LAS TÉCNICAS DE SOFT-COMPUTING EN LA RECUPERACIÓN DE INFORMACIÓN

José Angel Olivas Varela
Doctor Ingeniero en Informática.
E-mail: joseangel.olivas@uclm.es

1. Introducción.

En los últimos años, la cantidad de información digital ha crecido en forma vertiginosa, llegando a niveles jamás alcanzados antes. De hecho, estimaciones de la *International Data Corporation* indican que entre 1999 y 2003 se ha publicado en la Web un volumen de datos equivalente a toda la información generada por la humanidad desde la antigüedad hasta 1998. La Web, es una base de datos distribuida, textual y multimedia con un tamaño inmanejable en su globalidad.

Esto representa un desafío en muchos aspectos, en particular para manipular, buscar y recuperar la información y el conocimiento contenido, teniendo en cuenta que los datos que provienen de la Web además de incluir el contenido de las páginas y los enlaces entre ellas, también incluyen ficheros (*logs*) sobre el uso de la misma. Se podría decir entonces que en la Web hay esencialmente tres tipos distintos de datos: contenido semi-estructurado (texto, multimedia, etc), estructura de enlaces (grafos) y datos de uso y que todos estos datos no son estáticos, sino que cambian en el tiempo.

Los portales y motores de búsqueda actuales suelen ser muy eficientes, pero los resultados que proporcionan no suelen ser plenamente satisfactorios para los usuarios (demasiada información no buscada o errónea, falta de información relacionada recuperada), por ello se propone el desarrollo y pruebas de mecanismos más "inteligentes" de acceso, búsqueda, gestión y recuperación de información y conocimiento contenidos en la Web.

2. Estado de los conocimientos sobre el tema.

Desde que en el año 1965 el profesor Lotfi A. Zadeh introdujo la Lógica Borrosa o Difusa (Fuzzy Logic), muchas han sido sus aplicaciones, las más importantes en el campo del control industrial (lavadoras Bosch con sistema ECO-Fuzzy, ABS de Nissan, Aire Acondicionado Mitsubishi...). Pero, ya desde sus inicios, la intención del profesor Zadeh era introducir un formalismo capaz de representar y manipular la incertidumbre e imprecisión inherentes al lenguaje natural.

Por otro lado, la mayor parte de la inmensa cantidad de información contenida en Internet, está almacenada en documentos textuales en lenguaje natural en multitud de idiomas.

El profesor Zadeh describe qué es *Soft-computing* en estos términos[1]:

"What is soft computing?

Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty and partial truth. In effect, the role model for soft computing is the human mind. The guiding principle of soft computing is: Exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost. The basic ideas underlying soft computing in its current incarnation have links to many earlier influences, among them my 1965 paper on fuzzy sets; the 1973 paper on the analysis of complex systems and decision processes; and the 1979 report (1981 paper) on possibility theory and soft data analysis. The inclusion of neural network theory in soft computing came at a later point. At this juncture, the principal constituents of soft computing (SC) are fuzzy logic (FL), neural network theory (NN) and probabilistic reasoning (PR), with the latter subsuming belief networks, genetic algorithms, chaos theory and parts of learning theory. What is important to note is that SC is not a melange of FL, NN and PR. Rather, it is a partnership in which each of the partners contributes a distinct methodology for addressing problems in its domain. In this perspective, the principal contributions of FL, NN and PR are complementary rather than competitive.

Implications of soft computing.

The complementarity of FL, NN and PR has an important consequence: in many cases a problem can be solved most effectively by using FL, NN and PR in combination rather than exclusively. A striking example of a particularly effective combination is what has come to be known as neurofuzzy systems. Such systems are becoming increasingly visible as consumer products ranging from air conditioners and washing machines to photocopiers and camcorders. Less visible but perhaps even more important are neurofuzzy systems in industrial applications. What is particularly significant is that in both consumer products and industrial systems, the employment of soft computing techniques leads to systems which have high MIQ (Machine Intelligence Quotient). In large measure, it is the high MIQ of SC-based systems that accounts for the rapid growth in the number and variety of applications of soft computing - and especially fuzzy logic.

