Encontrar Documentos a través de las Palabras

Carlos G. Figuerola, Emilio Rodríguez Vázquez de Aldana
Ángel F. Zazo, José Luis Alonso Berrocal

Grupo REINA
Universidad de Salamanca
http://reina.usal.es

1. El problema de la Recuperación de Información

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 [rijsbergen1979information].

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, seleccionando 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, aún 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)[hooper1965indexer, stubbs2000internal]. Incluso el mismo indizador, en momentos diferentes, produce descripciones diferentes de los mismos documentos. Es difícil producir después una recuperación eficaz, partiendo de descripciones de contenidos inconsistentes: ¿cuál o cuáles materias se deberían buscar para satisfacer una determinada necesidad de información?

Lo cual nos lleva el 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.

1.2 La indización automática

Sin embargo, 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 por 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 documentos 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 nuesta necesidad de información, y esto puede ser un problema cuando los documentos encontrados son muchos.

1.3 Modelos Teóricos: el Modelo Vectorial

La superación o, al menos mitigación de estos problemas, ha dado lugar a numerosos modelos teóricos; algunos de ellos no han sido aplicados nunca en la práctica [belkin1987retrieval]. Otros, no obstante, son la base de los sistemas de recuperación más avanzados disponibles actualmente. Entre éstos, el más conocido y utilizado es el modelo del espacio vectorial o, simplemente, modelo vectorial.

Formulado inicialmente por G. Salton en los años 70 [salton1975vector], ha sido ampliamente aplicado desde entonces. Sin entrar en definiciones rigurosas, la idea básica es que un documento puede ser representado mediante un vector o lista de palabras. Ahora bien, cada una de esas palabras tiene, en ese documento, un peso, es decir un coeficiente que intenta expresar en qué medida esa palabra es representativa del contenido de ese documento [harman1992ranking]. Un usuario podría formular o expresar su necesidad de información redactando un texto en lenguaje natural; ese texto (consulta), puede ser representado de la misma manera que los documentos: un vector o lista de palabras, cada una de ellas con un peso determinado.

A partir de ahí, aplicando alguna de las funciones que permiten estimar la similitud entre dos vectores, podemos obtener on coeficiente que exprese el pareciso o similitud entre el vector consulta y cada uno de los vectores de cada uno de los documentos. Obviamente, los documentos con un coeficiente más alto responderán mejor a nuestra consulta.

1.4 La Investigación Experimental

La investigación experimental busca medir los resultados de aplicar tal o cual solución a algún problema planteado en RI; por ejemplo, aplicar un modo diferente de calcular el peso de los términos. Para ello es preciso disponer de medios para evaluar los resultados de la recuperación; esto es lo que nos permite comparar unas técnicas con otras. Los aspectos evaluables, no obstante, pueden ser diversos: rapidez, facilidad, comodidad; y también los que tienen que ver directamente con los resultados de la recuperación, es decir, con la efectividad de la recuperación. Este último aspecto es el que más nos interesa.

Para evaluar la efectividad en la recuperación es preciso disponer de algunos elementos. Uno de ellos son las llamadas colecciones experimentales, es decir, colecciones de documentos diseñadas expresamente para servir de base a la experimentación. Estas colecciones no solamente constan de documentos, sino también de baterías más o menos amplias de consultas, así como de estimaciones o juicios de relevancia:

expertos en las distintas materias han revisado los documentos y han determinado cuáles son realmente relevantes para cada consulta.

Otro elemento son las diversas medidas que permiten estimar la efectividad de la recuperación [harter1997evaluation]. De una forma u otra, todas se basan en algún tipo de comparación entre los documentos recuperados y lo que se conoce como documentos relevantes, esto es, los que sí satisfacen la necesidad de información expresada en la consulta. En este sentido, las medidas más utilizadas son la precisión y la exhaustividad, y otras derivadas de éstas.

La precisión mide la proporción de documentos que son realmente relevantes sobre los recuperados tras una consulta. Es un concepto inverso al ruido documental.

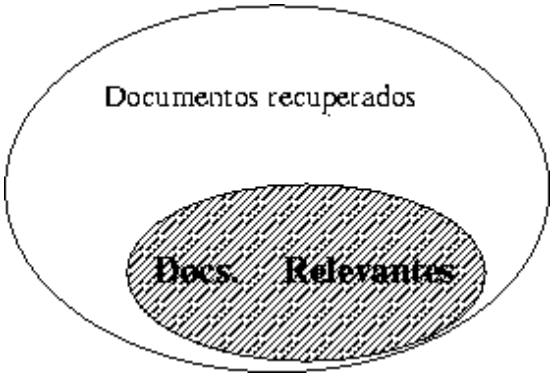

Fig. 1

La exhaustividad pretende medir la proporción de documentos relevantes que se recuperan sobre la cantidad de ellos que haya en la base de datos o colección documental. Es lo inverso al concepto de silencio documental.

Fig. 2

Estas dos medidas básicas se combinan en diversas formas, que permiten observar a un tiempo no sólo ambos aspectos, sino también otras componentes, como el hecho de que los documentos relevantes se recuperen antes o después, ya que, como se ha dicho antes, el orden en que los Sistemas de Recuperación devuelven los documentos recuperados es importante: en una búsqueda en la que obtenemos tal vez decenas o centenares de documentos, no es lo mismo que los relevantes aparezcan en los primeros lugares o que lo hagan en las últimas posiciones. En este sentido, una de las formas de evaluación más utilizada es el gráfico de precisión-exhaustividad interpolada, que pretende expresar todos estos extremos.

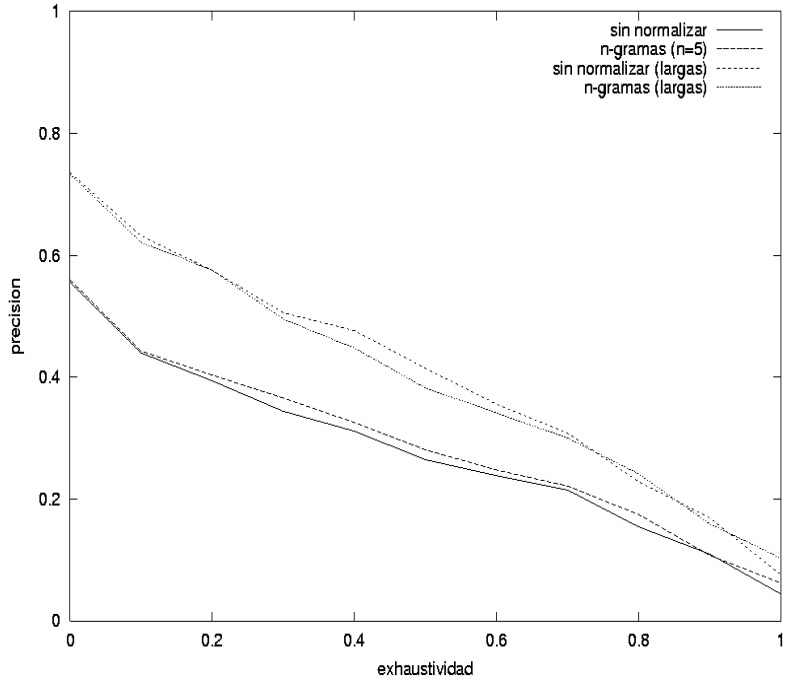

Fig. 3

2. Indización con aproximación lingüística

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, siendo uno de los destacados 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 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 a nivel de categoría gramatical como a nivel 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 de 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 [tzoukerman1997effective], pueden dividirse en propuestas en indexación morfológica, indexación sintáctica e indexación basada en el sentido de las palabras.

