposing temples of reading into vibrant community anchors.
¿Qué pila de caracteres se obtiene como resultado? Bajar el scroll hasta el final de la página, lugar en el que aparece esta información.
Términos como "libraries--and" podrían afectar negativamente a la recuperación. Justifica tu respuesta.
Cuál es el código hexadecimal del apóstrofe en la palabra "Chicago's"

Texto2
La Biblioteca Nacional de España no es sólo la principal biblioteca del país sino la más importante de todas las que existen en los países de habla española y el primer centro informativo y documental sobre la cultura escrita hispana. A lo largo de sus 300 años –tiene categoría de museo bibliográfico desde 1858- ha sabido adaptarse a los cambios políticos y sociales del país conservando siempre su principio básico: reunir, conservar y difundir el conocimiento de sus fondos. La institución cultural más antigua del país, fue fundada por Felipe V, quien quiso que los libros y las riquezas artísticas pasaran a disposición general pública para convertirse en instrumentos de renovación de la cultura nacional. El proyecto de fundación de la Biblioteca, preparado por el confesor del monarca y primer director Pedro Robinet, fue aprobado el 29 de diciembre de 1711. La que fuera "Real Librería" (Biblioteca Nacional en 1836), celebra hoy su Tricentenario en su espléndida madurez, con la mirada puesta siempre en el futuro. Aquí mostramos la historia, sus grandes joyas, la compleja maquinaria que nunca se detiene al servicio de los sonidos, las imágenes y las palabras; pero por delante está la innovación, las nuevas tecnologías, la constante adaptación al tiempo que está por venir. La conmemoración del Tricentenario de una de las más importantes bibliotecas nacionales del mundo requiere la implicación de toda la sociedad. Con la exposición La Biblioteca Nacional de España: 300 años haciendo historia, la Biblioteca no sólo abre sus puertas sino que sale al encuentro de los ciudadanos y así lo hará a lo largo de doce meses a través de congresos, conferencias, conciertos, exposiciones… El Tricentenario es un "acontecimiento de excepcional interés público" que dejará huella en la institución y en la sociedad española porque siempre han ido de la mano y porque la BNE, es tuya.
¿Qué pila de caracteres se obtiene como resultado? Bajar el scroll hasta el final de la página, lugar en el que aparece esta información.
A tenor de los resultados obtenidos ¿qué términos podrían causar problemas en la recuperación de información si no son tratados?
Cuál es el código hexadecimal del término "Tricentenario"

Ejercicio de normalización de caracteres

5. Prueba los siguientes textos en el ejercicio y responde a las preguntas que se formulan.

Texto1
Implementación del modelo FRBR en RDA, requisitos específicos para un posible perfil europeo, RDA como estándar de contenido e internacionalización, fueron las temáticas de los debates en la reunión técnica de EURIG que se celebró el pasado 27 de enero en la

Bibliotèque Nationale de France.

El EURIG Technical Meeting sirvió para analizar propuestas de cambio y enmienda de ciertas reglas para elevar al Joint Steering Comitte, órgano encargado de la revisión de RDA.
EURIG, Grupo Europeo de Interés en RDA, está presidido por Alan Danskin (British Library). La secretaria del Grupo es Laura Peters (Koninklijke Bibliotheek) y el vicepresidente Gildas Illien (Bibliothèque Nacionale de France).

Esta reunión es de carácter técnico, y tiene por objeto el debate sobre instrucciones RDA específicas y su análisis de cara a una posible implantación de RDA en las bibliotecas europeas. La gran acogida de la reunión denota el claro interés que despierta RDA en Europa, interés que se ve reflejado en que el principal debate se ha desplazado de adoptar o no RDA a cómo adoptar RDA.

Se reúnen un total de 36 miembros de distintas bibliotecas e instituciones procedentes de Austria, Croacia, República Checa, Dinamarca, Finlandia, Francia, Alemania, Gran Bretaña, Italia, El Vaticano, Letonia, Noruega, Polonia, Eslovaquia, España, Suecia, Suiza y Holanda.

La siguiente reunión de miembros se llevará a cabo probablemente en la BNE en el mes septiembre.

¿Qué cambios se obtienen en la columna token2 y hexadecimal2 con respecto a token1 y hexadecimal1?

¿Se eliminan todos los signos de puntuación?

¿Se eliminan todos los signos diacríticos?

¿Aparece el texto en minúsculas?

Qué significa "0A"

Texto2

Powdered earthworms and lots of white wine: that's the 17th century cure for piles on display, along with other science-related documents at *Particles of the Past* , an exhibition launched at the National Library today.
As part of the Dublin City of Science 2012, the exhibit brings together a set of unusual historical documents – carefully restored maps, a recipe for making ice cream in a bucket, a journal from the voyages of Captain Cook, among others – all tied together by the theme of science.

"Whether your interests lie in early medicine, photography, engineering, or you simply have an enquiring mind, *Particles of the Past* has something for everyone," says exhibition co-curator Riona McMorrow.

The exhibit also hopes to bring more attention to the crucial nature of the library's document conservation department, which uses scientific techniques to analyze and preserve items from the past.

From using UV light to analyze pigments and fibres, to the chemistry used in conservation, the department works with the State Laboratory to make sure that manuscripts will be available for generations to come.

Aside from oddments of outdated cure-alls, visitors are also encouraged to learn how to take care of their own important documents.

"This exhibition is more than just telling us about our past," said Minister for Arts Jimmy Deenihan.

Most people have old letters, deeds passed down through several generations. Bust most people also don't know how to take carer of them, Mr Deenihan said.

¿Se eliminan todos los signos de puntuación?

¿Se eliminan todos los signos diacríticos?

¿Aparece el texto en minúsculas?