The conceptual structure of soft computing suggests that students should be trained not just in neural network theory or fuzzy logic or probabilistic reasoning but in all of the associated

[1] http://www.cs.berkeley.edu/Research/Projects/Bisc/bisc.welcome.html.

methodologies, though not necessarily to the same degree. This is the principle which guides the BISC Seminar on Soft Computing and the course Fuzzy Logic, Neural Networks and Soft Computing which I teach at present. The same applies to journals, books and conferences. We are beginning to see the appearance of journals and books with soft computing in their title. A similar trend is visible in the titles of conferences."

A la hora de clasificar las diferentes líneas de investigación relacionadas con la recuperación de información, y más en concreto, con las posibilidades de las técnicas de *Soft-Computing* en la recuperación de Información en Internet, se pueden utilizar diferentes criterios.

Una posibilidad es realizar una clasificación en función de la parte del proceso de recuperación de información en la que están centradas. En este caso podemos distinguir los siguientes grupos de enfoques:

- Modelos de representación lógica de documentos.
- Lenguajes de especificación de consultas.
- Sistemas de evaluación de consultas.
- Sistemas de presentación y clasificación de resultados.
- Sistemas de retroalimentación de consultas.

Otra posibilidad es distinguir los enfoques según las técnicas utilizadas:
- Utilización de ontologías.
- Estudio de asociaciones y relaciones entre términos.
- Construcción y utilización de perfiles de usuarios.
- Utilización de algoritmos de *clustering* y clasificación.

Atendiendo al objetivo y el ámbito de aplicación de los sistemas de recuperación de información, podemos distinguir diferentes tipos de sistemas:
- Sistemas de búsquedas basados en consultas.
- Sistemas basados en directorios dinámicos.
- Sistemas de preguntas-respuestas (*Question-Answering systems*).
- Búsquedas conceptuales.

Todas estas categorías, podrían ser a su vez subdivididas en función de si en su realización se han considerado tecnologías típicas de *Soft-computing* como Lógica Borrosa y técnicas de Inteligencia Artificial.

Las posibilidades de estas técnicas en el campo de la recuperación de la información están claramente constatadas en numerosos estudios, como por ejemplo:

- F. Crestani, G. Pasi, Soft computing in information retrieval: Techniques and applications, Physica Verlag, Series studies in fuzziness, (2000).
- M. Kobayashi, K. Takeda, Information retrieval on the web, ACM Computing Surveys 32(2), (2000) 144-173.
- G. Pasi, Flexible information retrieval: some research trends, Mathware and Soft Computing 9, (2002) 107-121.
- M. Nikravesh, Fuzzy conceptual-based search engine using conceptual semantic indexing, Proceedings of the 2002 NAFIPS annual meeting, (2002) 146-151.

- E. Herrera-Viedma, G. Pasi, Fuzzy approaches to access information on the Web: recent developments and research trends, Proceedings of the Third Conference of the EUSFLAT, (2003), 25-31.

Por último señalar que el propio profesor Zadeh, en el seminario de BISC (*Berkeley Initiative in Soft-Computing*) resalta la necesidad y actualidad de la línea propuesta.

"Title: Web Intelligence, World Knowledge and Fuzzy Logic
Lotfi A. Zadeh
BISC Program
EECS- UC Berkeley
Date: September 14; Tuesday
Time: 4:00-5:30pm
Location: 405 Soda Hall

Existing search engines—with Google at the top—have many remarkable capabilities; but what is not among them is deduction capability—the capability to synthesize an answer to a query from bodies of information which reside in various parts of the knowledge base.

In recent years, impressive progress has been made in enhancing performance of search engines through the use of methods based on bivalent logic and bivalent-logic-based probability theory. But can such methods be used to add nontrivial deduction capability to search engines, that is, to upgrade search engines to question-answering systems? A view which is articulated in this note is that the answer is`No´.

The problem is rooted in the nature of world knowledge, the kind of knowledge that humans acquire through experience and education.

It is widely recognized that world knowledge plays an essential role in assessment of relevance, summarization, search and deduction. But a basic issue which is not addressed is that much of world knowledge is perception-based, e.g., "it is hard to find parking in Paris," "most professors are not rich," and "it is unlikely to rain in midsummer in San Francisco." The problem is that (a) perception-based information is intrinsically fuzzy; and (b) bivalent logic is intrinsically unsuited to deal with fuzziness and partial truth.