2.1 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" [krovetz1993viewing]. Una versión de un "s-temmer" 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" [figuerola2001spanish].

Estos problemas derivados de una "poda ciega" pueden evitarse contando con un lexicón computacional y un adecuado procesador morfológico, es decir, con los recursos y herramientas de PLN capaces de determinar para cada representación superficial de una palabra, todas –no lo olvidemos- las posibles categorías gramaticales, con su forma canónica correspondiente. La morfología computacional ha experimentado un fuerte desarrollo en la pasada década y ha adoptado técnicas muy eficientes desde el punto de vista computacional (morfología de estados finitos) [alegria1993morfologia] para llevar a cabo el reconocimiento de las palabras, hasta el punto de considerarse, al menos para las lenguas dominantes, un problema prácticamente resuelto.

No obstante, utilizar este procesador lingüístico para llevar a cabo la normalización de las palabras en el momento de la indexación obliga a incorporar un desambiguador categorial (Part of Speech Tagger), esto es, una herramienta capaz de asignar, para cada palabra de un texto, una única categoría gramatical, dado el contexto de aquélla. Como ejemplo de lo que venimos diciendo, supóngase el texto siguiente: "el príncipe no se casa". Centrándonos en la palabra "casa", la salida del lexicón y el procesador morfológico nos devolvería "casa" como "sustantivo" (S), cuya forma canónica sería la misma, y "casa" como una flexión del verbo (V) casar . Por tanto, para indexar dicho documento por sus formas canónicas, ha de determinarse previamente cuál es la correcta: casa/S o casar/V.

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 [chowdury1998improving].

Obviamente, también pueden pensarse ejemplos en los que una indexación de acuerdo a dicho par acarreen efectos negativos: por ejemplo, si discriminamos "diseño" como verbo y como sustantivo después de la desambiguación, al no tratar el fenómeno morfológico de la "derivación", obtenemos una representación diferenciada para ambas formas (diseño/S, diseñar/V), aunque en los textos a indexar, las apariciones de dicha palabra con diferente categoría estén siendo utilizadas para expresar el mismo concepto [KROVETZ 92]. Como factor en contra, además, al utilizar técnicas lingüísticas para acometer la indexación morfológica hay que contar con el considerable aumento de recursos computacionales que implica respecto del uso de las técnicas no lingüísticas antes mencionadas.

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 inglés; 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 de no lingüísticos, con lo que no resulta aconsejable el uso de las primeras dado la diferencia en el coste computacional. En [gonzalo1999lexical] se comprueba, incluso, que la indexación con POS-Tagging producía mejoras inapreciables frente a la indexación de palabras ortográficas. Siguiendo con el inglés, respecto si supone mejoras el uso de "stemmers" o no, en [harman1991???] se concluye que no y, por el contrario, en [krovetz1997homonymy] se comprueban mejoras para diferentes "algoritmos de poda". Respecto del español, los resultados obtenidos en

[figuerola2001spanish] parecen indicar que las técnicas de "stemming" producen efectos beneficiosos frente a los métodos que no realizan ninguna normalización y, por otro lado, en [vilares2003manejando] se realizan, entre otros, experimentos con técnicas de stemming y con herramientas lingüísticas para tratar la morfología flexiva, obteniendo mejores resultados utilizando estas últimas. 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 [kraaij1998comparing] [monz2002shallow], como no lingüísticas [mcname2001language].

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 95-97% o incluso superior [rodriguez1999tecnicas]), según se desprende de [gonzalo2002indexacion], no parecen relevantes.

2.3. 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 que:

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") [strazalkowski1999evaluating].

Se ha experimentado con diferentes métodos para evaluar los efectos de una indexación multipalabras. Básicamente, las técnicas empleadas pueden agruparse en estadísticas y lingüísticas. Las primeras se limitan a recolectar coocurrencias de pares de palabras en los textos (bigramas) [pickens2000????]. Las segundas utilizan métodos de análisis sintáctico superficial ("Shallow Parsing") para reconocer diferentes estructuras sintagmáticas, más o menos complejas, a veces no sólo para detectar determinadas secuencias de etiquetas gramaticales (con lo cual son más selectivos que la simple recolección de pares de palabras que efectúan los métodos anteriores), sino también para normalizar diferentes árboles sintácticos en patrones comunes a la hora de indexar [strazalkowski1999evaluating].

El "analizador sintáctico superficial" se añade a la salida de un POS-tagger, lo que supone utilizar un conjunto de recursos y herramientas que acarrean un coste computacional importante. Respecto de los resultados obtenidos por la investigación, en el sentido de si suponen una mejora o no respecto de la indexación basada en simples palabras ortográficas, éstos se muestran, cuando menos, contradictorios [peñas2004tecnicas]. Aunque ha sido un tema investigado desde fines de los 70, no fue, no obstante, hasta la última mitad de los 90 cuando se pudo experimentar la indexación sintagmática con grandes colecciones documentales aplicando las modernas técnicas de PLN de análisis robusto.

Las conclusiones obtenidos 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 [zhai1997evaluation], [strazalkowski1999evaluating].

2.4 Indexación 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, siendo los principales los "diccionarios" y la red semántica de palabras WordNet [peñas2004tecnicas].

