Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 19-43 DOI: 10.2436/15.2503.02.1

Encontrar documentos a través de las palabras y de los enlaces

José L. Alonso Berrocal
Departamento de Informática y Automática
Universidad de Salamanca

Resumen

Esta ponencia se centra en la recuperación de palabras y enlaces para encontrar documentos. En ella se exponen los métodos de recuperación de información, tanto los métodos teóricos como la indización práctica, y se resumen sus resultados. También se explica ampliamente en qué consiste el modelo vectorial de recuperación de información y, finalmente, se habla de las técnicas de recuperación en la Web y su relación con el llamado *spamdexing*, que es la capacidad de conseguir ocupar las primeras posiciones de los motores de búsqueda.

Palabras clave: *spamdexing*, indización, modelo vectorial de recuperación de información, recuperación de información.

Resum: *Trobar documents per mitjà de les paraules i dels enllaços*

Aquesta ponència se centra en la recuperació de paraules i enllaços per a trobar documents. S'hi exposen els mètodes de recuperació d'informació, tant els mètodes teòrics com la indexació pràctica, i se'n resumeixen els resultats. També s'explica àmpliament en què consisteix el model vectorial de recuperació d'informació i, finalment, es parla de les tècniques de recuperació en la Web i la relació amb l'anomenat *falsejament d'índexs* (en anglès, *spamdexing*), que és la capacitat d'aconseguir ocupar les primeres posicions dels motors de cerca.

Paraules clau: falsejament d'índexs, indexació, model vectorial de recuperació d'informació, recuperació d'informació.

Abstract: *Finding documents through words and links*

This paper focuses on the retrieval of words and links to find documents. It expounds information retrieval methods, both theoretical methods and practical indexing, and sum-

marises the results. A broad explanation is also provided as to what the vector model of information retrieval is, and the paper finally addresses Web retrieval techniques and the relationship with the so-called *spamdexing*, which is the capacity to occupy leading positions in the search engines.

KEY WORDS: spamdexing, indexing, vector model of information retrieval, information retrieval.

1. INTRODUCCIÓN

En la segunda mitad del siglo XX se produce lo que se ha dado en llamar *explosión documental*: un crecimiento exponencial de la masa de documentos, de todo tipo y en todo soporte. Esto ha puesto de relieve el problema de la recuperación de información. Es decir, la necesidad de seleccionar documentos concretos que resuelvan necesidades informativas concretas. El problema se centra fundamentalmente en seleccionar en función del contenido de los documentos; otro tipo de selección (por fechas, autores, etc.) ofrece menos problemas, al tratarse de información estructurada que puede procesarse mediante tecnología convencional (Van Rijsbergen, 1979). La vía clásica de abordar dicho problema de la recuperación de información es la indización manual: el contenido de los documentos es examinado y analizado por personas expertas, y descrito por éstas utilizando los llamados lenguajes documentales: una suerte de lenguajes artificiales controlados diseñados específicamente para describir el contenido temático de los documentos (las materias de éstos). El resultado de estas descripciones documentales puede ser almacenado de forma que se faciliten búsquedas posteriores entre estas descripciones, y seleccionar así los documentos que puedan responder a unas determinadas materias. En un principio esta forma de almacenamiento eran los clásicos ficheros en papel o cartulina, ordenados por diversos criterios; y, posteriormente, las bases de datos convencionales de los ordenadores. La indización manual, sin embargo, aun cuando se almacenen y gestionen sus resultados con ordenadores, tiene serios inconvenientes. En primer lugar, es un proceso caro y costoso: debe ser llevado a cabo por personal especializado y se trata de una tarea que requiere tiempo; no se trata, pues, de una cuestión solamente de elevados costes económicos: el tiempo necesario para indizar los documentos es mayor que el que éstos tardan en producirse. Es imposible procesar ni siquiera una mínima parte de los documentos que se producen; el alto grado de obsolescencia de buena parte de la documentación actual agrava este problema. El segundo gran problema de la indización manual es el de la inconsistencia. Se ha comprobado experimentalmente que distintos indizadores describen el mismo documento de maneras distintas (a pesar de utilizar el mismo lenguaje controlado para ello) (Hooper, 1965; Stubbs *et al.*, 2000). Incluso el mismo indizador, en momentos diferentes, produ-

ce descripciones diferentes de los mismos documentos. Es difícil producir después una recuperación eficaz, partiendo de descripciones de contenidos inconsistentes: ¿qué materias se deberían buscar para satisfacer una determinada necesidad de información? Lo cual nos lleva al tercer problema: para seleccionar los documentos que resuelvan una necesidad de información, es preciso describir dicha necesidad, y hacerlo con el mismo lenguaje controlado que se utilizó para describir los documentos; si para esto fue necesario utilizar personal especializado, para formalizar las necesidades de información también será preciso. El usuario deberá recurrir a intermediarios, a ese personal especializado, para obtener resultados satisfactorios.

2. Métodos en la recuperación de información

En la actualidad, buena parte de los documentos están disponibles en formato electrónico. En ocasiones, documentos en soporte papel están también en formato electrónico, pues han sido elaborados mediante máquinas electrónicas (por ejemplo, con un procesador de texto); en otros casos, existen sola y directamente en soporte electrónico. Sea como fuere, este hecho introduce un cambio sustancial, pues, al estar el documento completo en un soporte legible por ordenador, puede ser procesado por programas informáticos y es posible plantearse una indización totalmente automática. La indización automática, sin embargo, no está exenta de problemas. El principal de ellos es que un documento contiene mucha información, pero débilmente estructurada; al menos, estructurada de una forma que no es lo suficientemente explícita como para que los programas informáticos actuales puedan entenderla. Una solución simple a este problema es lo que se ha venido conociendo como *búsquedas en texto libre*, o también como *búsquedas de subcadenas*. Esto es, la selección por parte de un programa informático de aquellos documentos que contienen tal o cual palabra. Normalmente se podrá buscar más de una palabra, y, en ese caso, se podrán indicar restricciones adicionales mediante operadores booleanos, operadores de proximidad, truncamientos, etc. Esta solución simple tiene sus inconvenientes: los más importantes son los derivados de la sinonimia y la polisemia. Dado que un mismo concepto puede expresarse con palabras distintas (sinónimos), no siempre se puede saber cuál de ellas habrá sido utilizada en cada documento; de otro lado, puesto que una misma palabra puede referirse a conceptos diferentes, podemos encontrarnos con que muchos documentos que contienen una determinada palabra en realidad tratan sobre temas que nada tienen que ver con lo que nos interesa. El uso de operadores booleanos, de proximidad, etc. puede ayudar, pero hace que las búsquedas sean difíciles de realizar por el usuario no experto, sin llegar a paliar, sin embargo, los problemas apuntados. En todo caso, las búsquedas por palabras contenidas en los documen-

tos producen un resultado en el cual todos los documentos encontrados lo son en la misma medida: no hay forma de saber qué documentos pueden ser mejores para satisfacer nuestra necesidad de información, y esto puede ser un problema cuando los documentos encontrados son muchos.