To come to grips with the fuzziness of world knowledge, new tools are needed. The principal new tool—a tool which is briefly described in their note—is Precisiated Natural Language (PNL). PNL is based on fuzzy logic and has the capability to deal with partiality of certainty, partiality of possibility and partiality of truth. These are the capabilities that are needed to be able to draw on world knowledge for assessment of relevance, and for summarization, search and deduction."

El Uso de Información Lingüística Borrosa en la RI

Los SRID difusos basados en información lingüística difusa (Bolc y otros, 1985; Bordogna, Pasi, 1993; 1995; Kraft y otros, 1994) son diseñados usando el concepto de *variable lingüística* (Zadeh, 1975) para representar mejor la información cualitativa en el subsitema de consultas. Estos SRID cuentan con lenguajes de consultas ponderados lingüísticos que mejoran la

interacción SRID-usuario. Estos lenguajes de consulta, por un lado, incrementan las posibilidades de expresión de los usuarios porque con ellos es posible asignar pesos a los términos de las consultas indicando importancia relativa o umbrales de satisfacción, y por otro, facilitan a los usuarios la expresión de sus necesidades de información porque pueden expresar los pesos mediante valores lingüísticos más propios del lenguaje humano. Algunos miembros del equipo han propuesto diferentes modelos de SRID lingüísticos usando una aproximación lingüística difusa ordinal que facilita la expresión y el procesamiento de los pesos de las consultas (Herrera-Viedma 2000a;2000b;2001a; 2001b; 2002).

La principal limitación de los anteriores SRID lingüísticos es que cuentan con pocos medios para que el usuario pueda expresar sus necesidades de información o su idea del concepto de relevancia, i.e., solamente presentan la ponderación lingüística de los términos. Ellos no contemplan la posibilidad de ponderar los distintos elementos de una consulta (términos, sub-expresiones, conectivos Booleanos y la consulta completa) simultáneamente, lo cual podría contribuir a incrementar las posibilidades de expresión de los usuarios.

AG y su Uso en RI para la Retroalimentación por Relevancia:

Los AG son algoritmos de búsqueda de propósito general que se basan en principios inspirados en la genética de las poblaciones naturales para llevar a cabo un proceso evolutivo sobre soluciones de problemas. Fueron propuestos por Holland (Holland, 1975) y constituyen una herramienta óptima para proporcionar una búsqueda robusta en espacios complejos, siendo muy válidos para solucionar problemas de optimización que requieran una búsqueda eficiente y eficaz. Los AG son una de las técnicas de SC que más se han aplicado en RI, tal y como lo demuestra el gran número de publicaciones aparecidas recientemente en la literatura especializada (Chen, 1995; Chen y otros, 1998; Gordon, 1991; Horng, Yeh, 2000; Kraft y otros, 1997; Martín-Bautista, 1999; Robertson, Willet, 1994, 1996; Sanchez y otros, 1995; Smith, Smith, 1997; Vrajitoru, 1998; Yang, Korfhage, 1994).

En este trabajo se está interesado en el uso de los AG en RI para la retroalimentación por relevancia, como una técnica útil para la optimización de consultas, que puede contribuir a mejorar la exhaustividad y precisión de los SRID. La retroalimentación por relevancia con AG se ha aplicado principalmente en SRID vectoriales y Booleanos. En los primeros, para aprender los términos y los pesos de los términos en las consultas (Chen, 1995; Chen y otros, 1998; Horng, Yeh, 2000; Robertson, Willet, 1996; Yang, Korfhage, 1994) y en los segundos, para aprender los términos, los pesos de los términos, y los operadores Booleanos (Kraft y otros, 1997; Sanchez y otros, 1995; Smith, Smith, 1997). En general, muchas de las técnicas de retroalimentación por relevancia con AG ven disminuido su rendimiento por alguna de las siguientes causas: simplicidad del AG, uso inapropiado del AG, y mala definición de la función de adaptación y de los operadores genéticos.

Sistemas de recuperación de información

Un sistema de recuperación de información se define como el proceso que trata la representación, almacenamiento, organización y acceso de elementos de información [Salton, 83]. Es decir, es un sistema capaz de almacenar, recuperar y mantener información [Kowalsky, 97].