¿A qué se deben los posibles errores?

Ejercicio de eliminación de palabras vacías

6. Prueba los siguientes textos en el ejercicio y responde a las preguntas que se formulan.

Texto1

Con el cierre de Megaupload, cada día aparecen más sitios acusados de realizar actividades que se califican como piratería de archivos con copyright, y eso es precisamente lo que ocurrió hace unos días en el sitio library.nu, el cual fue reportado por la Association of American Publishers por contener 400 mil textos digitales sin permiso de distribución por parte de las editoriales, además de recaudar unos 10 millones de dólares por la publicidad que aparece en el sitio.

En total son 17 las editoriales que pidieron el cierre del sitio, entre las que se encuentran HaperCollins, Oxford University Press y Macmillan, debido a que en el sitio hay links a miles de copias ilegales de libros en archivos PDF desde diciembre de 2010. De acuerdo a la declaración de uno de los abogados de la asociación de editoriales, estos archivos se alojan en el sitio iFile.it.

Este representa el primer cierre de un sitio que contiene enlaces de libros después de que Megaupload fuese cerrado y se procesara legalmente a sus trabajadores. Según Kim Dotcom, fundador de Megaupload, el cierre del sitio que fundó representa “el mayor desprecio a los derechos humanos básicos en Internet“.

La corte encontró en el caso de library.nu, problemas legales con 10 títulos de libros por cada una de las compañías editoriales que realizaron la acusación, y cada link a esos libros podría significar una multa de 250,000 euros y unos seis meses en la cárcel, por lo que un representante de iFile.it declaró que están haciendo todo lo posible por sacar de su sitio cualquier archivo que pueda violar los derechos de autor de esas obras editoriales.

Después de cierre de library.nu, iFile.it sigue en línea pero solamente permite que los usuarios registrados sean quienes puedan subir archivos al sitio, y eliminó cualquier vínculo con los archivos que se subieron a library.nu. Por ahora, los sitios que permiten a sus usuarios subir material que podría infringir copyright, deberá de estar muy alerta o podría meterse en un problema muy grande, y eso sin que aún hayan entrado en vigor legislaciones como SOPA.

¿Qué resultados se obtienen?

¿Se eliminaron todas las palabras vacías?

En caso negativo, ¿qué término falló y a qué fue debido?

¿Cuántos términos tenía el texto original y cuántos tiene el resultado? ¿Qué porcentaje de palabras vacías se eliminó? ¿Se cumple la ley de Luhn para este caso?

Texto2

Zum Abschluss gab es eine Überraschungsgeschichte, bei der sich alle gemeinsam mit Annie Vollmers das Ende ausgedacht haben. Daran anschließend haben die Kinder in einer zweiten Malaktion die Überraschungsgeschichte gemalt.

Die Aktion wurde fortgesetzt in der Diezer Bibliothek. Die Kinder der Kita "Kinderhafen" liefen gemeinsam mit der ersten Klasse und deren Lehrerin Strumpelmeier in die Bibliothek.

"Die Bibliothek ist für alle da, ihr seid herzlich willkommen", so begrüßte die Leiterin Monika Scharf die jungen Gäste. Mit einem Guten-Morgen-Lied der ersten Klasse begrüßten sich die Kinder gemeinsam.

Nun konnte es losgehen. Annie Vollmers fing gleich mit dem "Buchtitel-Spiel" an. Es machte den Kindern riesigen Spaß. Danach durften sie in den zahlreichen Kisten und Regalen stöbern, und einige fanden auch die Bücher vom "Buchtitel-Spiel" (Pipi Langstrumpf, Sams . . .) wieder. Nun kam die vierte Klasse der Grundschule mit Frau Weidenfeller dazu. Gemeinsam mit Annie Vollmers wurde wieder gelesen und erzählt. Die Kinder der ersten Klasse hatten ihre Lieblingsbücher gemalt, die Annie Vollmers allen vorstellte. In ihrer Lesung waren Geschichten, Märchen, Zungenbrecher, Fingerspiele wie zum Beispiel "Der Bücherwurm".

Mit einem Buch-Quiz hat Annie Vollmers die Kinder mit kniffligen Fragen getestet, ob sie denn auch gut zugehört hatten. Alle Kinder bekamen eine Urkunde von der Bibliotheksleiterin. Zum Abschluss konnten alle Kinder Bilder ausmalen oder in den Bücherkisten und Regalen Bücher aussuchen und schmökern.
¿Qué resultados se obtienen?
¿Cuántos términos tenía el texto original y cuántos tiene el resultado? ¿Qué porcentaje de palabras vacías se eliminó? ¿Se cumple la ley de Luhn también para este caso?

Práctica4. Indexación y recuperación con Lemur

Hasta el momento se han estudiado y probado los distintos procedimientos para la generación de una colección de documentos, su preparación, depuración y reducción de cara a la indexación.

En la presente práctica se pondrá a prueba el funcionamiento de "Lemur", un indexador desarrollado por el Centro para la Recuperación Inteligente de Información, del Departamento de Ciencias de la Computación de la Universidad de Massachusetts Amherst. Este sistema permite indexar miles de documentos, aplicando depuración de palabras vacías y stemming basados en el algoritmo de Porter y Krovetz.

Instalación de Lemur

La versión disponible de Lemur permite su instalación en S.O. Windows, incluye autoinstalador y su paradigma de desarrollo está basado en lenguaje Java. Para poder llevar a cabo la práctica, se recomienda su instalación en el directorio "*Mis documentos (My documents)*".