La indexación 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" [MILLER 91]. Sirva como un pequeño ejemplo de indexación basada en los sentidos el siguiente: supóngase que la palabra "coche", por simplificar, tiene dos sentidos, de "vehículo a motor" (1) y de "coche de bebé" (2) y la palabra "automóvil" uno sólo, el primero que hemos asignado a "coche". Indexando de acuerdo a "diccionarios", todas las ocurrencias de ambas palabras en documentos y preguntas se indexarían, p.e., en la forma, *palabra#número de sentido*. Es decir, "coche#1" y "automóvil#1" significaría que en el documento o pregunta a indexar apareció dicha palabra con el sentido de "vehículo a motor" y "coche#2" con el sentido de "coche de bebé". Con lo que se adoptan representaciones diferenciadas para las distintas acepciones de "coche" (resolviendo el problema de la polisemia), pero distintas para las apariciones de "coche" como vehículo a motor y "automóvil" (con lo que no resolvemos la sinonimia). Por el contrario, indexando de acuerdo a WordNet, dónde cada "synset" –para las categorías gramaticales nombre, adjetivo, verbo y adverbio- tiene un identificador único, la representación puede ser en la forma *categoría#número de synset*. Esto es, todas las apariciones del concepto "vehículo a motor", cuyo "synset", supongamos, se identifica en la base de datos como "N#123" quedarían representadas de acuerdo a dicho identificador y las referidas al concepto "coche de bebé", pongamos por caso, al identificador "N#322".

Naturalmente, el proceso de asignar el sentido correcto a las palabras de los textos debe realizarse de forma automática. La desambiguación del sentido de las palabras (Word Sense Disambiguation) es un problema computacionalmente complejo que se aborda desde diferentes planteamientos [veronis1998word], pero que aún necesita perfeccionarse. A falta de conocer los resultados de la conferencia SENSEVAL de 2004 (no estaban disponibles en línea al escribir estas páginas), el mejor sistema de la celebrada en 2001 en la modalidad "grano fino" para el inglés conseguía una precisión del 69% (ver resultados en http://www.sle.sharp.co.uk/senseval2/Results/guidelines.htm).

En cuanto a los experimentos aplicados a la indexación, resumiendo, se han concentrado en dos aspectos principales [gonzalo1999lexical]:

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) [smeaton1996experiments], con lo que los resultados no pueden extrapolarse a colecciones de grandes volúmenes de texto.

Diversos trabajos se han centrado en el segundo aspecto. Para establecer la precisión mínima exigida a la desambiguación en tareas de RI, se han ideado métodos artificiales (las pseudo-palabras de M. Sanderson [sanderson00????]), se ha recurrido a la desambiguación manual de pequeños pasajes de texto (dado el coste tiempo que ello supone) o se han utilizado "corpus" desambiguados (p.e. el "SEMCOR"). Por otro lado, al mismo tiempo, la indexación unas veces se ha realizado según diccionarios y otras de acuerdo a los "synsets". Por ello, los datos ofrecidos por los diferentes trabajos de investigación no se aceptan como concluyentes o, incluso, se ponen en entredicho. Así, p.e. en [sanderson????] se concluye que si se desambigúa con una precisión inferior al 90%, la recuperación se degrada, pero queda por establecer si la ambigüedad artificialmente introducida con las pseudo-palabras es comparable con las palabras reales, y en los experimentos de [gonzalo1999lexical] se fija la precisión sobre el 60%, concluyéndose que la diferencia se debe, por una lado, a que la indexación con WordNet es más tolerante a errores y, por otro, al tratar con palabras reales. Sin embargo, estos últimos resultados se cuestionan en [sanderson2000retrieving], por no utilizar una colección estándar de evaluación y por la forma de crear documentos a partir de los pasajes del SEMCOR. 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 [krovetz1997homonymy].

3. Expansión de consultas

En el apartado anterior se ha descrito la importancia de la indización en RI y cómo puede mejorarse ésta con diversos mecanismos. Sin embargo, aunque se realice un buen proceso de indización, a menudo los usuarios no encuentran respuestas adecuadas a sus necesidades informativas, frecuentemente ello se debe a la manera en que ésta se representa.

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.

El interés para realizar expansión de consultas se centra en consultas con muy pocos términos, pues las consultas largas suelen proporcionar buenos resultados de recuperación, al incluir más términos comunes con los documentos relevantes. De hecho, la mayor parte de las consultas que se realizan en buscadores y sistemas de información en Internet tienen de uno a tres términos [wolfram2001public].

Para expandir la consulta deben utilizarse palabras o frases con significado similar a aquellos de la consulta inicial. La idea es que si varios términos están semánticamente relacionados entre sí, cuando un usuario está interesado en uno de ellos, probablemente también lo estará en los otros, y los documentos indizados con éstos también serán relevantes para el usuario. Para desarrollar la expansión de consultas hay que resolver tres aspectos importantes:

1. En primer lugar, hay que establecer la relación entre términos. Lo inmediato es utilizar tesauros o diccionarios en los que aparecen relaciones entre términos.
2. Después hay que seleccionar, de todos los términos relacionados con los de la consulta, cuáles son más adecuados para ser añadidos a dicha consulta.
3. Por último, teniendo en cuenta el sistema de recuperación utilizado, hay que determinar el mecanismo de pesado de los nuevos términos, y ello depende del criterio de selección previo.

Para realizar la expansión lo más rápido sería utilizar tesauros o diccionarios generales ya existentes. Aunque a este respecto suele citarse con cierta frecuencia los malos resultados que obtuvo Voorhees en 1994 [vorhees1994query] con una herramienta general como WordNet, la realidad es que otros estudios demuestran que la procedencia del tesauro no es tan importante [mandala2000query]; parece tener mayor importancia los otros dos aspectos: la manera de seleccionar los términos y el mecanismo de pesado.

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:

1. *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*".

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

   a) *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 clustering local (tesauros locales de términos).

   b) *Análisis global*. Utiliza información de toda la colección de documentos para expandir la consulta. Se suelen emplear mecanismos de clustering 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 [qiu1993concept], tesauros construidos a partir de la asociación de términos y frases (*phrase-finder*) [jing1994assocation], y tesauros basados en información sintáctica [grefenstette1992use].

Muchas de estas técnicas requieren un conocimiento lingüístico más o menos sofisticado, como la técnica *Phrase-finder*, en la que se determinan agrupamiento de nombres, nombres y adjetivos, etc., o la que aplica tesauros basados en información sintáctica. En general, este tipo de técnicas no son sustancialmente mejores que las que utilizan análisis puramente estadístico [gonzalo2002indexacion].

Es importante señalar que pueden ser consideradas técnicas de expansión otros mecanismos aparentemente alejados de la clasificación que acabamos de ver. Uno de los ejemplos más evidentes es la lematización (véase el apartado anterior). Al aplicar la lematización, cada palabra de la consulta se expande con aquellas que poseen su mismo lema.