Pero, ¿qué representa el concepto de información?. En este contexto, información puede ser cualquier elemento apto para su recuperación, como por ejemplo texto (incluidos números y fechas), imágenes, audio, video y otros objetos multimedia [Kowalsky, 97]. El principal tipo de objeto recuperable ha sido siempre el texto, debido a su facilidad de tratamiento en comparación con los objetos multimedia, pero recientemente están surgiendo sistemas capaces de recuperar otros tipos de objetos en Internet. De hecho, Google[2] ha incluido un sistema de recuperación de imágenes en su buscador.

Existen propuestas formales, como la propuesta por Baeza-Yates [Baeza-Yates, 99], que define un modelo de sistema de recuperación de información como una cuádrupla [D, Q, F, $R(q_i, d_j)$] donde:

- D es un conjunto de vistas lógicas (o representaciones) de los documentos que forman la colección.
- Q es un conjunto compuesto por vistas lógicas (o representaciones) de las necesidades de información de los usuarios. Estas vistas se denominan consultas (queries).
- F es una forma de modelar la representación de los documentos, consultas y sus relaciones.
- $R(q_i, d_j)$ es una función de evaluación que asigna un número real al par formado por una consulta q_i ? Q y la representación de un documento d_j ? D. Este valor determinará el orden de aparición de los documentos de una consulta q_i.

Evidentemente la forma de modelar la representación influye de forma notable en el resto de elementos que componen el sistema de recuperación de información. Algunos ejemplos clásicos son el modelo boleano y el modelo vectorial, entre otros. Esta definición está claramente orientada a información textual, donde los documentos serán páginas web u otros objetos (como imágenes) representados en forma textual.

Recientemente se ha comenzado también a tener en cuenta al usuario, debido a la aparición de Internet y a las nuevas posibilidades que plantea. Como se verá posteriormente, la forma tradicional de determinar la satisfacción del usuario con una búsqueda va unida al concepto de relevancia. Pero este concepto está inherentemente influido por la subjetividad del usuario. Según Karen Spark Jones, la recuperación trata con dos elementos "inaccesibles", la necesidad de información que el usuario tiene y el contenido de la información que el sistema le proporciona, y la relación entre ambas, reflejada en el grado en el que un documento es relevante al usuario [Spark Jones, 99]. Este planteamiento centrado en la figura del usuario ha fomentado la aparición de nuevos sistemas de recuperación de información. Este tipo de sistemas intentan conocer los intereses de sus usuarios para adelantarse a sus necesidades de información. Otra tendencia es la aparición de sistemas colaborativos para la recuperación de información.

Recuperación de información vs recuperación de datos.

La principal diferencia radica en la utilización en los sistemas de recuperación de datos de un lenguaje con una estructura y semántica precisas. Por tanto, un sistema de recuperación de datos intenta recuperar todos los objetos que satisfacen claramente unas condiciones definidas expresadas mediante una expresión regular o una expresión del álgebra relacional [Baeza-Yates, 99]. Al estar los datos perfectamente estructurados y definidos por una serie de atributos (como

[2] http://www.google.com]

sucede en las bases de datos relacionales) es posible utilizar un lenguaje de definición de lo que se quiere recuperar, como es SQL (Structured Query Language).

Sin embargo, para un sistema de recuperación de información, el objeto recuperado puede no adaptarse de forma exacta a las peticiones de búsqueda. La razón fundamental es que la información que gestiona un sistema de recuperación de información está en lenguaje natural, sin estructurar, por lo que puede ser semánticamente ambigua. Este es uno de los principales problemas que presentan, ya que tratan de interpretar la relevancia de un documento ante una consulta.

Proceso de recuperación de información.