2.1. *Los modelos teóricos*

La superación o, al menos la mitigación de estos problemas, ha dado lugar a numerosos modelos teóricos; algunos de ellos no han sido aplicados nunca en la práctica. Otros, no obstante, son la base de los sistemas de recuperación más avanzados disponibles actualmente.

Un esquema de los principales modelos para la representación y búsqueda es el que se puede ver a continuación (figura 1), cuyas características desarrollamos a continuación:

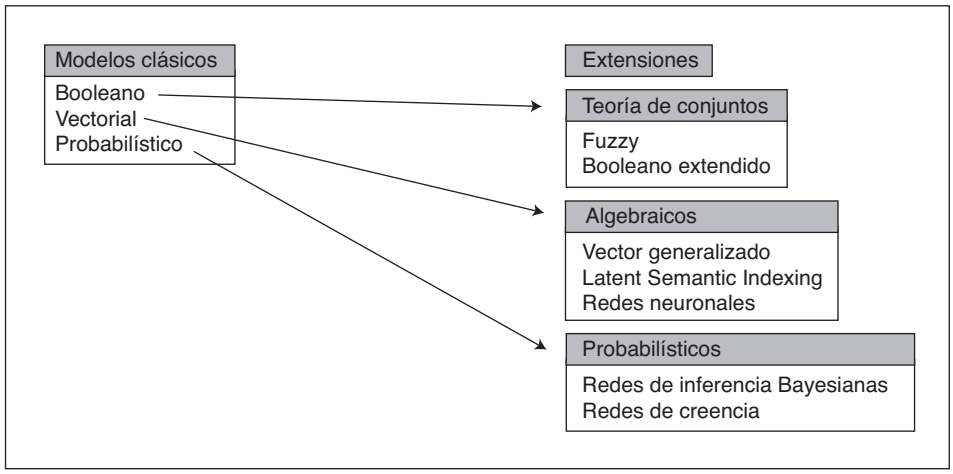


FIGURA 1. Modelos para la representación y búsqueda de palabras

a) Las características más importantes del modelo booleano son:

— Documentos.
  • Suele realizarse indización manual: a partir de la lectura y comprensión del texto, el indizador decide asignar los mejores términos que representen su contenido: descriptores.

— Consultas.
  • Las consultas se formulan utilizando los términos índice (descriptores) y una serie de operadores (booleanos, de proximidad, selección, truncamiento, etc.) y facilidades (índices, tesauros, etc.).

— El sistema de recuperación es sencillo, todo el esfuerzo recae en el usuario a la hora de plantear la consulta.
— Típico en bibliotecas, OPAC, etc.

b) En el modelo vectorial y probabilístico las características esenciales son:
— Documentos.
• Se lleva a cabo una indización automática: proceso complejo que trata de asignar automáticamente los mejores términos índice a los documentos (selección y extracción de términos).
— Consultas.
• Las consultas se realizan en lenguaje natural.
• El mismo proceso de indización automática se aplica a la consulta para obtener los términos índice que la representan.
— El sistema de recuperación es complejo. Todo el esfuerzo recae en él.
— Típico en motores de búsqueda de Internet (los mejores motores añaden información de enlaces, ej. Google).

2.2. *Indización manual vs. indización automática*

En el proceso de indización lo que pretendemos es obtener un conjunto de términos o procedimientos sintácticos (frases nominales) y convencionales para representar el contenido de un documento, con el fin de permitir su recuperación. Para ello nos basamos en el concepto de *término índice*: palabra o conjunto de palabras que tiene significado propio y que se utiliza para representar un concepto y en la idea de que tanto los documentos como la necesidad informativa pueden representarse utilizando términos índice.

Podríamos decir que *la indización* es el proceso de análisis que obtiene la representación de un documento / necesidad informativa utilizando términos índice.

Las características tipológicas de indización són las siguientes:

a) En el caso de la indización manual las características más importantes serían:
— Indización: conjunto de términos o procedimientos sintácticos (frases nominales) y convencionales para representar el contenido de un documento, con el fin de permitir su recuperación.
• Muy costosa en tiempo: muy lenta, mucho más que la producción de documentos.
• Muy costosa en dinero.

— Problemas de inconsistencia inevitable entre indizadores (sinonimia, polisemia, etc.), se requieren índices de concordancia y control de autoridades.
  • Dos personas pueden asignar diferentes palabras al mismo concepto, y la misma palabra puede aparecer en documentos que traten temas diferentes:

vendo coche usado vs. *automóvil de segunda mano*

b) En el caso de la indización automática algunas de sus características serían:
— Proceso complejo que asigna automáticamente los mejores términos índice a los documentos.
— Se persigue que las consultas puedan realizarse en lenguaje natural (texto libre).
— Problemas:
  • Información pobremente estructurada.
  • Formatos de documentos.
  • Codificación de la información.
  • Problemas de detección y conversión.
  • Normalización de términos (mayúsculas/minúsculas, acentos…).

Los pasos fundamentales que es necesario dar serían los siguientes:

1) Análisis del texto para determinar el tratamiento que se realizará sobre números, guiones, signos de puntuación, tratamiento de mayúsculas y/o minúsculas, etcétera.

2) Eliminación de palabras vacías, muy frecuentes y muy poco frecuentes. Se reduce el número de términos con valores muy pocos significativos para la recuperación.

3) Aplicación de lematización sobre los términos resultantes para eliminar variaciones morfosintácticas y obtener lemas.

4) Selección de términos que serán considerados términos índice (sustantivos, nombres propios).

5) Utilización de tesauros. Puede ayudar tanto en el proceso de indización como en el de búsqueda de información (expansión de consultas).

Analicemos brevemente cada uno de estos pasos.

2.2.1. *Análisis del texto (tokenización)*

Los elementos a tener en cuenta en esta fase son:

— Separación de palabras y «localización».
  • Carácter espacio, punto, comas, etc.
— Caracteres de puntuación.
  • A veces forman parte de términos (TVE-1, sub'21, Canal+, *e-mail*).
— Tratamiento de acentos.
  • Importante en otras fases del proceso léxico.
— Tratamiento de números.
— Detección de sintagmas y grupos nominales.
  • Nombres propios y expresiones multipalabra.
— Almacenamiento en mayúsculas/minúsculas.