A continuación presentamos las técnicas de expansión que no requieren conocimiento lingüístico para llevarse a cabo. Para comprobar la bondad de cada uno de ellas, hemos realizado varias pruebas sobre una colección documental de 215.000 documentos y 50 preguntas cortas (2,64 términos de media por consulta). Hemos utilizado el valor de precisión media ($\overline{P}$) y precisión a 10 documentos vistos (*P@10*). Utilizamos esta última medida ya que frecuentemente los usuarios no suelen mirar más que la primera pantalla de resultados, esto es, los primeros 10 documentos.

3.1 Aplicación de lematización

Utilizando un s-stemmer que elimina las terminaciones -os, -as, -es, -o, -a, -e de las palabras (sin considerar su categoría gramatical) hemos obtenido una mejora del 11,46 y 10,84 % en $\overline{P}$ y *P@10* respectivamente. Se trata de un proceso con muy poco gasto computacional que se realiza en el proceso de indización.

3.2 Realimentación de consultas con criterios de relevancia del usuario

El usuario puede visualizar los documentos recuperados, y marcar los que considera relevantes y no relevantes para su necesidad informativa. El sistema elabora una nueva consulta teniendo en cuenta los términos de la consulta original, y los términos de los documentos relevantes y no relevantes. En general se suele utilizar el algoritmo de Rocchio [rocchio1971relevance], ec. (1), en el que es preciso ajustar tres coeficientes (α, β y γ). En nuestro caso hemos utilizado valores de α y β mayores que γ. Los resultados son extraordinarios, con una mejora en $\overline{P}$ y *P@10* del 300,1% y del 301,2%, respectivamente. El mayor inconveniente es que requiere de la presencia del usuario.

$$\vec{q'} = \alpha\vec{q} + \frac{\beta}{n_{rel}} \sum_{d_j \in rel} \vec{d_j} - \frac{\gamma}{n_{norel}} \sum_{d_j \in norel} \vec{d_j} \qquad (1)$$

3.3 Pseudo realimentación de consultas

La idea es utilizar la realimentación de consultas de manera automática, esto es, suponiendo que los primeros documentos recuperados son relevantes. Suele utilizarse el algoritmo de Rocchio con el coeficiente γ = 0. En general este método consigue mejorar los valores medios de $\overline{P}$ y *P@10*, si bien,

algunas consultas empeoran, precisamente aquellas que obtienen malos resultados sin realizar expansión. En nuestro experimento hemos considerado relevantes los primeros 5 documentos recuperados, y hemos lanzado la nueva consulta (q') con los 40 términos de más peso, obteniendo una mejora en $\overline{P}$ y *P@10* del 10,73% y 8,43%, respectivamente.

3.4 Utilización de tesauros

Un tesauro es una matriz que mide relaciones entre términos. Estas relaciones pueden calcularse automáticamente computando relaciones de coocurrencia, tanto de términos como de documentos: si dos términos coocurren en el mismo documento, de alguna forma estarán relacionados (tesauro de asociación); si dos documentos poseen términos comunes, también estarán relacionados (tesauro de similitud). Asimismo, se pueden construir tesauros globales (con la información de toda la colección documental), o tesauros locales (con información de los documentos recuperados).

Experimentalmente se comprueba que los tesauros de asociación y similitud obtienen resultados de expansión muy parecidos entre sí, si bien, el computo de los de similitud es extraordinariamente elevado en comparación con los de asociación. Las tres funciones más utilizadas para medir el grado de asociación entre dos términos t_i y t_k son las de Tanimoto (Jaccard), Coseno y Dice, ec. (2), donde n_i y n_k son el número de documentos en los que aparece el término t_i y t_k, respectivamente, y n_{ik} el número de documentos en los que coocurren t_i y t_k .

$$\begin{aligned} \text{Tanimoto}(t_i, t_k) &= \frac{n_{ik}}{n_i + n_k - n_{ik}} \\ \text{Coseno}(t_i, t_k) &= \frac{n_{ik}}{\sqrt{n_i \cdot n_k}} \\ \text{Dice}(t_i, t_k) &= \frac{2 \cdot n_{ik}}{n_i + n_k} \end{aligned} \tag{2}$$

Como ejemplo, en la Tabla 1 aparecen los 20 mejores términos relacionados con la entrada “espacial” en el tesauro de asociación global.

Tanimoto		Coseno		Dice	
espacial	1.0000	espacial	1.0000	espacial	1.0000
nasa	0.2788	astronautas	0.4860	nasa	0.4361
astronautas	0.2670	nasa	0.4755	astronautas	0.4215
espaciales	0.2583	espaciales	0.4472	espaciales	0.4105
orbita	0.2336	cañaveral	0.4002	orbita	0.3788
transbordador	0.2026	astronauta	0.3935	transbordador	0.3369
astronauta	0.1846	orbita	0.3838	astronauta	0.3117
cañaveral	0.1809	orbital	0.3609	cañaveral	0.3063
orbital	0.1540	transbordador	0.3443	orbital	0.2669
nave	0.1339	shuttle	0.3250	nave	0.2361
shuttle	0.1310	discovery	0.3088	shuttle	0.2317
aeronautica	0.1250	soyuz	0.2461	aeronautica	0.2222
cohete	0.1196	transbordadores	0.2450	cohete	0.2136
satelites	0.1153	nave	0.2377	satelites	0.2067
transbordadores	0.1152	hubble	0.2372	transbordadores	0.2066
discovery	0.1121	cosmonautas	0.2335	discovery	0.2016
satelite	0.1106	cohete	0.2284	satelite	0.1991
experimentos	0.1049	cosmonauta	0.2283	experimentos	0.1899
jirl	0.0970	espacio	0.2242	jirl	0.1768
kennedy	0.0928	aeronautica	0.2225	kennedy	0.1698
endeavour	0.0874	atlantis	0.2221	endeavour	0.1607

Tabla 1. Ejemplo de expansión para el término “espacial”.

Un aspecto muy importante es la selección de términos para realizar la expansión. Hemos comprobado que los resultados óptimos se obtienen cuando la expansión se realiza considerando los mejores términos relacionados *con todos los términos de la consulta original*, y no con cada uno de ellos por separado. Algunos autores [peat1991limitations] han denostado la utilización de tesauros en tareas de expansión de consultas, sin duda, ello se ha motivado por la manera de seleccionar los términos añadidos: se incluían todos los relacionados con cada término de la consulta, o aquellos que superaban un cierto umbral.

En relación con la utilización de tesauros locales o globales, podemos decir que la aplicación de los locales proporciona mejores resultados, curiosamente, cuando más local es el tesauro. En las Figura 2 y 3 puede verse el grado de mejora con ambos métodos.