El proceso de recuperación de información desde el punto de vista del usuario consiste en realizar una pregunta al sistema y obtener un conjunto de documentos. Pero es necesario una serie de procesos previos y diferenciados para su correcto funcionamiento. Existen una serie de procesos necesarios en todo sistema de recuperación de información, que son los siguientes:

1. Proceso de indexación
2. Proceso de consulta
3. Proceso de evaluación
4. Proceso de retroalimentación del usuario

Estos procesos son los clásicos en todo sistema de recuperación de información. Esta clasificación no está cerrada, sino que dependiendo del sistema de recuperación y de sus mecanismos de funcionamiento se pueden distinguir nuevos procesos. Los sistemas que utilicen estructuras de conocimiento adicionales a los índices de términos clásicos suelen tener procesos adicionales encargados de construir, mantener o actualizar dichas estructuras. Por ejemplo, en los sistemas de recuperación de información basados en la utilización de perfiles de usuario se puede distinguir un nuevo proceso denominado proceso de construcción y actualización del perfil de usuario. Este proceso almacena los términos que representan las preferencias de los usuarios con la finalidad de mejorar el comportamiento del sistema en futuras consultas.

Para que el proceso de recuperación pueda realizarse es necesario que exista una base de datos que contenga la información de los objetos que el sistema es capaz de recuperar. Si un objeto no esta en esta base de datos no podrá ser recuperada. Esta base de datos contiene una representación de los objetos recuperables, y no los objetos en sí mismos. Los objetos se representan en algún mecanismo que permita su modelado, típicamente indexando un conjunto de términos relevantes del mismo, que puede ser un subconjunto o la totalidad de los términos del documento. Los SRI poseen mecanismos que permiten introducir un nuevo objeto en la base de datos, o bien utilizan mecanismos automáticos que se encargan de explorar el espacio de búsqueda (típicamente Internet) introduciendo los nuevos documentos.

Proceso de indexación.

Los algoritmos suelen seleccionar como documentos relevantes aquellos que poseen los términos que forman la consulta del usuario. Por tanto, es necesaria alguna forma de representación que permita determinar la existencia o no de estos términos en el documento.

Una primera posibilidad consiste en almacenar los objetos y buscar en cada uno de ellos la existencia o no de los términos. Este planteamiento hoy por hoy es inviable debido a la enorme cantidad de información que es necesario manejar para convertirse en un sistema eficaz. Por

tanto, es necesaria la utilización de unas estructuras que almacenen información sobre los objetos y que permitan acelerar las búsquedas. Estas estructuras se denominan índices.

Normalmente, los documentos suelen ser preprocesados antes de ser indexados para reducir el número de términos y por tanto mejorar la eficiencia de la recuperación. Por el contrario, se pierde información sobre los documentos lo que conlleva una serie de limitaciones a la hora de recuperarlo.

Preprocesado de documentos.

Los algoritmos clásicos utilizados en el paso previo a la indexación son los siguientes:

1) *Eliminación de signos de puntuación*: se suelen eliminar los acentos, comas, puntos y demás signos de puntuación para tratar los términos de forma uniforme. Este proceso tiene el inconveniente de que se pierde esta información y no se podrán utilizar signos de puntuación en las consultas de los usuarios.

2) *Eliminación de palabras prohibidas (stop words):* todos los idiomas tienen un conjunto de palabras de frecuente aparición que se utilizan para garantizar la concordancia sintáctica de las frases. Estas palabras no aportan ningún significado a un documento, sino que solo se utilizan para seguir las reglas del idioma. Este es el caso de las preposiciones, conjunciones, determinantes, etc. A la hora de indexar, suele existir una lista que contiene estas palabras, denominada stoplist, y que sirve como referencia para excluir palabras a la hora de indexar.