Figura 29. Proyecto Lemur

- Sitio web oficial: http://www.lemurproject.org/
- Descargar Lemur 5.2 exe: http://sourceforge.net/projects/lemur/files/latest/download

Entre los contenidos instalados figurará el directorio "*User\Mis documentos\Indri 5.2\lib*" en el que se encuentran los siguientes archivos ejecutables "jar", necesarios para manejar el sistema:

- **IndexUI.jar** - Se utiliza para efectuar el proceso de indexación de contenidos a partir de una colección de documentos de prueba.

- **RetUI.jar** - Empleado para llevar a cabo pruebas de recuperación de información sobre los contenidos indexados, permitiendo determinar la eficacia y eficiancia del proceso de indexación.

Colección de prueba

Para llevar a cabo la práctica es necesario probar una colección de prueba con "Lemur" y responder a las preguntas definidas en la plantilla correspondiente. La colección que se utilizará ha sido desarrollada específicamente para probar la capacidad de reducción, eliminación de palabras vacías y recuperación en el entorno de las noticias de actualidad de los medios de comunicación españoles, de los que se han extraido cerca de 24.000 noticias, transcritas en documentos "txt".

- Descargar colección de prueba: http://www.mblazquez.es/blog-ccdoc-recuperacion/documentos/news-24000-es_blazquez.zip

Indexación

1. Crear proyecto de indexación “prueba1” a partir de los documentos de la colección de prueba. El formato de los documentos es “txt”. No se añadirán palabras vacías. Aplicar Stemming con el algoritmo de Krovetz. El límite de memoria será de 1024MB.

Tamaño del directorio prueba1	
Tamaño del fichero inverso	

2. Crear proyecto de indexación “prueba2” a partir de los documentos de la colección de prueba. El formato de los documentos es “txt”. No se añadirán palabras vacías. Aplicar Stemming con el algoritmo de Porter. El límite de memoria será de 1024MB.

Tamaño del directorio prueba2	
Tamaño del fichero inverso	

Recuperación

3. Reseña los resultados obtenidos para las siguientes consultas en cada proyecto de indexación.

Consulta	Prueba1 (10 docs más relevantes)	Prueba2 (10 docs más relevantes)
reestructuración del sector inmobiliario		
política fiscal		
Energía nuclear		

4. ¿Se obtienen los mismos resultados con Prueba1 y Prueba2? ¿A qué puede ser debido?

5. En las consultas anteriores, ¿estaban presentes todos los términos de la consulta en el documento?

6. En las consultas anteriores, ¿estaban presentes todos los términos de la consulta en el documento?

7. Al mostrar la columna “Score” ¿qué tipo de valor se muestra?

8. Probar en el proyecto “prueba1” la consulta “rne” con 25000 número de documentos y determinar cuál es la frecuencia de aparición del primer resultado y la del último.

Práctica5. Calculando pesos TF-IDF

Con el objetivo de asimilar los conocimientos referidos a la ponderación de términos, mediante el cálculo de pesos TF-IDF, se propone la siguiente práctica. Se tendrán que obtener los pesos referidos a una serie de términos y finalmente ordenar los resultados para conocer cuáles son los términos con mayor poder discriminatorio y representativo.

- Calculadora de pesos TF-IDF
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/ponderacion01.php

Término	N	DF	TF doc1	Cálculo por defecto con correctivo y aplicando logaritmo en base 10		Sin correctivo y aplicando logaritmo en base 2	
				IDF	TF-IDF	IDF	TF-IDF
Universalizar	14520	150	1				
Legado		29	2				
Albert		70	5				
Einstein		70	5				
Universidad		2586	4				
Investigador		3658	3				
Isaac		50	1				
Newton		50	1				
Explotación		4585	8				
Referencias		3215	15				
Tecnológico		1987	12				
Propiedad		6885	3				
Intelectual		7845	3				
Nobel		180	2				
Física		789	7				

1. ¿Cuál es el término con mayor y menor factor IDF? Indica cuáles, así como sus correspondientes factores.

2. ¿Cuál es el término con mayor y menor peso TF-IDF? Indica cuáles, así como sus pesos. ¿Se cumple la propiedad de que su poder discriminatorio es inversamente proporcional? Justifica tu respuesta.

3. ¿Qué diferencias se encuentran a la hora de utilizar el cálculo por defecto y el cálculo sin correctivo? Explica los motivos y razones.

Práctica6. Probando el modelo booleano

A partir de los conocimientos adquiridos sobre el modelo booleano, se propone poner a prueba la resolución de una consulta booleana compleja en modo manual a partir de un fichero diccionario. Por otro lado se proporciona acceso a un simulador de búsquedas booleanas a través del que se resolverán diversas consultas que tienen como objetivo poner de manifiesto la operativa del modelo. El corpus del sistema lo conforman 20.000 registros relativos a noticias de medios de comunicación de prensa, radio y televisión.

- Simulador de consulta booleana
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/modelobooleano.php

1. A partir del siguiente fichero diccionario crear una tabla matriz binaria y resolver las siguiente consulta booleana:
   a. Q = ((Biblioteca AND España) OR (Electricidad AND Ciencia)) XOR (Geografía OR Población)

ID término	Término	ID documentos
T1	Clima	{ 1, 4, 6, 7, 8, 9, 10 }
T2	Biblioteca	{ 2, 4, 5, 6, 10 }
T3	Universidad	{ 1, 2, 3, 7, 8, 9 }
T4	Alcalá	{ 2, 5, 9, 10 }
T5	España	{ 1, 2, 3, 9, 10 }
T6	Libros	{ 4, 8, 10 }
T7	Geografía	{ 2, 3, 5, 7, 10 }
T8	Población	{ 3, 4, 6, 8, 9 }
T9	Electricidad	{ 1, 3, 5, 7, 9 }
T10	Ciencia	{ 1, 5, 7, 8, 10 }
T11	Social	{ 2, 3, 4, 8, 9 }
T12	Luz	{ 5, 6, 7, 10 }
T13	Unamuno	{ 4, 8, 10 }
T14	Física	{ 1, 3, 5, 7, 8 }
T15	Fluidos	{ 1, 2, 7, 9 }

Diccionario	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10

Ejecución de la consulta

Resultado

2. Utilizar el simulador de búsqueda booleana disponible en: http://mblazquez.es/blog-ccdoc-recuperacion/programas/modelobooleano.php y realizar la siguiente tabla de consultas, reseñando y analizando los resultados que se solicitan.