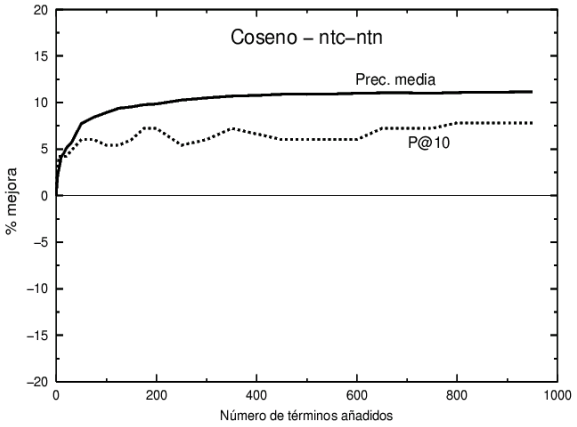

Figura 4. Tesauro de asociación global.

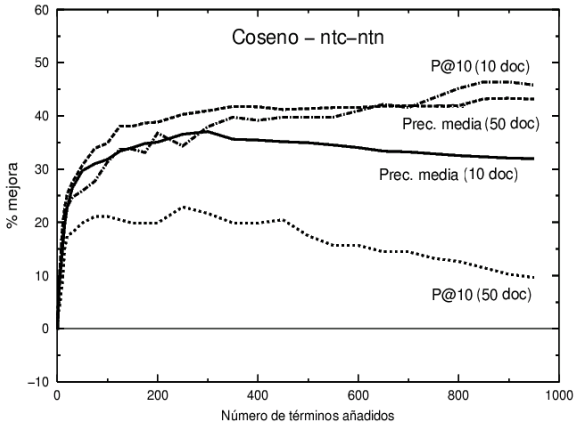

Figura 5. Tesauro de asociación local (utilizando 50 términos y 10 documentos.)

Como conclusión de este apartado, debemos decir que la expansión de consultas es una técnica muy válida para aumentar la eficacia de los sistemas de recuperación de información. Las técnicas que hemos descrito no implementan conocimiento lingüístico, y en general son *fáciles* de aplicar. Si dejamos al margen la aplicación de la lematización, que suele realizarse en muchos sistemas, destaca sobre todas ellas la pseudo-realimentación de consultas, de hecho es la más utilizada, al no requerir la presencia del usuario. Recordemos que el objetivo es que el usuario encuentre documentos más adecuados a sus necesidades informativas.

4. La Recuperación de Información en el web

4.1 Introducción

Las técnicas de Recuperación de Información que se han empleado en el 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[1] o el enorme tamaño que deben soportar los índices, haciendo más difícil su adecuada gestión mediante el empleo de los modelos tradicionales.

Baeza-Yates [baeza1999modern] afirma que en el web existen básicamente tres formas de buscar información:

[1]Los creadores de páginas web insertan en la descripción de las mismas términos que no tienen nada que ver con el contenido. Ello implica que al buscar se puedan obtener estas páginas sin corresponder con la temática deseada por el usuario

Emplear motores de búsqueda: El funcionamiento básico de los motores de búsqueda consiste en indexar una porción importante de los documentos residentes en la globalidad del web y posteriormente permiten la localización de la información a través de la formulación de una pregunta.

Empleo de directorios: Cuando se utilizan los directorios, se realiza una clasificación de los documentos web por materias. Posteriormente podemos navegar por las diferentes secciones o buscar en los índices realizados al efecto.

Buscar explotando la estructura hipertextual: Esta modalidad no está del todo disponible.

A continuación mostramos una comparativa entre los dos primeros mecanismos comentados con anterioridad.

	Descubrimiento de recursos	Representación del contenido	Representación de la consulta	Presentación de los resultados
Motores de búsqueda	Automática por robots	Indización automática	Explícita (palabras clave, operadores)	Páginas creadas dinámicamente en cada consulta. Exhaustivos y poco precisos
Directorios	Lo realizan las personas	Clasificación manual	Implícita (navegación por categorías)	Páginas creadas antes de la consulta. Poco exhaustivos, muy precisos.

Tabla 2: Comparativa Motores vs. Directorios. Fuente:Delgado Domínguez, A. Mecanismos de recuperación de información en la WWW [En línea]. Palma de Mallorca, Universitat de les Illes Balears, 1998. http://servidorti.uib.es/adelaida/tice/modul6/memfin.pdf [Consulta: 2 de febrero de 2005]

4.2 Explotando la estructura hipertextual

Este 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.

La búsqueda dinámica es "'equivalente a la búsqueda secuencial en textos"' [baeza1999modern]. Se realiza una búsqueda online para descubrir información relevante siguiendo los enlaces de las páginas recuperadas.

Algunos autores [chang2001mining] añaden a los tres sistemas comentados los siguientes:

Metabuscadores: Son sistemas desarrollados para mitigar el problema de tener que acceder a varios motores de búsqueda con el fin de recuperar una información más completa sobre la temática de interés. La petición realizada por los usuarios se envía a todos los motores de búsqueda contemplados por el metabuscador, obtiene los resultados y algunos detectan las URL duplicadas y eliminan la redundancia.

Filtrado de información: Se trata de un complemento de los motores de búsqueda más que un modelo alternativo. El concepto de filtrado tiene que ver con la decisión de considerar (a priori) si un documento es relevante o no, eliminándolo del índice en caso contrario. El filtrado de términos mejora la calidad del índice del motor y se acelera la velocidad de la recuperación de información.

En el siguiente esquema podemos ver el proceso de filtrado.

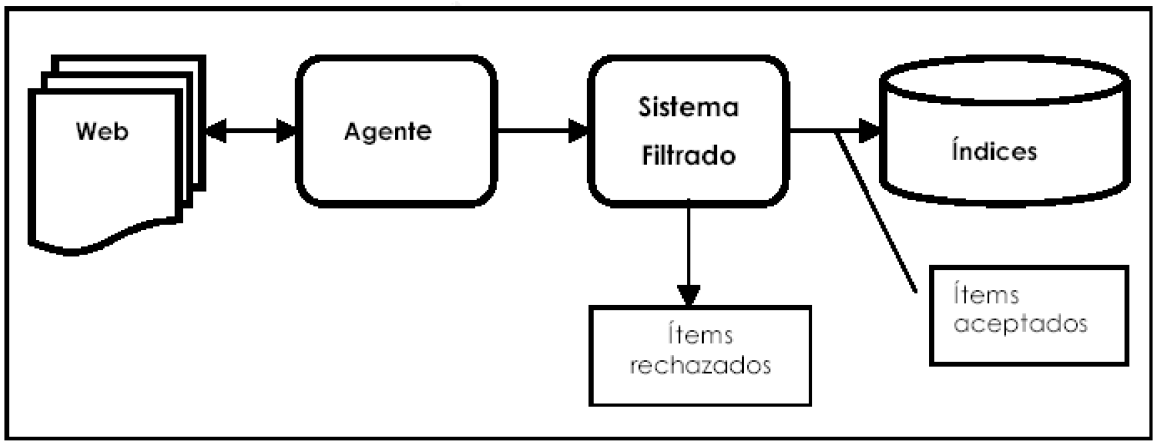

Fig. 6

4.3 Funcionamiento de los motores de búsqueda

Es importante hacer referencia al mecanismo empleado en la arquitectura robot-indexador, que tiene dos fases:

- La recopilación de la información por parte del robot.
- La fase de indexación para una posterior recuperación de la información.