3) *Stemming o lematización:* Los algoritmos de lematización consisten en obtener la raíz de una palabra, denominada raíz o stem, ignorando a la hora de indexar las múltiples variaciones morfológicas que puede tener. En la mayoría de los casos la raíz será una palabra sin significado, por ejemplo, la raíz de "plaza" y "plazoleta" sería "plaz". Normalmente se suele aplicar para eliminar el sufijo de una palabra, sin aplicarse al prefijo. Las premisas en las que se basa son que la raíz relaciona el significado del concepto con la palabra, y que los sufijos introducen ligeras modificaciones del concepto o que se utilizan para propósitos sintácticos. El objetivo original de la lematización fue mejorar el rendimiento y reducir el número de palabras que un sistema tenía que almacenar para consumir menos recursos del sistema.
   Otra de sus características es que permite aumentar el recall de la recuperación a costa de una reducción en la precisión. Es decir, se recuperan palabras relacionadas conceptualmente al tener la misma raíz y por tanto se obtiene un conjunto más rico de términos, evitándose así perder términos potencialmente relevantes. Pero este fenómeno conlleva riesgos en la precisión de la recuperación, debido a que los lenguajes naturales no son regulares en sus construcciones. Es posible que se indexen juntas palabras no relacionadas conceptualmente al tener la misma raíz, y por tanto no sea posible diferenciarlas a la hora de la recuperación. Por ejemplo, podría suceder que se indexen bajo la raíz "pec" los términos "pecado" y "peces". Este problema se denomina sobrelematización. También es posible que el algoritmo de lematización falle y obtenga raíces distintas para dos palabras conceptualmente similares. A esta situación se la denomina bajolematización. Este caso se podría dar por ejemplo con dos variaciones del verbo "tener", obteniendo raíces distintas ("tene" y "teng") para las palabras "tenemos" y "tengo", indexándose bajo raíces distintas.
   Otro problema es que este método es dependiente del idioma, y por lo tanto sería necesario a la hora indexar utilizar un algoritmo específico para cada idioma. Esta situación lleva asociado la utilización de un algoritmo para determinar el idioma. Además, este tipo de

algoritmos funcionan bien con idiomas que tengan una sintaxis no excesivamente complicada, como el inglés, pero en cambio fallan mucho más con otro tipo de idiomas más complejos, como el castellano. Por tanto, la lematización difiere dependiendo de los distintos idiomas.

Hay diferentes técnicas utilizadas en este tipo de algoritmos, destacando la utilización de reglas y diccionarios. Existen multitud de algoritmos de lematización basados reglas, la mayoría de ellos en inglés, de los cuales el más sencillo es el lematizador S [Hull, 96], que se limita a remover las terminaciones plurales. El algoritmo de lematización más famoso es el algoritmo de Porter [Porter, 80], y elimina cerca de 60 terminaciones en cinco etapas. En cada paso se elimina un tipo concreto de terminación corta eliminándolo sin más o transformando la raíz. También cabe destacar los algoritmos de Lovins [Lovins, 68] y Paice [Paice, 90]. Los algoritmos basados en diccionario, como KSTEM [Krovetz, 93], intentan eliminar los errores anteriormente descritos aumentan. Se estima que se puede aplicar la lematización correctamente al 40% de las palabras, al 20% no se puede ("universidad" y "universo" por ejemplo), y en el resto de los casos la exactitud de la lematización depende del contexto.

La efectividad de la lematización ha sido discutida y no existe un consenso sobre la misma, habiendo diferentes estudios en contra de su utilización, mientras que otros autores como Krovetz [Krovetz, 93] afirman que esta técnica mejora el recall e incluso la precisión cuando los documentos y consultas son cortas.

Por último es necesario nombrar los n-gramas. Los n-gramas ignoran el aspecto semántico de las palabras y son algoritmos que se basan en que dos palabras relacionadas semánticamente suelen contener los mismos caracteres.

4) *Eliminación de documentos duplicados:* Se estima que en Internet existen muchas páginas Web duplicadas, aproximadamente el 30%. La eliminación de documentos duplicados permite mejorar el rendimiento y reducir espacio de almacenamiento. Pero la tarea de identificar documentos similares no es trivial, ya que pueden darse diferentes situaciones que compliquen esta labor, como por ejemplo, el formato del documento. Dos documentos pueden ser idénticos en contenido pero estar en diferentes formatos (html, postcript, pdf o word).

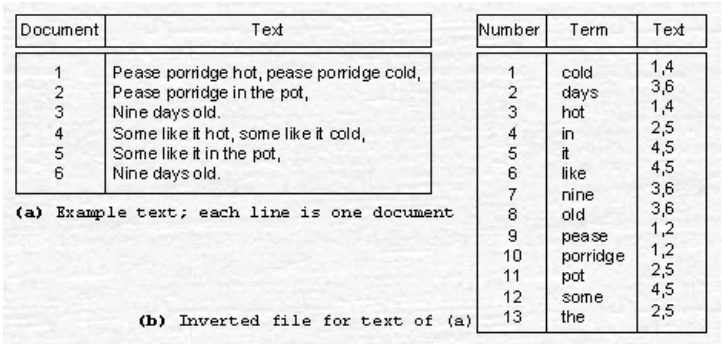
   Una de las posibles formas de detectar la similitud de los documentos consiste en convertir todos los documentos a un mismo formato, normalmente texto plano, utilizando alguna herramienta de conversión estándar. Cada documento se divide en una colección de partes o trozos formados por pequeñas unidades de texto (por ejemplo líneas o sentencias). Después, a cada trozo se le aplica una función *hash* para obtener un identificador único. Si dos documentos comparten un número de trozos con igual identificador por encima de un umbral T, entonces se consideran documentos similares.