ID consulta	Consulta	Nº de resultados obtenidos	Tiempo de ejecución	Nº de docs. Relevantes	Nº de docs. Irrelevantes
1	Política AND Nacional			[no]	[no]
2	Política OR Nacional			[no]	[no]
3	Política XOR Nacional			[no]	[no]
4	Crisis AND Economía				
5	Biblioteca AND Deporte				
6	Elecciones AND Europa				
7	Museo AND Biblioteca				
8	Europa AND Laboral				

3. Las consultas 1, 2 y 3 demuestran las propiedades de intersección, unión y complemento. Demuéstralo aplicando la lógica matemática con los resultados obtenidos.

4. Los resultados de la consulta 4 contienen en su texto todos los términos de búsqueda. En caso negativo ¿a qué puede ser debido? ¿qué resultados no tienen que ver con la crisis económica? Justifica tus respuestas.

5. ¿Los resultados de la consulta 5 son relevantes? ¿Por qué sí o no? ¿Qué tipo de resultados se muestran, a qué corresponden?

6. ¿Todos los resultados de la consulta 6 tienen que ver con las elecciones europeas? ¿Por qué no se recuperan exactamente las noticias que versen sobre las elecciones precisadas? ¿Qué tipo de operador solucionaría este problema? ¿existe dicho operador en este simulador? ¿existe quizás en Google y otros buscadores?

7. ¿Cuántos resultados de la consulta 7 versan sobre bibliotecas de museo?

8. ¿Cuál es el motivo de los resultados irrelevantes de la consulta 8?

9. Elabora 3 consultas de prueba utilizando 4 términos y conectivas distintas.

ID consulta	Consulta	Nº de resultados obtenidos	Tiempo de ejecución	Nº de docs. Relevantes	Nº de docs. Irrelevantes
1					
2					
3					

Práctica7. Prueba manual del modelo vectorial

A partir de las formulaciones del modelo vectorial, se propone poner a pruena la resolución de una consulta vectorial en modo manual a partir de un fichero diccionario. Se obtendrán los resultados del producto escalar, similaridad del coseno, similaridad de Dice y Jaccard para finalmente comparar y discutir los resultados obtenidos.

1. A partir del siguiente fichero diccionario calcular los coeficientes de similaridad, preparando la consulta del usuario:

Cadena de consulta original del usuario						
La obra de Unamuno y sus recursos literarios						
Depuración de la consulta del usuario. Incluir al lado de cada palabra los siguientes pesos entre paréntesis (3,234) (4,678) (1,123) (2,345).						
Fichero diccionario		**Documento1**	**Documento2**	**Documento3**	**Documento4**	**q = pesos de la consulta del usuario**
Id	**Término**	**Peso TF-IDF**	**Peso TF-IDF**	**Peso TF-IDF**	**Peso TF-IDF**	
T1	Literarios	1,452	0	3,111	3,544	
T2	Biblioteca	0	2,093	0	1,234	
T3	Universidad	2,122	0	2,345	0	
T4	Alcalá	3,564	0	1,234	0	
T5	España	4,123	4,245	0	0	
T6	Libros	0	1,234	2,234	0	
T7	Sinonimia	0	0	5,332	0	
T8	Obra	2,342	0	0	5,322	
T9	Hipérbole	0	0	4,234	0	
T10	Ciencia	0	0	1,123	0	
T11	Social	0	2,345	2,234	0	
T12	Luz	1,975	0	0	0	
T13	Unamuno	4,543	2,135	1,111	2,345	
T14	Metáfora	0	0	3,234	0	
T15	Fluidos	6,134	0	0	0	
T16	Literatura	2,234	3,456	4,324	0	
T17	Semántico	0	1,212	3,232	0	
T18	Recursos	0	1,234	1,234	1,234	

2. Calcula la similaridad de los documentos 1, 2, 3 y 4, con respecto a la consulta Q, según su producto escalar binario.

Consulta	**Cálculo del producto escalar binario**	**Resultado**
Documento1		
Documento2		
Documento3		

Documento4		

3. Calcula la similaridad de los documentos 1, 2, 3 y 4, con respecto a la consulta Q, según su producto escalar TF-IDF.

Consulta	Cálculo del producto escalar TF-IDF	Resultado
Documento1		
Documento2		
Documento3		
Documento4		

4. Calcula la similaridad del coseno en los documentos 1, 2, 3 y 4, con respecto a la consulta Q.

Consulta	Cálculo de similaridad del coseno	Resultado
Documento1		
Documento2		
Documento3		
Documento4		

5. Calcula la similaridad de Dice en los documentos 1, 2, 3 y 4, con respecto a la consulta Q.

Consulta	Cálculo de similaridad de Dice	Resultado
Documento1		
Documento2		
Documento3		
Documento4		

6. Calcula la similaridad de Jaccard en los documentos 1, 2, 3 y 4, con respecto a la consulta Q.

Consulta	Cálculo de similaridad de Jaccard	Resultado
Documento1		
Documento2		
Documento3		
Documento4		

7. A tenor de todos los resultados obtenidos cuál crees que es el método de similaridad más preciso para la consulta formulada por el usuario. Razona tu respuesta.