De forma gráfica lo resumimos en la siguiente imagen:

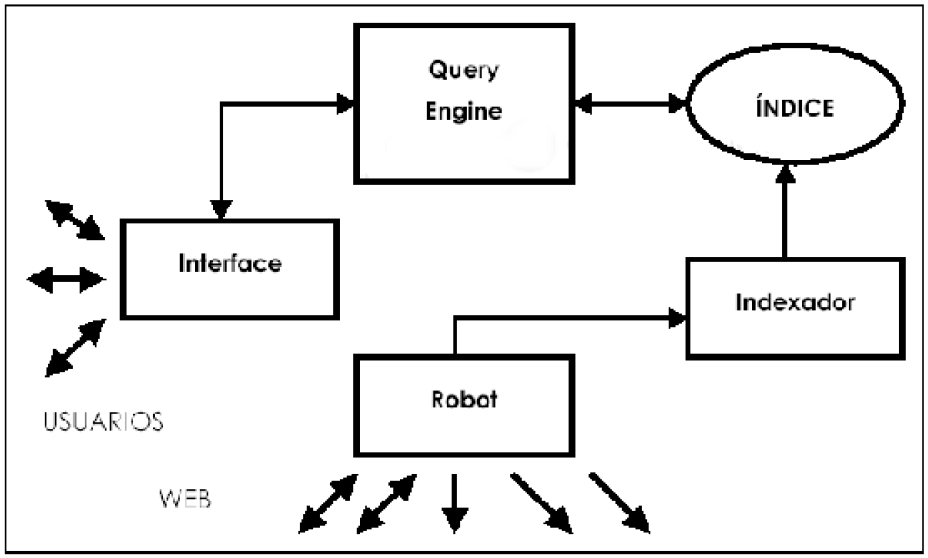

Fig. 7

Los robots son programas que de forma automática permiten rastrear el web. Inician el rastreo a partir de una dirección URL y se siguen los enlaces contenidos en esa URL [baldi2003modeling], [chakrabarti2003mining].

Existen otras modalidades que indicamos a continuación:

Knowbots: programados para localizar referencias hipertexto dirigidas hacia un documento. Permiten evaluar el impacto de las diferentes aportaciones de áreas del conocimiento.

Wanderers (vagabundos): encargados de realizar estadísticas.

Worms (gusanos): encargados de la duplicación de directorios ftp.

WebAnts (hormigas): conjunto de robots alejados físicamente, que cooperan.

Una vez realizada la fase de recogida de datos con el robot se crea el índice, verdadero corazón del motor de búsqueda. Este índice normalmente consiste en una lista de palabras asociadas a sus correspondientes documentos y para ello se suele emplear un fichero inverso similar al mostrado en la siguiente imagen:

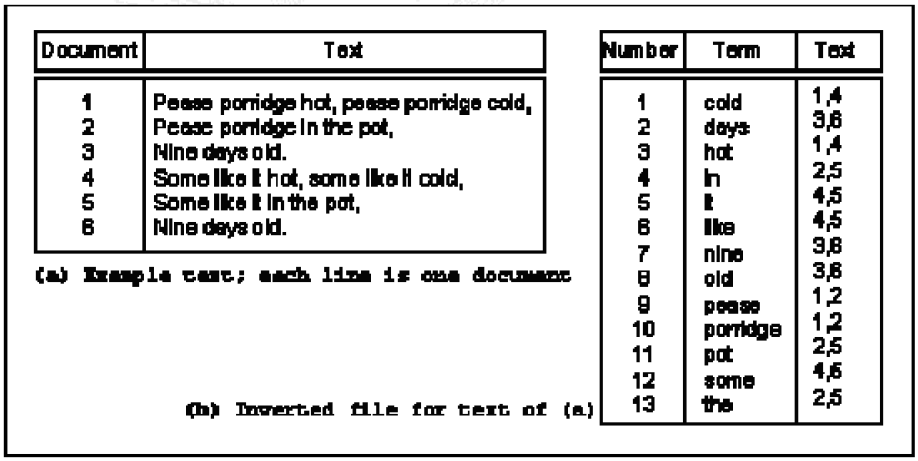

Fig. 8

El fichero inverso se convierte en una enorme estructura de datos con problemas de gestión, debiendo recurrir a las técnicas más avanzadas para poder mejorarla. Normalmente se debe acudir a técnicas que nos permitan simplificar el tamaño de los índices, como pueden ser:

- Conversión de textos a minúsculas.
- Stemming.
- Supresión de palabras vacías.
- Compresión de textos.

Lógicamente el proceso de indización debe ser automático y la información para generar estos índices se puede obtener de la información suministras por los creadores o editores de las páginas web, mediante la etiqueta <title> o bien mediante la información de los metadatos suministrada. Además podemos extraer la información directamente del documento. Normalmente se mezclan ambas posibilidades.

Adicionalmente, se pueden asignar pesos a los términos obtenidos, en función de diferentes criterios:

- Si aparece en el título o no.
- Según la frecuencia absoluta.
- Si la página es grande en tamaño.
- Si aparece la información en los metadatos.

4.4 Técnicas de ranking

Finalmente, un aspecto muy importante a tener en cuenta es el ranking, es decir, el orden en el que se presentan los resultados al usuario, en función de la relevancia de los documentos respecto a la pregunta realizada.

Esta discriminación permite que aparezcan en primer lugar los documentos más relevantes, facilitando el acceso a la información. Se desconoce como se realizan estas tareas en la mayoría de los buscadores, pero se emplean de forma habitual.

Existen dos grandes variantes en los algoritmos de ranking:

1. Variantes del modelo vectorial o booleano.
2. Las que siguen el principio de extensión de los enlaces.

De la primera variante existen tres métodos:

- Booleano extendido.
- Vectorial extendido.
- Más citado.

De la segunda variante existen bastantes métodos, pero destacaremos:

- WebQuery.
- HITS. Se puede ampliar información en [kleinberg1999authoritative]
- PageRank. Se puede ampliar información en [brin1998anatomy]

El que mayor éxito tiene en la actualidad es el PageRank, utilizado en el buscador Google.