2.2.2. Palabras vacías, muy frecuentes y muy poco frecuentes *(stop word)*

Se pretende reducir el ruido que pueda introducir la indización de todos los términos de un documento, y esta reducción se consigue suprimiendo:

— Palabras vacías:
  • Poseen muy poca capacidad semántica.
— Palabras muy frecuentes:
  • Si un término aparece en casi todos los documentos no sirve para diferenciar unos de otros.
— Palabras muy poco frecuentes.
  • Suelen ser errores de tecleado o palabras muy específicas (la probabilidad de que un usuario las solicite es muy baja).

2.2.3. Proceso de lematización *(stemming)*

En el proceso de lematización se tienen en cuenta los aspectos siguientes:

— «En un diccionario o repertorio léxico, elegir convencionalmente una forma para remitir a ellas todas las que derivan de su misma familia por razones de economía» (DRAE, 22.ª ed.).
— Palabras que son variaciones morfológicas con un significado prácticamente idéntico.
— Tratamiento:
  • Simple: eliminación de plurales (*s-stemmer*) o sufijos.
  • Complejo: sofisticadas técnicas de análisis procedente del PLN.
— Se basan en:
  • Aplicación de reglas.
  • Autómatas finitos.

2.2.4. Selección de términos índice

Con el objetivo de reducir la carga computacional, se intentan seleccionar los mejores términos índice.

Posibilidades de la selección de términos índice:

a) Valor de discriminación: capacidad de un término para discriminar unos documentos de otros. Tiene un coste computacional muy elevado. Además está relacionado con la frecuencia de aparición del término en toda la colección de documentos.

b) Naturaleza morfosintáctica del término: las palabras que actúan como nombres tienen mayor contenido semántico. Se pueden emplear técnicas del PLN para esta tarea, pero su coste computacional es muy elevado en comparación con sus beneficios.

2.2.5. Aplicación de tesauros

Un tesauro es un diccionario de términos controlados que contiene relaciones entre términos.

Los usos en recuperación de información (RI) son los siguientes:

— Indización (generalmente manual):
  • Los tesauros proporcionan un vocabulario controlado para la normalización de conceptos.
— Consultas:
  • Los tesauros se utilizan para plasmar con mayor exactitud la necesidad informativa del usuario, o bien, para reducir o ampliar los resultados del sistema en función de la jerarquía de términos presentes en el tesauro.
  • Expansión de consultas: trata de plantear una nueva consulta añadiendo nuevos términos relacionados con los de la consulta original (es necesario realizar un recálculo de pesos).

3. La aproximación lingüística[1]

Durante la década de los noventa, la disciplina conocida como procesamiento del lenguaje natural (PLN) experimentó un fuerte impulso que posibilitó el desarrollo de técnicas de análisis robustas, es decir, aplicables a textos sin restricciones de dominio, lo que, a su vez, permitió ampliar sus campos de aplicación. Uno de los destacados es el de la recuperación de información (RI).

Desde el campo del PLN no tardó en observarse cómo el método de indexación comúnmente adoptado en RI era resultado de un análisis muy superficial del

1. Figuerola *et al.*, 2006.

texto, y que éste podía perfeccionarse empleando las nuevas herramientas de análisis desarrolladas, para solucionar o, cuando menos, atemperar los efectos que más se denunciaban en RI —y que aún padecemos hoy en día en nuestra búsqueda cotidiana en Internet como determinantes a la hora de aumentar la efectividad en los sistemas de recuperación de información: los derivados de la ambigüedad léxica, tanto en el ámbito de la categoría gramatical como en el de significado. Como se explicó en el apartado anterior, la representación de documentos y preguntas consistía —y consiste, aún hoy día, en la mayoría de los sistemas en uso— en la detección de las «palabras ortográficas» (al menos para las lenguas con nuestros convenios ortográficos) de los textos, la normalización de las mismas a su forma mayúscula y minúscula (con eliminación de acentos y diacríticos) y la supresión de las que están incluidas en lo que se conoce como *listas de parada* o *listas de palabras vacías*.

Independientemente del método de «pesado» adoptado y de la «función o métrica de comparación» de preguntas y documentos que cada sistema implemente —que determinará, como se ha dicho también en el apartado anterior, los documentos a recuperar y el orden en que se devuelven al usuario—, el conjunto inicial de documentos candidatos susceptibles de ser recuperados será seleccionado entre aquellos que contengan, dependiendo del sistema de recuperación, todas las mismas palabras de la consulta (caso, por ejemplo, de Google), o al menos una parte de las mismas palabras de dicha consulta (caso de los sistemas basados en el modelo vectorial).

Repasamos a continuación los diferentes experimentos que se han planteado sobre colecciones monolingües y que, siguiendo a Tzoukerman *et al.* (1997), pueden dividirse en propuestas en indexación morfológica, indexación sintáctica e indexación basada en el sentido de las palabras.

3.1. *Indización morfológica*

En RI se han propuesto y experimentado técnicas no lingüísticas para intentar indexar las palabras de los documentos y de las preguntas por su raíz (técnicas de *stemming*). Estos métodos no lingüísticos, sencillos y eficientes computacionalmente, simplemente realizan una poda indiscriminada de, normalmente, determinados fines de palabra.

Se han propuesto métodos que van desde un simple *s-stemmer*, es decir, aquél que, para el inglés, elimina de toda palabra el carácter final *s* (con lo que se busca que los plurales y singulares de las palabras de documentos y preguntas se indexen por un mismo patrón), hasta otros más sofisticados para intentar tratar la morfología derivativa. Obviamente estas eliminaciones ciegas de ciertos sufijos producen anomalías en el intento de obtención de la raíz tanto por exceso como por defecto. Una versión del conocido algoritmo de Porter normaliza a la forma *organ* las palabras *organization*, *oganism* y *organ* (Krovetz, 1993). Una versión de un *s-stem-*

mer para el español que elimina los sufijos *as*, *es*, *os*, *a*, *e* y *o* de todas las palabras tiene, por ejemplo, como efecto transformar tanto *capa*, *capo* (y versiones plurales) y *cape* en *cap* (Figuerola *et al.*, 2002).