Práctica8. Prueba automática del modelo vectorial

Efectuada la prueba manual del modelo vectorial, se pueden advertir los detalles que operan durante el cálculo de los coeficientes de similaridad para obtener un rango de documentos ordenados según su relevancia con respecto a una consulta dada. Esta visión se corresponde con un enfoque teórico matemático perfectamente delimitado. No obstante el modelo de similaridad del coseno aplicado a una colección de prueba real, puede ofrecer resultados que se salen fuera de los parámetros habituales de ejecución. Para ello se pondran a prueba diversas consultas planteadas por un hipotético usuario, utilizando el simulador de consulta vectorial, desarrollado ex-profeso para esta práctica. La colección de prueba utilizada es la misma que la utilizada en el simulador de consulta booleana.

- Simulador de consulta vectorial
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/modelovectorial.php

1. Utilizar el simulador de consulta vectorial disponible en: http://mblazquez.es/blog-ccdoc-recuperacion/programas/modelovectorial.php y realizar la siguiente tabla de consultas, reseñando y analizando los resultados que se solicitan.

ID	Consulta	Nº resulta-dos	Tiempo de Ejecu-ción	Nº de docs. Pertinentes	Nº de docs. Irrelevantes	Título del documento con mayor similaridad, coeficiente	Título del documento con menor similaridad, coeficiente
1	– Biblioteca(1) – Nacional(1) – Literatura(1)						
2	– Biblioteca(11) – Nacional(11) – Literatura(-11)						
3	– Biblioteca(1) – Nacional(1) – Literatura(-11)						
4	– Crisis(3) – Económica(3) – Mundial(3)						
5	– Crisis(3) – Económica(10) – Mundial(3)						

6	– Partido(1) – Político(1)						
7	– Partido(10) – Político(10)						
8	– Economía(10) – China(2) – Asia(2)						
9	– Economía(2) – China(2) – Asia(2)						
10	– Tecnología(10) – Energía(2) – Solar(2)						
11	– Tecnología(3) – Energía(3) – Solar(3)						
12	– Universidad(3) – Complutense(3) – Madrid(3)						

2. Se recuperan a partes iguales los contenidos de literatura y bibliotecas en la consulta 1 que en la consulta 2 y 3. Describe los efectos que producen los pesos.

3. ¿Qué efecto produce la ponderación negativa? Investiga si es posible este tipo de ponderación. Utiliza todos los recursos a tu alcance para ofrecer una explicación plausible de ello. Documenta con fuentes tu respuesta.

4. ¿El resultado con mayor coeficiente de similaridad de la consulta 4, responde a la pregunta del usuario? ¿Cómo perfeccionar la consulta para obtener resultados más pertinentes? Explica todos los pasos dados y efectúa una prueba que lo demuestre.

5. ¿Qué cambios se producen en la consulta 5 con respecto a la anterior? Explica la influencia por pesos de cada término.

6. Las consultas 6 y 7 plantean una búsqueda sobre partidos políticos, siendo tan general la consulta, ¿se encuentran resultados pertinentes en ambos casos?

7. ¿Cómo evitar que aparezcan resultados relativos a partidos de fútbol o deportes en las consultas 6 y 7?

8. Si el objetivo de la consulta 8 y 9 es encontrar noticias relativas a la economía de China y Asia en general, porqué se obtienen mejores resultados cuando la ponderación del término economía se iguala al resto de términos. ¿Qué efecto se conseguiría elevando la ponderación uniformemente en todos los términos (probar distintos valores 3, 5, 8, 10)?

9. Si el objetivo de la consulta 10 y 11 es obtener noticias sobre tecnología solar, ¿por qué al ponderar el término “tecnología” más que los demás se obtienen peores resultados?

10. En la consulta 12 se pretende recuperar noticias relativas a la Universidad Complutense de Madrid. ¿Cómo mejorar el posicionamiento de las noticias relevantes, qué pesos se deben establecer? Efectúa varias pruebas hasta encontrar la solución.

11. Elabora 4 consultas de prueba. En la primera utilizar 1 término, en la segunda 2 términos, en la tercera 3 términos y en la cuarta 4 términos. Cada una de ellas deberá contener los términos de la anterior con el objetivo de precisar un contenido determinado en la colección de prueba. Explicar cuál es el objeto de recuperación de las consultas y porqué se han asignado los pesos utilizados.

ID consulta	Consulta	Nº de resultados obtenidos	Tiempo de ejecución	Nº de docs. Pertinentes	Nº de docs. Irrelevantes
1					
2					
3					
4					

Práctica9. Prueba manual del modelo probabilístico

A partir de las formulaciones del modelo probabilístico, se propone poner a pruena la resolución de una consulta con el modelo probabilístico en modo manual a partir de un fichero diccionario. Se obtendrán los resultados de similaridad para cada documento propuesto para finalmente comparar y discutir los resultados obtenidos.