En este algoritmo, la importancia de una página viene dada por la importancia de las páginas que la enlazan y se determina de la siguiente manera:

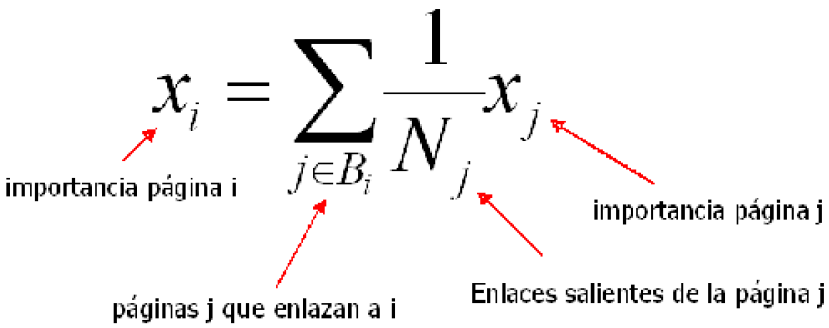

Fig. 9

En este esquema es fundamental la importancia de las páginas que enlazan a una dada. Si las páginas que la apuntan son páginas importantes, el valor del PageRank aumenta, con respecto a páginas de menor importancia. Es importante estar enlazado por buenas páginas, que poseerán un alto PageRank.

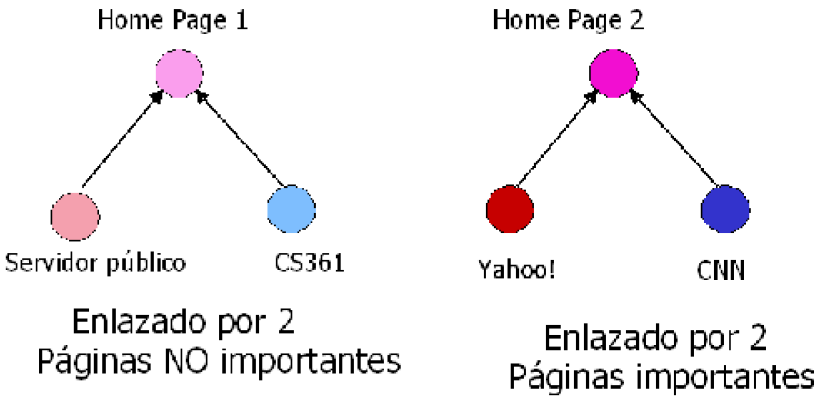

Fig. 10

5 Conclusiones

El crecimiento espectacular de los documentos disponibles en formato electrónico ha puesto de relieve el problema de la Recuperación de la Información. Los enfoques más utilizados para abordar este problema se basan en la utilización de los términos o palabras que aparecen en los documentos, complementados en ocasiones con otros elementos de información, como los hiperenlaces en el caso de las páginas web. En cualquier caso, este tipo de enfoques sugieren la aplicación de diversas técnicas y tratamientos a palabras y demás elementos informativos, como puedan los las características hipertexto de algunos documentos.

En este trabajo hemos revisado los procesos y tratamientos más importantes, sus bases teóricas y las opciones de implementación más importantes. Algunas de tales tienen una base meramente estadística, otras se apoyan en planteamientos lingüísticos, y otras ambas cosas. La recuperación de documentos o

páginas web permite tomar en cuenta otros elementos como los hiperenlaces y la estructura de grafo del web.

Se han descrito también los sistemas de medición y evaluación normalmente aplicados, que permiten estimar los efectos de la aplicación de las diversas técnicas. Parte de ellas se utilizan con éxito en sistemas de recuperación operativos; otras son objeto de experimentación todavía aunque los resultados disponibles parecen prometedores.

REFERENCIAS:

[alegria1996morfologia] Alegría, I.: "Morfología de estados finitos", SEPLN (18), págs. 1-26 (1996).

[baeza1999modern] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. ACM Press, New York, 1999.

[baldi2003modeling] Pierre Baldi, Paolo Frasconi, and Padhraic Smyth. Modeling the Internet and the Web: probabilistic methods and algorithms. Wiley, Chichester, 2003.

[belkin1987retrieval] Belkin, N. J. Y Croft, W. B. (1987): "Retrieval Techniques", en Annual Review of Information Science and Technology, 22, 109-145

[berrocal2002cibermetria] José L. Alonso Berrocal, Carlos G. Figuerola, Ángel F. Zazo, and Emilio Rodríguez. La cibermetría en la recuperación de información en el Web. In Primeras Jornadas de Tratamiento y Recuperación de Información, JOTRI-2002, Valencia, España, 4 y 5 de Julio de 2002, pages 117-124. Facultad de Informática. Universidad Politécnica de Valencia, 2002. ISBN: 84-9705-199-8.

[berrocal2004cibermetria] José Luis Alonso Berrocal, Carlos G. Figuerola, and Ángel F. Zazo. Cibermetría: nuevas técnicas de estudio aplicables al Web. Trea S.L., Gijón, 2004. ISBN: 84-9704-114-3.

[brin1998anatomy] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998.

[chakrabarti2003mining] Soumen Chakrabarti. Mining the Web. Morgan Kaufmann, Amsterdam, 2003.

[chang2001mining] G. Chang et al. Mining the World Wide Web: an information search approach. Kluwer Academic Publishers, Norwell, Massachusetts, 2001.

[chowdury1998improving] Chowdhury, A. y McCabe, M.C.: "Improving Information Retrieval using Part of Speech Tagging" (url:citeseer.ist.psu.edu/256084.html) (1998).

[figuerola2001spanish] Figuerola, C.G.; Gómez, R.; Zazo, A.F. y Alonso, J.L.: "Spanish Monolingual Track: The Impact of Stemming on Retrieval, en Carol P. (ed): Working notes for the CLEF 2001 workshop, Springer (2001).

[gonzalo1999lexical] Gonzalo, J.; Peñas, A. y Verdejo, F.: "Lexical ambiguity and Information Retrieval revisited". En Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC, Maryland, pp. 195-202 (1999).

[gonzalo2002indexacion] Gonzalo, J.; Peñas, A. y Verdejo, F.: "La indexación con técnicas lingüísticas en el modelo clásico de recuperación de información". En Primeras Jornadas de Tratamiento y Recuperación de Información, JOTRI-2002, págs. 97-106. (2002)

[grefenstette1992use] G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. En "Proceedings of the 15th ACM-SIGIR Conference", págs. 89-97. ACM (1992).