Como quiera, además, que una misma palabra puede tener, para diferentes categorías gramaticales, también la misma forma canónica (por ejemplo, *bajo* es la misma forma canónica cuando es adjetivo, sustantivo y preposición), se ha de buscar una forma de representación, en el momento de la indexación, diferenciada (*bajo*/A, *bajo*/P, *bajo*/S). De este ejemplo que hemos puesto puede colegirse fácilmente que el efecto de la desambiguación categorial puede ser beneficioso, pues con el par canónica/categoría gramatical se discriminan diferentes usos (acepciones) de la cadena de caracteres *bajo*. Otros efectos positivos que pueden obtenerse de utilizar técnicas de *POS-Tagging* en la indexación son: una eliminación coherente de las palabras vacías (por ejemplo, desechar *bajo* como preposición como palabra de indexación) y una posibilidad de reducción del tamaño de los índices (Chowdhury y McCabe, 1998).

En cuanto a los resultados obtenidos en los distintos experimentos de indexación morfológica en el momento de la recuperación, lógicamente han sido dependientes del lenguaje de la colección documental, pues los diferentes fenómenos morfológicos (flexión, derivación y composición) no se manifiestan con la misma intensidad en todas las lenguas (el inglés, p. e., es un idioma muy pobre a nivel flexivo en comparación con el español; el alemán, por otro lado, es un idioma muy aglutinativo). Así, por ejemplo, para el inglés, la conclusión obtenida es que la indexación con técnicas lingüísticas no aporta mejoras respecto de los métodos no lingüísticos, con lo que no resulta aconsejable el uso de las primeras dado la diferencia en el coste computacional. Respecto del español, los resultados obtenidos en Figuerola *et al.* (2002) parecen indicar que las técnicas de *stemming* producen efectos beneficiosos frente a los métodos que no realizan ninguna normalización. Para otros idiomas, como por ejemplo el holandés y el alemán, se ha comprobado que tratar la descomposición de palabras ortográficas en las correspondientes gramaticales produce efectos beneficiosos, tanto utilizando técnicas lingüísticas (Kraaij y Pohlmann, 1998; Monz y Rijke, 2002) como no lingüísticas (McNamee y Mayfield, 2002). En cuanto a la evaluación de los efectos que pudieran derivarse de los errores en la desambiguación categorial (la precisión de los *POS-Taggers* se sitúa entre el 95 % y el 97 %, o incluso superior), según se desprende de Gonzalo *et al.* (2002), no parecen relevantes.

3.2. *Indización sintáctica*

El método de indexación por palabras aisladas implícitamente asume la independencia de éstas respecto de los textos de las que se extraen y, por tanto, obvia lo siguiente:

1) Muchos conceptos se construyen concatenando, en determinadas lenguas, varias palabras ortográficas. Ese conjunto de palabras puede tener, para determinados dominios semánticos, una gran relevancia y, sin embargo, aisladamente, ese conjunto de palabras, por ser muy utilizadas en la colección documental, adquirir un peso irrelevante. Además, el orden de las palabras en la frase implica una variación del significado (*college junior*, vs. *junior in college* vs. *junior college*).

2) Por otra parte, determinados conceptos pueden expresarse con diferentes construcciones sintácticas que sería conveniente, a la hora de indexar, buscar una representación común (*Poland is attacked by Germany* vs. *Germany attacks Poland*).

Las conclusiones obtenidas por los grupos de investigación que más han experimentado en la indexación de sintagmas (grupo Xerox, grupo Clarit y Strazalkowski *et al.*, fundamentalmente) con técnicas lingüísticas pueden resumirse en las siguientes: en la indexación por sintagmas aunque se obtienen mejores resultados utilizando técnicas lingüísticas que meramente estadísticas, las diferencias son escasas; las mejoras entre una indexación por sintagmas con técnicas lingüísticas y una indexación por simples palabras ortográficas son inapreciables si las preguntas son cortas, aunque si las preguntas son largas sí se aprecian; la indexación por sintagmas no debe suplir a la indexación de los elementos simples que los componen; no es fácil determinar qué peso dar a los compuestos detectados.

3.3. *Indización basada en el sentido de las palabras*

Se han propuesto varios métodos para indexar documentos y preguntas de acuerdo al significado de las palabras que los componen, con el objetivo de medir los efectos que pudieran producirse al resolver los problemas de la ambigüedad léxica semántica. Para ello, se han utilizado diferentes recursos, principalmente los diccionarios y la red semántica de palabras WordNet (Peñas, 2004). La indización basada en los sentidos de acuerdo a un diccionario, dada su forma de organización, permite la representación diferenciada de los diferentes significados de un mismo significante. Esto es, posibilita el tratamiento de la polisemia y la homonimia. Utilizando una red semántica como WordNet, organizada en *synsets* (conceptos), es posible el tratamiento no sólo de los fenómenos anteriores sino también el de la sinonimia, además de la meronimia, hiponimia… dado que en la base de datos también se almacenan dichas relaciones entre los *synsets*. En cuanto a los experimentos aplicados a la indexación, resumiendo, se han concentrado en dos aspectos principales (Gonzalo *et al.*, 1999):

1) Evaluar si producen mejoras y en qué medida en la recuperación de información.

2) Fijar el umbral de error en la precisión de la desambiguación a partir del cual se produce una degradación en la efectividad de la recuperación de información.

De los resultados obtenidos del primer tipo de experimentos, los primeros efectuados cronológicamente, no era posible establecer unas conclusiones, dadas las tasas de precisión de los desambiguadores utilizados. Efectivamente, no se podía determinar si era beneficiosa o no en RI la indexación por sentidos, pues no era posible establecer la degradación que producía la desambiguación incorrecta. Otros experimentos han utilizado la estrategia de la desambiguación manual, pero para ello han recurrido a textos muy breves, p. e., pies de página (Smeaton y Quigley, 1996), con lo que los resultados no pueden extrapolarse a colecciones de grandes volúmenes de texto. El problema parece aún abierto, aunque más bien se ha pospuesto hasta que la tecnología en desambiguación madure. Independientemente de estos problemas enunciados, también se ha planteado el de la «granularidad» de los sentidos tanto en diccionarios como en WordNet. Un «grano muy fino» (trabajar con muchas acepciones diferentes para una entrada léxica), puede ser, muchas veces, contraproducente en RI, dado que al indexar separamos sentidos que pueden estar semánticamente muy cercanos.