1. A partir del siguiente fichero diccionario calcular el coeficiente de similaridad

Cadena de consulta original del usuario							
La obra de Unamuno y sus recursos literarios							
Depuración de la consulta del usuario. N = 4000							
Obra Unamuno Recursos Literarios							
Fichero diccionario		**Documento1**	**Documento2**	**Documento3**	**Documento4**	**q = pesos de la consulta del usuario**	**Freq. Aparición de del término de la consulta en los documentos**
Id	**Término**	**Peso TF-IDF**	**Peso TF-IDF**	**Peso TF-IDF**	**Peso TF-IDF**		
T1	Literarios	1,452	0	3,111	3,544		158
T2	Biblioteca	0	2,093	0	1,234		
T3	Universidad	2,122	0	2,345	0		
T4	Alcalá	3,564	0	1,234	0		
T5	España	4,123	4,245	0	0		
T6	Libros	0	1,234	2,234	0		
T7	Sinonimia	0	0	5,332	0		
T8	Obra	2,342	0	0	5,322		589
T9	Hipérbole	0	0	4,234	0		
T10	Ciencia	0	0	1,123	0		
T11	Social	0	2,345	2,234	0		
T12	Luz	1,975	0	0	0		
T13	Unamuno	4,543	2,135	1,111	2,345		256
T14	Metáfora	0	0	3,234	0		
T15	Fluidos	6,134	0	0	0		
T16	Literatura	2,234	3,456	4,324	0		
T17	Semántico	0	1,212	3,232	0		
T18	Recursos	0	1,234	1,234	1,234		1800

2. Calcula la similaridad de los documentos 1, 2, 3 y 4, con respecto a la consulta Q, según su producto escalar binario.

Consulta	**Cálculo del producto escalar binario**	**Resultado**
Documento1	Pesos de los términos de la consulta:	
	Similaridad:	
Documento2	Pesos de los términos de la consulta:	
	Similaridad:	

Documento3	Pesos de los términos de la consulta:	
	Similaridad:	
Documento4	Pesos de los términos de la consulta:	
	Similaridad:	

3. Calcula la similaridad de los documentos 1, 2, 3 y 4, con respecto a la consulta Q, según su producto escalar TF-IDF.

Consulta	Cálculo del producto escalar TF-IDF	Resultado
Documento1	Pesos de los términos de la consulta:	
	Similaridad:	
Documento2	Pesos de los términos de la consulta:	
	Similaridad:	
Documento3	Pesos de los términos de la consulta:	
	Similaridad:	
Documento4	Pesos de los términos de la consulta:	
	Similaridad:	

Práctica10. Prueba automática del modelo probabilístico

Efectuada la prueba manual del modelo probabilístico, se pueden advertir los detalles que operan durante el cálculo de los coeficientes de similaridad, obteniendo en esencia los pesos ajustados de los términos de la consulta del usuario. Este proceso también se emplea en muchos sistemas de recuperación, completamente automatizados, para tener un punto de vista diferente, se propone el desarrollo de la práctica con un simulador desarrollado ex-profeso par testar el comportamiento del cálculo de pesos de la consulta con una colección real. Algunos de los factores correctores de la formulación empleada, fueron modificados para mejorar los resultados, obtenidos. No obstante la filosofía del modelo sigue intacta y se podrá comprobar el mecanismo de retroalimentación por relevancia que lo caracteriza.

- Simulador de consulta probabilística
  http://mblazquez.es/blog-ccdoc-recuperacion/programas/modeloprobabilistico.php

1. Utilizar el simulador de consulta vectorial y realizar la siguiente tabla de consultas, reseñando y analizando los resultados que se solicitan.

ID	Consulta	Nº resulta-dos	Tiempo de Ejecu-ción	Nº de docs. Pertinen-tes	Nº de docs. Irrelevan-tes	Título del documento con mayor similaridad y coeficiente	Título del documento con menor similaridad y coeficiente
1	– Biblioteca – Nacional – Literatura						
2	– Crisis – Económica – Mundial						
3	– Partido – Político						
4	– Economía – China – Asia						
5	– Tecnología – Energía – Solar						
6	– Universidad – Complutense – Madrid						

2. En relación a la consulta1, ¿Qué pesos se obtuvieron para los términos de la consulta? ¿Existen documentos con la misma similaridad? ¿A qué se debe ese efecto de idéntica similaridad en los documentos? ¿Cómo lo solucionarías? ¿Muestra resultados sobre bibliotecas y literatura? ¿Resulta más efectivo este algoritmo probabilístico que el modelo vectorial o el booleano ante una consulta análoga? Pruébalo y confírmalo.

3. En la consulta2, ¿se encuentran en las primeras posiciones, resultados relevantes? Qué pesos obtuvo los términos de la consulta…

4. La consulta3, ¿consigue evitar entre las primeras posiciones los resultados relativos a los partidos deportivos? Para mejorar la consulta, marca los contenidos que consideres relevantes (no más de 4 para evitar el bloqueo del servidor, ya que disponemos de un servicio de alojamiento limitado) y comprueba si los resultados mostrados han mejorado y en qué aspectos. ¿Se modificaron los pesos de los términos de la consulta, qué valor tenían y qué valor tienen ahora? Trata de marcar nuevos contenidos relevantes, eleva su número de 4 a 8, ¿Se han visto modificados los pesos? En caso negativo, se te ocurre alguna explicación a tenor de la formulación que se observa…

Práctica11. Método de evaluación de un sistema de recuperación de información

La evaluación de sistemas de información y recuperación tiene una aplicación muy clara en el apartado de los algoritmos de recuperación y clasificación de contenidos. En este sentido, la práctica propuesta supone un caso real del proceso de evaluación de un algoritmo de recuperación y clasificación automática.

Partiendo de una serie de categorías temáticas que comprenden unas páginas web de resultados, se deberá comprobar que los contenidos recuperados bajo el paraguas de la categoría temática asignada, lo están correctamente. Esto implica determinar distintos valores:

- **Botón Marcar Relevante** - Relevancia del 100% significa que el contenido está conforme con la categoría asignada automáticamente por el sistema.

- **Botón Marcar Irrelevante** - Relevancia del 0% significa que el contenido es completamente opuesto a la categoría asignada automáticamente por el sistema.

- **Botón Marcar Grado de Relevancia** - Relevancia del 80% - 60% - 40% y 20% significa que se ha considerado un documento parcialmente relevante en la medida porcentual que se indica.