[harman1992ranking] Harman, D. K.: "Ranking Algorithms", en Frakes, W. B. y Baeza-Yates, R., eds.: Information Retrieval: Data Structures and Algorithms, Englewood Cliffs (NJ): Prentice-Hall Inc., 1992, pp. 363-392

[harter1997evaluation] Harter, S. P. Y Hert, C. A. (1997): Evaluation of Information Retrieval Systems: Approaches, Issues and Methods" en Annual Review of Information Science and Technology, 32, 3-94

[hooper1965indexer] Hooper, R. S.: Indexer consistency test - origin, measurements, results and utilization, Bethesda, MD., 1965

[jing1994association] Y. Jing y W. B. Croft. An association thesaurus for information retrieval. En "Proceedings of RIAO-94", pág. 146-160, New York, US (1994).

[kleinberg1999authoritative] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, pages 668-677, 1999.

[kraaij1998comparing] Kraaij, W. Y Pohlmann, R.: "Comparing the effect of syntactic vs. statistical phrase index strategies for Dutch". En Proceedings ECDL'98, pp. 605-617 (1998).

[krovetz1992lexical] Krovetz, R. y Croft, B.W.: "Lexical Ambiguity and Information Retrieval", ACM Transactions on Information Systems, 10(2), págs. 115-141 (1992).

[krovetz1993viewing] Krovetz, R.: "Viewing morphology as an Inference Process". En Proceedings of the 6th ACM/SIGIR, pp. 191-203 (1993)

[krovetz1997homonymy] Krovetz, R.: "Homonymy and Polisemy in Information Retrieval". En Proceedings of the 33th Meeting of the ACL, pp. 72-79 (1997).

[mandala2000query] R. Mandala, T. Tokunaga y H. Tanaka. Query expansion using heterogeneous thesauri. Information Processing & Management 36(3), 361-378 (2000).

[mcname2001language] McNamee, P. y Mayfield, J.: "A Language-Independent Approach to European Text-Retrieval". En Peters, C. (ed): Cross-Language Information Retrieval and Evaluation, Workshop, CLEF 2000, Springer, pp. 129-139 (2000)

[miller1991lexical] Miller, G.: WordNet: "A lexical database for english", Communications of the ACM, 38 (11), pp. (-), (1991)

[monz2002shallow] Monz, C. y Rijke, M.: "Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German and Italian". En Peters, C.: Cross-Language Information Retrieval Systems, CLEFF 2001, Springer, pp. 262-277 (2002).

[peat1991limitations] H. J. Peat y P. Willet. The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the American Society for Information Science 42(5), 378-383 (1991).

[peñas2004tecnicas] Peñas Padilla, A.: Técnicas lingüísticas aplicadas a la búsqueda textual multilingüe: ambigüedad, variación terminológica y multilingüismo, SEPLN, 2004.

[qiu1993concept] Y. Qiu y H.-P. Frei. Concept-based query expansion. En "Proceedings of the 16th ACM-SIGIR Conference", págs. 160-169. ACM (1993).

[rijsbergen1979information] Rijsbergen, K. Van: Information Retrieval, London: Butterwoths, 1979

[rocchio1971relevance] J. J. Rocchio. Relevance feedback in information retrieval. En G. Salton, editor, "The SMART Retrieval System. Experiments in Automatic Document Processing", págs. 313-323. Prentice Hall, Englewoods Cliffs, N. J. (1971).

[rodriguez1999tecnicas] Rodríguez Hontoira, H.: "Técnicas estadísticas en el tratamiento del lenguaje natural", en Blecua, J.M. (ed): Filología e informática: nuevas tendencias en los estudios filológicos, Barcelona: UAB, pp. 111-140 (1999).

[salton1975vector] Salton, G., Wong, A. y Yang, C. S.(1975): "A Vector Space Model for Automatic Indexing", en Communication of the ACM, 18, 613-620

[sanderson2000retrieving] Sanderson, M.: "Retrieving with good sense", Information Retrieval, 2 (1), pp. 49-69 (2000).

[smeaton1996experiments] Smeaton, A.F. y Quigley, A.: "Experiments on using semantics distances between words in image caption retrieval". En Proceedings of the 19th International Conference on Research and Development in Information Retrieval (1996).

[spark1999role] Spark Jones, K: "What is the Role of NLP in Text Retrieval?", en : Strazalkowski (ed.): Natural Language Information Retrieval, Kluwer Academic Press, Elsevier, págs. (Xx-xx) (1999).

[sproat1992morphology] Sproat, R.: Morphology and Computation. The MIT Press, 1992

[strazalkowski1999] Strazalkowski, T.; Lin, F.; Wang, J.; y Pérez-Carballo, J.: "Evaluating NLP techniques in Information Retrieval. A TREC Perspective", en : Strazalkowski (ed.): Natural Language Information Retrieval, Kluwer Academic Press, Elsevier, págs. 113-145 (1999).

[stubbs2000internal] Stubbs, E. A., Mangiaterra, N. E. y Martínez, A.M (2000): "Internal quality audit of indexing: a new application of interindexer consistency", en Cataloguing & Classification Quaterly, 28(4), 53-70

[tzoukerman1997effective] Tzoukerman, E.; Klavans, J.L.; Jacquemin,C.: "Effective Use of Natural Language Processing of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing". En Proceedings of 20th ACM/SIGIR, pp. 148-155 (1997).

[veronis1998word] Ide, N. y Veronis, J. : "Word sense disambiguation. The state of the art", Computational Linguistics, 24 (1), pp. 1-40 (1998).

[vilares2003manejando] Vilares Ferro, J.; Barcala Rodríguez, F.M.; Fernández Lanza, S. y Pombo Otero, J.: "Manejando la variación morfológica y léxica en Recuperación de Información Textual", SEPLN (30), págs. 99-106 (2003).

[voorhees1994query] E. M. Voorhees. Query expansion using lexical-semantic relations. En "Proceedings of the 17th ACM-SIGIR Conference", págs. 61-69. ACM (1994).

[voorhees1999natural] Voorhees, E.M.: "Natural Language Processing and Information Retrieval" en Pazienza, M.T.(ed): Information Extraction: Towards Scalable, Adaptable Systems, New York: Springer, pp. 32-48 (1999).

[wolfram2001public] D. Wolfram, A. Spink, B. J. Janses y T. Saracevic. Vox populi: The public searching of the web. Journal of the American Society for Information Science and Technology 52(12), 1073-1074 (2001).

[zhai1997evaluating] Zhai, C.; Tong, X.; Mili-Frayling, N. y Evans D.: "Evaluation of syntactic phrase indexing - CLARIT TREC5 NLP track report", en The Fifth Text Retrieval Conference (TREC-5), NIST Special Publication, (1997).