3.4. *Expansión de consultas*

Uno de los problemas más importantes en RI consiste en formular la consulta para que plasme adecuadamente la necesidad informativa del usuario. Aparte de los requerimientos del sistema para formalizar la consulta, el mayor problema consiste en determinar el conjunto de palabras que expresen semánticamente esa necesidad. El problema se agrava debido al efecto de inconsistencia en la asignación subjetiva de términos a conceptos. Figuras como la sinonimia o la polisemia (u otras menos importantes, como la homonimia, la antonimia, la hiperonimia, la hiponimia, o la anáfora) hacen que el mismo concepto pueda expresarse con palabras diferentes y una misma palabra pueda aparecer en documentos que tratan sobre temas distintos. En esta situación no es de extrañar que el usuario tenga que replantear su consulta para obtener mejores resultados. De hecho, es ésta una de las acciones más habituales de los usuarios que utilizan motores de búsqueda en Internet. Se han propuesto diversos mecanismos para construir la nueva consulta. En general, en todos ellos se realiza una ampliación de nuevos términos a la consulta inicial y un recálculo de la importancia de cada término en la nueva consulta. Esto es lo que se conoce como expansión de consultas. Se pretende ampliar el número de términos que mejor definan la necesidad informativa del usuario de acuerdo a la colección documental y al modelo de recuperación utilizado. Para realizar la expansión lo más rápido sería utilizar tesauros o diccionarios generales ya existentes. Podemos realizar una clasificación de técnicas de expansión dependiendo de si requieren o no de la presencia del usuario. Según este punto de vista se distinguen dos grandes enfoques:

a) Realimentación de consultas utilizando criterios de relevancia del usuario (*user relevance feedback*). Requiere una buena interfaz con el usuario, pero es el mecanismo que mejores resultados proporciona. También se utiliza en motores de búsqueda en Internet, con la opción «páginas similares» o «*more like this*».

b) Expansión automática de consultas. No requieren de la presencia del usuario. Se pueden dividir a su vez en dos tipos:

— Análisis local. La expansión utiliza exclusivamente información de los documentos recuperados con la consulta inicial. Destacamos, por sus buenos resultados, la denominada pseudo-realimentación de consultas (*pseudo relevance feedback*). También se utilizan técnicas de agrupamiento local (tesauros locales de términos).

— Análisis global. Utiliza información de toda la colección de documentos para expandir la consulta. Se suelen emplear mecanismos de agrupamiento global con el objetivo de crear tesauros de términos. Destacamos varias técnicas: tesauros construidos a partir de la medida simple de coocurrencias, tesauros de similitud construidos realizando la transposición de la matriz documentos-términos (Qiu y Frei, 1993), tesauros construidos a partir de la asociación de términos y frases (*phrasefinder*), y tesauros basados en información sintáctica.

3.5. *Resumen de los resultados experimentales*

a) Aplicar lematización. Mejoras de 11,46 % y 10,85 % ($\bar{p}$ y *P*@10).

b) Realimentación de consultas con relevancia del usuario. El usuario visualiza los resultados y marca los relevantes y no relevantes y se reelabora la consulta. Hay mejoras del 300,1 % y 301,2 % ($\bar{p}$ y *P*@10).

c) Pseudo-realimentación de consultas. De forma automática se consideran los primeros documentos recuperados como relevantes. Algunas consultas mejoran y otras empeoran. Considerando los 5 primeros documentos recuperados y con 40 términos de más peso tenemos mejoras del 10,73 % y 8,43 % ($\bar{p}$ y *P*@10).

d) Tesauros. Las relaciones se pueden calcular automáticamente computando relaciones de coocurrencia tanto de términos como de documentos (tesauros de asociación); o si dos documentos poseen términos comunes (tesauro de similitud).

Además:

— Podemos utilizar tesauros globales (toda la colección) o locales (sólo los documentos recuperados).

— Podemos utilizar tesauros globales (toda la colección) o locales (sólo los documentos recuperados);

— los tesauros de asociación y los de similitud obtienen resultados similares, pero los de similitud tienen un tiempo de cómputo elevado;

— la expansión es mejor cuando se consideran los mejores términos relacionados con todos los términos de la consulta original;

— el empleo de tesauros locales obtiene mejores resultados.

4. El modelo vectorial

El modelo vectorial fue definido por Salton (Salton, 1968) hace ya bastantes años, y es ampliamente usado en operaciones de RI, así como también en operaciones de categorización automática, filtrado de información, etc. En el modelo vectorial se intenta recoger la relación de cada documento D_i, de una colección de N documentos, con el conjunto de las m características de la colección. Formalmente un documento puede considerarse como un vector que expresa la relación del documento con cada una de esas características.

En el modelo vectorial:

— cada documento es representado por un vector de términos;

— las consultas, formuladas en lenguaje natural, son representadas también como un vector de términos;

— es fácil aplicar alguna función de similitud que estime la semejanza entre el vector de la consulta y el de cada uno de los documentos.

Planteemos el problema de una manera más formal:

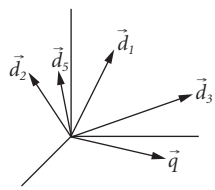
— cada documento d_j de la colección de N documentos se representa por un vector de m componentes, siendo m el número de términos índice de la colección;

— la consulta q se plantea al sistema en lenguaje natural, y también se representa por un vector;

— cada elemento del vector expresa la importancia que tiene el término índice en el documento o en la consulta: peso.

$$d_j \rightarrow \vec{d}_j = (w_{1j}, w_{2j}, \ldots., w_{mj})$$

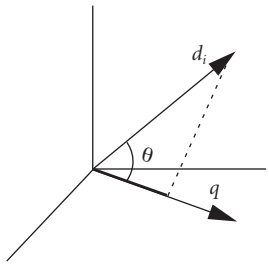
$$q \rightarrow \vec{q} = (w_{1q}, w_{2q}, \ldots., w_{mq})$$



— Para calcular la similitud entre documentos y consultas se supone que la distancia semántica entre ellos coincide con la distancia entre los vectores que las representan;

— normalmente esa distancia se mide por el coseno del ángulo que forman.

Los documentos se ordenan por orden de similitud con la consulta (ranking) y se presentan los primeros al usuario.



$$\frac{\sum_{i=1}^{m} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{m} w_{ij}^2 \cdot \sum_{i=1}^{m} w_{iq}^2}}$$

Podemos utilizar los denominados vectores binarios, para ello mostremos con un ejemplo su utilización.

Una colección de documentos en la que el total de términos distintos fuese $n = 4$.

TABLA 1. *Matriz de documento por término*

	$Term_1$	$Term_2$	$Term_3$	$Term_4$
Doc_1	0	1	1	0
Doc_2	1	0	1	0
Doc_3	1	1	0	1
Consulta	0	1	0	1

Cada vector tiene $n = 4$ elementos, uno por cada término posible. El valor de cada elemento es 0 o 1, dependiendo de si el término aparece o no en el documento. Cualquier consulta puede ser tratada de la misma forma.