Se recomienda hacer click en un único botón. En caso de equivocación, marcar seguidamente el botón correcto. Esto deshará la operación anterior y validará como buena la última efectuada. También se advierte, que según se evalúa el contenido, automáticamente los botones cambian de color, permitiendo al evaluador distinguir fácilmente los resultados que quedan por evaluar.

El proceso de evaluación se registra automáticamente en la base de datos, de tal manera que posteriormente se pueda comprobar si el algoritmo de clasificación acertó o no en la categorización de los contenidos en todas las áreas de conocimiento o por el contrario falló en algunas, así como determinar el motivo del fallo. Esto significa que no es

necesario enviar ninguna práctica a través del campus virtual, ya que según se lleva a cabo la evaluación, ésta se va completando.

Se tiene una responsabilidad importante en la consecución de este trabajo, ya que su supervisión y evaluación servirá para determinar el grado de corrección y precisión del sistema de información, por ello se solicita la mayor concentración posible durante este proceso.

Lista de asignación

- http://mblazquez.es/testbench/evaluacion/prueba1-es/
- http://mblazquez.es/testbench/evaluacion/prueba1-mx/

11. Índice de tablas

12. Índice de figuras

13. Bibliografía y referencias

- ADAM, G.; BOURAS, C.; POULOPOULOS, V. 2009. CUTER: an Efficient Useful Text Extraction Mechanism. Disponible en: http://ru6.cti.gr/ru6/publications/3267PID838806.pdf

- BAEZA YATES, R.; RIBEIRO NETO, B. 2005. Modelling: Boolean model. En: Modern Information Retrieval. Disponible en: http://grupoweb.upf.es/WRG/mir2ed/pdf/slides_chap03.pdf

- BAEZA YATES, R.; RIBEIRO-NETO, B. 1999. Modern Information Retrieval. Addison Wesley.

- BERRY, M.W.; BROWNE, M. 2005. Understanding search engines: mathematical modeling and text retrieval. Disponible en: http://www.bookf.net/p/7539-understanding-search-engines

- BOOTH, A. D. 1967. A Law of Occurrences for Words of Low Frequency. Information and control, 10(4):386-393. Disponible en: http://www.sciencedirect.com/science/article/pii/S001999586790201X

- CROFT, W. B.; HARPER, D. J. 1979. Using probabilistic models of document retrieval without relevance information. Journal of Documentation. 35(4): pp.285-295

- CUNNINGHAM, H.; BONTCHEVA, K.; TABLAN, V. [et.al.] 2012. Gate: General Architecture for text engineering. Disponible en: http://gate.ac.uk/

- DROST. I.; INGERSOLL, G.; MARGULIES, B. [et.al.] 2010. Apache OpenNLP. Disponible en: http://incubator.apache.org/opennlp/

- FIGUEROLA, C.G.; ALONSO BERROCAL, J.L.; ZAZO RODRÍGUEZ, A.F.; RODRÍGUEZ, E. Algunas Técnicas de Clasificación Automática de Documentos. En: Cuadernos de Documentación Multimedia, (15). Disponible en: http://multidoc.rediris.es/cdm/viewarticle.php?id=28&layout=html

- GANJISAFFAR. Y. 2012. Crawler4j. Disponible en: http://code.google.com/p/crawler4j/

- GROSSMANY, D.A.; FRIEDER, O. 2004. Information Retrieval, Algorithms and Heuristic. Springer.

- JIMÉNEZ SALAZAR, H.; PINTO, D.; ROSSO, P. 2005. Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos. En: Procesamiento del Lenguaje Natural. 35: pp. 383-390. Disponible en: http://www.sepln.org/revistaSEPLN/revista/35/47.pdf

- JIMÉNEZ SALAZAR, H.; PINTO, D.; ROSSO, P. 2005. Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos.

En: Procesamiento del Lenguaje Natural. 35: pp. 383-390. Disponible en: http://www.sepln.org/revistaSEPLN/revista/35/47.pdf

- LÓPEZ, D. 2011. Information extraction in the WWW: technology and tools for problem solving = Extracción de información en la web, tecnología y herramientas para resolver la problemática. En: SISOB Observatorium for Science in Society based in Social Models. Disponible en: http://sisobproject.wordpress.com/2011/11/18/information-extraction-in-the-www-technology-and-tools-for-problem-solving-extraccion-de-informacion-en-la-web-tecnologia-y-herramientas-para-resolver-la-problematica

- LUHN, H. P. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research Development, 2(2): pp.159-165

- LUHN, H.P. 1960. Keyword-in-context index for technical literature. American Documentation, 11(4). pp. 288–295

- MANNING, C.D.; RAGHAVAN, P.; SCHÜTZE, H. 2008. Introduction to Information Retrieval. Cambridge University Press. 107-114 pp.

- MARTÍNEZ COMECHE, J.A. 2006. Los modelos clásicos de recuperación de información y su vigencia. En: Tercer Seminario Hispano-Mexicano de investigación en Bibliotecología y Documentación, UNAM, Centro Universitario de Investigaciones Bibliotecológicas. pp.187-206. Disponible en: http://eprints.rclis.org/bitstream/10760/9662/1/Modelos_RI_vers_def.pdf

- MOONEY, R.J.; NAHM, U.Y. 2005. Text Mining with Information Extraction. En: Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium. pp. 141-160. Disponible en: http://www.cs.utexas.edu/~ml/papers/discotex-melm-03.pdf

- NIKIC, V. 2010. Web Harvest. Disponible en: http://web-harvest.sourceforge.net/

- OSWALD, D. 2006. HTML Parser. Disponible en: http://htmlparser.sourceforge.net/

- PASTERNACK, J.; ROTH, D. 2009. Extracting Article Text from the Web with Maximum Subsequence Segmentation. En: WWW 2009 MADRID!, Track: XML and Web Data. Disponible en: http://www2009.eprints.org/98/1/p971.pdf

- POHL, S.; ZOBEL, J.; MOFFAT, A. 2010. Extended Boolean retrieval for systematic biomedical reviews. En: ACSC '10 Proceedings of the Thirty-Third Australasian Conferenc on Computer Science - Volume 102. Disponible en: http://dl.acm.org/citation.cfm?id=1862212

- POPESCU, A.M. 2007. Information Extraction from Unstructured Web Text. Disponible en: http://turing.cs.washington.edu/papers/popescu.pdf

- PORTER, M.F. 1980, An algorithm for suffix stripping, Program, 14(3) pp 130−137.