Si aplicamos una función de similitud simple, como el producto entre los vectores de la consulta y de cada documento:

TABLA 2. *Matriz con función de similitud*

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	
Doc_1	0	1	1	0	simil. = 1
Doc_2	1	0	1	0	simil. = 0
Doc_3	1	1	0	1	simil. = 3
Consulta	0	1	0	1	

Obtenemos una lista de los documentos similares a la consulta, ordenados por similitud.

El que más se ajusta a la consulta es Doc_3, seguido de Doc_1.

Pero no solamente podemos utilizar el vector binario, lo más interesante es poder utilizar pesos, de esta forma:

— podemos registrar más información, no solamente la aparición de términos en documentos;

— un término puede ser más significativo en un documento que otro;

— podemos asignar a cada término un peso en cada uno de los documentos, en función de su importancia en cada documento;

— ese peso se puede estimar de diversas formas (por su frecuencia de aparición, por el lugar o campo del documento en que aparece, etc.);

— podemos representarlo mediante un valor numérico.

Un ejemplo mediante el empleo de pesos sería el siguiente:

TABLA 3. *Matriz de pesos con función de similitud*

	$Term_1$	$Term_2$	$Term_3$	$Term_4$	
Doc_1	0	0,7	0,2	0	simil. = 0,35
Doc_2	0,5	0	0,6	0	simil. = 0
Doc_3	0,6	0,4	0	0,2	simil. = 0,26
Consulta	0	0,5	0	0,3	

El documento que más se ajusta a la consulta es Doc_1.

El cálculo de los pesos puede hacerse por tres factores:

1) Si un término se repite mucho en un documento debe ser muy representativo de su contenido.

Operación: contar el número de veces que aparece un término en un documento (*tf*).

2) Si un término aparece en casi todos los documentos no sirve para diferenciar unos de otros.

Operación: contar el número de veces que aparece el término en toda la colección documental (*idf*).

3) Efectos laterales de documentos largos (muchos términos) frente a documentos cortos (pocos términos):

Operación: aplicar un factor corrector de normalización que es necesario porque:

— no todos los documentos tienen el mismo tamaño;

— conviene normalizar los pesos obtenidos con la frecuencia y el *idf*;
— el peso de un término *t* en un documento *d* se obtiene con estos tres elementos.

$$\frac{tf \times idf}{\text{normalización}}$$

Para poder trabajar con estos planteamientos se diseñaron diferentes sistemas de pesado, de forma que:
— se han propuesto diferentes formas de calcular cada uno de los tres componentes;
— cada una de esas formas se denomina o representa mediante una letra;
— las combinaciones posibles se denominan esquemas de pesado;
— ejemplo: BNN, NTC, ATU.

Para el cálculo de la frecuencia las formas son (en negrita la letra que se aplica al esquema):

none n_{tD}
binary 1

max-norm $\dfrac{n_{tD}}{\text{máx}\, n_D}$

aug-norm $0,5 + 0,5 \left(\dfrac{tf}{\text{máx}\ n_D} \right)$

square n_{tD}^2
log $ln(n_{tD}) + 1,0$

Donde:

n_{tD} n.º de veces que el término *t* aparece en el documento *D*
máx n_D n.º de veces del término que más aparece en el documento *D*

Para el cálculo del *idf* las formas son (en negrita la letra que se aplica al esquema):

none 1

tfidf $\log\left(\dfrac{N}{nd_t} \right)$

prob $\log\left(\frac{N - nd_t}{nd_t}\right)$

freq $\frac{1}{N}$

squared $\log\left(\frac{N}{nd_t}\right)^2$

Donde:

N número de documentos en la colección
nd_t número de documentos en que aparece el término t

Para el cálculo del normalizador las formas son (en negrita la letra que se aplica al esquema):

none 1

sum $\sum_{i=1}^{n} peso_{t_i D}$

cosine $\sqrt{\sum_{i=1}^{n} peso_{t_i D}^2}$

fourth $\sum_{i=1}^{n} peso_{t_i D}^4$

max máx $peso_{tD}$

Donde:

$peso_{tiD}$ peso del término i en el documento D
n número de términos en el documento D
máx $peso_{tD}$ peso del término con más peso en el documento D

Por ejemplo, si el esquema seleccionado fuera ntc-ntc (esquema en el proceso de indización y en el de consulta, que puede ser distinto), el cálculo sería:

Peso. Esquema ntc-ntc $\frac{tf \times idf}{\text{normalización}}$

— *tf* (*term frecuency*): número de veces que aparece un término en el documento/consulta.

— *idf* (*inverse document frecuency*): $\log\left(\frac{N}{nd_t}\right)$

N número de documentos en la colección
nd_t número de documentos en que aparece el término t

— normalización: se consigue haciendo que los vectores sean unitarios.

5. La recuperación en la Web

Las técnicas de recuperación de información que se han empleado en la Web, han procedido en su mayor parte de los sistemas de RI tradicionales. Por ello han surgido grandes problemas, debido a que el entorno de trabajo no es exactamente el mismo y además las características de los datos almacenados difieren considerablemente. Además han surgido nuevos problemas como el *spamming* o el enorme tamaño que deben soportar los índices, haciendo más difícil su adecuada gestión mediante el empleo de los modelos tradicionales. Las páginas web poseen una característica que las hace especiales. Prescindiendo de imágenes, sonido, elementos de captación de datos y demás ornamentos, las páginas web tienen enlaces con otras páginas. Estos enlaces son los que confieren su particular carácter a la documentación web (Alonso Berrocal *et al.*, 2003).

A partir de esos enlaces el espacio Web puede ser considerado como un grado dirigido, en el que los nodos serían las diferentes páginas existentes y los arcos, los hipervínculos que enlazan un nodo con otro (Alonso Berrocal *et al.*, 2004).

La explotación de la estructura hipertextual (Alonso Berrocal *et al.*, 1999) como método de recuperación incluye los lenguajes de consulta a la Web y la búsqueda dinámica, ideas que no están aún suficientemente implantadas. Los lenguajes de consulta a la Web pueden utilizarse para localizar todas las páginas web que tengan al menos una imagen y que sean accesibles al menos desde otras tres páginas, empleando para ello diversos modelos.

Este tipo de planteamientos se extrapola a la Web, considerado como una colección de documentos y por lo tanto se le aplican los modelos comentados. Pero le añadimos el matiz que nos suministran los enlaces, dándole un contenido semántico que podemos emplear en el modelo vectorial (Figuerola *et al.*, 2000).

Los trabajos más interesantes con enlaces están seguramente en el empleo de técnicas de posicionamiento.

5.1. *Técnicas de posicionamiento*

Las técnicas de posicionamiento, las podemos entender como el conjunto de procedimientos que permiten colocar un sitio o una página web en un lugar óptimo entre los resultados proporcionados por un motor de búsqueda. Estas técnicas han tenido y tienen un campo de trabajo y estudio muy activo y en el que se trabaja de forma constante.

Existen dos grandes variantes en los algoritmos de ranking:

— variantes del modelo vectorial o booleano

— los que siguen el principio de extensión de los enlaces.

De la primera variante hay tres métodos:

— booleano extendido

— vectorial extendido

— más citado.

De la segunda variante hay tres métodos:

— WebQuery

— HITS

— PageRank.

Algunas de las técnicas más utilizadas han sido las siguientes.

5.1.1. HITS

Este algoritmo desarrollado por Kleinberg (Kleinberg, 1999) depende de la consulta y considera el conjunto de páginas *S* que *apuntan a* o *son apuntadas por* la respuesta:

— páginas que tienen muchos links que apuntan a ellas en *S* son $\boldsymbol{A}(_P)$= llamadas autoridades (*authorities*);

— páginas que tienen muchos links de salida son llamadas conectores $h(p)$ = conectores (*hubs*).

Mejores páginas *authorities* vienen de links de entrada desde buenos conectores (*hubs*) y buenos *hubs* vienen de enlaces de salida de buenas *authorities*.

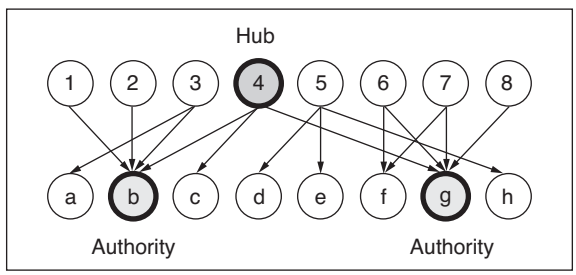


FIGURA 2. Ilustración del algoritmo HITS

5.1.2. PageRank

El PageRank (Page *et al.*, 1998) es la técnica de posicionamiento de mayor éxito y aunque se han descrito diversos problemas en el mecanismo básico de obtención del PageRank, se han planteado soluciones a los mismos (Sung Jin y Sang Ho, 2002) y constantemente se publican artículos sobre la mejora del mismo. La técnica del PageRank ha demostrado suficientemente sus características como técnica de posicionamiento en los procesos de recuperación de información (Dominich y Skrop, 2005).

El PageRank simula un usuario que navega aleatoriamente en la Web, quien salta a una página aleatoria con probabilidad q o que sigue un hyperlink aleatorio (en la página actual) con probabilidad $1 - q$.

Este proceso se modela como una cadena de Markov, en que se puede calcular la probabilidad estacionaria de estar en cada página.

La importancia de una página viene dada por la importancia de las páginas que la enlazan.

$$PR(a) = q + (1 - q) \sum_{i=1}^{n} \frac{PR(p_i)}{C(p_i)}$$

6. Webspam

Un campo de trabajo de gran actualidad son las investigaciones sobre w*eb spam*. No podemos decir con certeza que exista una única definición para *web spamming*, referido por muchos autores (Gyongyi y Garcia-Molina, 2005) como s*pamdexing*, y muchas veces definido como una práctica para conseguir una posición elevada en los resultados de los motores de búsqueda, utilizando técnicas para engañar a los algoritmos de clasificación.

El término *spam* según Castillo *et al.* (2006) ha sido utilizado en los últimos años referido a los mensajes no solicitados (normalmente comerciales).

El *spamdexing* es definido por Gyongyi y Garcia-Molina (2005) y referido por Castillo *et al.* (2006), como «cualquier acción con la intención de conseguir un aumento injustificado de la relevancia o importancia de una página web, considerando su valor real».

Cualquiera que sea la definición es cierto que el *spam* se refiere a algo indeseable, incluso perturbador, con una influencia negativa en el proceso HTTP, que al basarse en el paradigma solicitud-respuesta imposibilita el envío directo de las páginas por los *spammers* hacia los usuarios finales. Para superar esta defensa del

protocolo los *spammers* utilizan otras técnicas y medios. La más utilizada es a través de mensajes, aparentemente unidireccionales, vía *e-mail*.

Pero si nos centramos en el modo de operar de los *spammers* sobre los sistemas de recuperación de información en la Web, veremos que es diferente del resto. En este caso los principales destinatarios son los motores de búsqueda y la forma de engañar y minar las relaciones de confianza establecidas entre los usuarios de los motores de búsqueda (Gyongyi y Garcia Molina, 2005).

Estas técnicas de *spam* orientadas a los motores de búsqueda, pretenden obtener la atención de los usuarios finales, con fines normalmente comerciales. Una de las razones que subyacen están en las dificultades de los usuarios finales en distinguir las informaciones fiables de las no fiables debido al éxito de los motores en las últimas décadas (Metaxas y DeStefano, 2005).

Los usuarios han ido aumentando su confianza en los motores de búsqueda como medio de obtención de información, y los *spamers* han logrado, con éxito, conducir esa confianza a los resultados de cada consulta.

Para que sea posible continuar con la confianza en los resultados de las consultas, los constructores de motores de búsqueda deben realizar un gran esfuerzo para proporcionar respuestas sin *spam*. Realizarán sofisticadas estrategias de ranking que, junto a los algoritmos que permitan la detección del *spam,* lo eliminarán de las respuestas (Becchetti *et al.*, 2008).

De forma general algunas de las formas de realizar web *spam* se resumiría en la siguiente lista:

— *Spamdexing*
  • *keyword stuffing* (relleno)
  • *link farms* (granjas)
  • *spam blogs* (*splogs*)
  • *cloaking.*

6.1. SEO vs. *spam*

La optimización para motores de búsqueda (SEO, por sus siglas en inglés) tiene que ver con asegurarse de que un sitio sea encontrable por los buscadores. Los servicios que ofrecen los *spammers* incluyen la creación de miles o millones de páginas falsas que tienen como propósito el engañar a las máquinas de búsqueda y a sus usuarios.

En cualquier caso, la relación entre el administrador de un sitio web que intenta tener un alto posicionamiento y el administrador de la máquina de búsqueda es una relación entre adversarios en un juego de suma cero. Cada ganancia inmerecida de ranking para una página es una pérdida de precisión para la máquina de búsqueda.

Técnicas SEO legítimas (≈ técnicas de sombrero blanco):

— objetivo, aparecer en lo más alto cuando un cliente está buscándolos;

— en contraposición a una página elaborada por personas que odian a su cliente;

— más eficaz, pregunta a los sitios web legítimos para vincularse al cliente.