- PORTER, M.F. 2006. The Porter Stemming Algorithm. Disponible en: http://tartarus.org/~martin/PorterStemmer/

- PORTER, M.F.; BOULTON, R. 2010. Snowball. Disponible en: http://snowball.tartarus.org/

- RAMOS, J. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. En: The First instructional Conference on Machine Learning. Disponible en: https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf

- RIJSBERGEN, C.J. 1979. Information Retrieval. Disponible en: http://www.dcs.gla.ac.uk/Keith/Preface.html

- RIJSBERGEN, C.J.; [et.al.] 1979. Information Retrieval. Disponible en: http://www.dcs.gla.ac.uk/Keith/Chapter.2/Ch.2.html

- RIJSBERGEN, C.J.; Robertson S.E.; PORTER, M.F. 1980. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587).

- ROBERTSON, S. 2004. Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation. Vol.60: (5), 503-520 pp.

- ROBERTSON, S.E. 1977. The probability ranking principle in IR. Journal of Documentation, 33(4): pp.294-304

- ROGERS, J.D.; TANIMOTO, T.T. 1960. A Computer Program for Classifying Plants. Science. pp1115-1118. Disponible: http://www.sciencemag.org/content/132/3434/1115.full.pdf

- SALTON, G.; McGILL, M.J. 1983. Introduction to Modern Information Retrieval. New York: Mc Graw Hill.

- SALTON, G.; WONG, A.; YANG, C.S. 1975. A vector space model for automatic indexing. En: Communications of the ACM, vol. 18, nr. 11, pp. 613–620. Disponible en: http://www.cs.uiuc.edu/class/fa05/cs511/Spring05/other_papers/p613-salton.pdf

- SCHULTZ, C.K. 1968. H.P. Luhn: Pioneer of Information Science - Selected Works. Macmillan.

- SEEGER, M. 2010. Building blocks of a scalable web crawler. Department of Computer Science and Media, Stuttgart University. Disponible en:

http://blog.marc-seeger.de/assets/papers/thesis_seeger-building_blocks_of_a_scalable_webcrawler.pdf

– SHARP, M. 2001. Text Mining. En: Seminar in Information Studies, Prof. Tefko Saracevic. Disponible en: http://comminfo.rutgers.edu/~msharp/text_mining.htm

– SHI, S.; XING, F.; ZHU, M. [et.al.] 2009. Anchor Text Extraction for Academic Search. En: Proceedings of the 2009 Workshop on Text Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, pages 10-18. Disponible en: http://dl.acm.org/citation.cfm?doid=1699750.1699753

– SINGHAL, A. 2001. Modern Information Retrieval: A Brief Overview. En: Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. Disponible en: http://singhal.info/ieee2001.pdf

– SPARCK JONES, K. 1979. Search term relevance weighting given little relevante information. Journal of Documentation. 35(1): pp. 30-48

– SPARCK JONES, K.; WILLET, P. 1997. Readings in Information Retrieval, San Francisco: Morgan Kaufmann.

– URBIZAGÁSTEGUI ALVARADO, R. 1999. Las posibilidades de la ley de zipf en la indización automática. En: B3 Bibliotecología, Bibliotecas, Bibliotecólogos. Disponible en: http://b3.bibliotecologia.cl/ruben2.htm

– URBIZAGÁSTEGUI ALVARADO, R.; RESTREPO ARANGO, C. 2011. La ley de Zipf y el punto de transición de Goffman en la indización automática. En: Investigación Bibliotecológica. 25(54): pp. 71-92. Disponible en: http://www.journals.unam.mx/index.php/ibi/article/download/27482/25470

– VELASCO, I.; DÍAZ, J.; LLORÉNS, A. 1999. Algoritmo de filtrado multi-término para la obtención de relaciones jerárquicas en la construcción automática de un tesauro. En: Revista Española de Documentación Científica, 22(1): pp. 34-49 Disponible en: http://redc.revistas.csic.es/index.php/redc/article/view/333/542

– VILARES, J. 2008. El Modelo Probabilístico: Características y Modelos derivados. Disponible en: http://www.grupolys.org/docencia/ln/2008-09/tutorial_modelo_probabilistico_apuntes.pdf/tutorial_modelo_probabilistico_apuntes.pdf

– WENINGER, T.; HSU, W.H. 2010. Text Extraction from the Web via Text-to-Tag Ratio. Disponible en: http://www.cs.illinois.edu/homes/weninge1/pubs/WH_TIR08.pdf

– YANG, E.Z. 2012. HTML Purifier. Disponible en: http://htmlpurifier.org/

– ZAZO, A.F.; BERROCAL, J.L.; FIGUEROLA, C.G.; RODRÍGUEZ, E. 2004. Estudio de usuarios de Datathéke: Propuestas de mejora utilizando expansión de consultas. Disponible: http://reina.usal.es/papers/zazo2004estudio.pdf

– ZIPF, G. K. 1949. Human behaviour and the principle of least effort. Addison-Wesley.