Spam (≈ técnicas de sombrero negro): crear lotes artificiales de los sitios web que enlazan a una página que promueve un producto (p. e. Viagra).

La separación en el «sombrero blanco» y el «sombrero negro» es una línea muy delgada.

7. Bibliografía

Alonso Berrocal; J. L.; Figuerola, C. G; Zazo, A. F. (2004). *Cibermetría*: *nuevas técnicas de estudio aplicables al Web.* Gijón: Trea.

— (1999). «Representación de páginas web a través de sus enlaces y su aplicación a la recuperación de información». *Scire*, vol. 5, n. 2, p. 91-98.

Alonso Berrocal, J. L. [*et al.*] (2003). «Agentes inteligentes: recuperación autónoma de información en la WEB». *Revista Española de Documentación Científica*, vol: 26, n. 1, p. 11-20.

Becchetti, L. [*et al.*] (2008). «Link analysis for web spam detection». *ACM Transactions on the Web,* vol. 2, n. 1, p. 1-42.

Castillo, C. (2006). A reference collection for web spam. *SIGIR Forum, vol. 40, núm. 2.*

Chowdhury, A.; McCabe, M. (1998). *Improving information retrieval using part of speech tagging* [en línia]. <*citeseer.ist.psu.edu/256084.html*> [Consulta: 29 mayo 2009].

Dominich, S.; Skrop, A. (2005). «Pagerank and interaction information retrieval». *Journal of the American Society for Information Science and Technology*, vol. 56, n. 1, p. 63-69.

Figuerola, C. G.; Alonso Berrocal, J. L.; Zazo Rodríguez, A. F. (2000). «El contenido semántico de los enlaces de las páginas web desde el punto de vista de la recuperación de información». A: Cabré, M. T.; Codina, L; Estopà, R (ed.). *Terminologia y Documentació.* I Jornada de Terminologia y Documentació (Barcelona, maig 2000). Barcelona: Institut Universitari de Lingüística Aplicada, 2000, p. 71-79.

Figuerola, C. G. [*et al.*] (2002). «Spanish monolingual track: The impact of stemming on retrieval». A: *Evaluation of Cross-Language Information Retrieval Systems.* Second Workshop of the Cross-Languge Evaluation Forum (Darmstadt, setembre 2001), Springer, vol. 2406, p. 253-261.

— (2006). «Encontrar documentos a través de las palabras». A: Fuentes, T.; Torres, J. (ed.). *Nuestras Palabras*: *Entre el Léxico y la Traducción*. Lingüística Iberoamericana, p. 147-174.

Gonzalo, J.; Peñas, A.; Verdejo, F. (1999). *Lexical ambiguity and information retrieval revisited.* 1999 Joint SIGDAT Conference on EMNLP and VLC (Maryland*, 1999),* p. 195-202.

GONZALO, J.; PEÑAS, A.; VERDEJO, F. (2000). *La indexación con técnicas lingüísticas en el modelo clásico de recuperación de información*. A: SANCHÍS, E.; MORENO, L.; GIL, I. (ed.). Primeras Jornadas de Tratamiento y Recuperación de Información (València, 4-5 juliol 2002). València: Universitat Politècnica de València. Facultat d'Informàtica, p. 97-106.

GYONGYI, Z.; GARCIA MOLINA, H. (2005). *Web spam taxonomy*. First International Workshop on Adversarial Information Retrieval on the Web.

HOOPER, R. S. (1965). *Indexer Consistency Test - Origin, Measurements, Results and Utilization.* Bethesda: MD.

KLEINBERG, J. M. (1999). «Authoritative sources in a hyperlinked environment». *Journal of the ACM*, p. 668-677.

KRAAIJ, W.; POHLMANN, R. (1998). *Comparing the effect of syntactic vs. statistical phrase index strategies for dutch*. Proceedings of ECDL'98 (setembre 1998), p. 605-617.

KROVETZ, R (1993). *Viewing morphology as an inference process*. A: KORFHAGE, R.; RASMUSSEN, E. M.; WILLET, P. (ed.). 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, 27 junio - 27 julio 1993). ACM, p. 191-203.

MCNAMEE, P.; MAYFIELD, J. (2002). «Language-Independent Approach to European Text-retrieval». A: *Cross-Language Information Retrieval Systems*. Springer, p. 29-139.

METAXAS, P. T.; DESTEFANO, J. (2005). «Web spam, propaganda and trust». AIRWeb2005, (10-14 maig).

MONZ, C.; RIJKE, M. (2002). «Shallow Morphological Analysis in Monolingual Information retrieval for Dutch, German and Italian». A: *Cross-Language Information Retrieval Systems.* Springer, p. 262-277.

PAGE, L. [*et al.*] (1998). *The PageRank citation ranking: Bringing order to the web* [informe técnico]. Stanford Digital Library Technologies Project.

PEÑAS, P. (2004) *Técnicas lingüísticas aplicadas a las búsqueda textual multilingüe: ambigüedad, variación terminológica y multilingüismo.* SEPLN.

QIU, Y.; FREI, H. P. (1993) *Concept-based query expansion.* A: KORFHAGE, R.; RASMUSSEN, E. M.; WILLET, P. (eds.). 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. (Pittsburgh, 27 junio - 27 julio 1993). ACM, p. 160-169.

RIJSBERGEN, C. J. VAN (1979). *Information Retrieval.* Glasgow: University of Glasgow. Department of Computer Science.

SALTON, G. (1968). *Automatic Information Organization and Retrieval.* Nova York: McGraw-Hill.

SMEATON, A.; QUIGLEY, I. (1996). *Experiments on using semantic distances between words in image caption retrieval*. A: FREI, H. P. [*et al.*] (ed.). 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Zuric, 18-22 agosto 1996). ACM, p. 174-180.

STUBBS, E. A.; MANGIATERRA, N. E.; MARTÍNEZ, A. (2000). «Internal quality audit of indexing: A new application of interindexer consistency». *Cataloguing & Classification Quaterly*, vol. 28, n. 4, p. 53-70.

SUNG JIN, K.; SANG HO, L. (2002). «An improved computation of the pagerank algorithm». *Lecture Notes in Computer Science*, vol. 2291. Springer, 2002, p. 73-85.

TZOUKERMAN, E.; KLAVANS, J.; JACQUEMIN, C. (1997). *Effective use of natural language processing of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsong*. Proceedings of 20th ACM/SIGIR (2 mayo 1997), p. 148-155.