Estructuras de indexación clásicas.

En los primeros sistemas, los índices se limitaban a contener un conjunto de palabras clave representativos del documento, pero actualmente se utiliza un mayor número de términos del documento. Normalmente en la indexación no se utilizan todos los términos (aunque hay excepciones), sino que se suelen utilizar un subconjunto de términos y se almacena aparte el documento completo en repositorios o caches si es posible o simplemente almacenar su ubicación, normalmente la URL (*Universal Resource Locator*) de algún documento de Internet.

La estructura clásica utilizada en la indexación de documentos es el archivo invertido, formada por dos componentes: el vocabulario y las ocurrencias (véase Figura 2.1). El vocabulario es el

conjunto de todas las palabras diferentes del texto. Para cada una de las palabras del vocabulario se crea una lista donde se almacenan las apariciones de cada palabra en un documento. El conjunto de todas estas listas se llama ocurrencias [Baeza Yates, 99]. Este mecanismo no es el único, sino que existen otros muchos como los ficheros de firmas, basados en técnicas *hash*, árboles PAT y grafos.



Document	Text
1	Pease porridge hot, pease porridge cold,
2	Pease porridge in the pot,
3	Nine days old.
4	Some like it hot, some like it cold,
5	Some like it in the pot,
6	Nine days old.

(a) Example text; each line is one document

Number	Term	Text
1	cold	1,4
2	days	3,6
3	hot	1,4
4	in	2,5
5	it	4,5
6	like	4,5
7	nine	3,6
8	old	3,6
9	pease	1,2
10	porridge	1,2
11	pot	2,5
12	some	4,5
13	the	2,5

(b) Inverted file for text of (a)

Ejemplo de fichero invertido (Fuente: [Frakew, 92])

Herramientas de búsqueda en Internet.

Hoy en día en Internet existen tres tipos principales de herramientas utilizadas en la búsqueda web: los directorios, los buscadores y los meta-buscadores. El cuarto tipo de herramienta, los agentes de búsqueda, depende para su viabilidad de la implantación de la Web Semántica.

Directorios.

Los directorios o índices temáticos son listados de recursos organizados según una jerarquía de temas. La jerarquía sigue una estructura de árbol, de forma que vaya desde las categorías más generales hacia categorías más específicas conforme bajamos en la estructura. Tradicionalmente, los documentos que forman parte de este tipo de sistemas son clasificados por indexadores humanos o por los propios autores de la página.

Recientemente, están comenzando a aparecer algoritmos automáticos de clasificación que realizan esta tarea de forma automática. Por ejemplo, Kim [Kim, 03] propone un algoritmo que utiliza la lógica difusa para obtener a partir de una colección de documentos una jerarquía. Otros mecanismos de categorización automática son TAPER (*a Taxonomy And Path Enhanced Retrieval system*) desarrollado por IBM. Este algoritmo construye una taxonomía en forma de árbol formada por términos que sean buenos discriminantes de la temática del documento, agrupados por clases. Posteriormente, se evalúa cada documento para obtener los términos discriminantes y posteriormente se someten a un proceso de evaluación para determinar a que clase corresponde el documento. Una vez determinada la clase, ese documento se asocia con el término que describe la clase. Este algoritmo se mejoró mediante la utilización de la información que proporcionan los enlaces de cada documento.

Los directorios además suelen permitir también búsquedas por palabras clave. La ventaja de este mecanismo reside en que se pueden restringir las búsquedas a categorías particularmente relevantes, pudiéndose de esta forma mejorar la relevancia de los documentos obtenidos.

Otro intento de automatizar la clasificación de los documentos dentro de un índice temático es el sistema OpenGrid [Lifantsev, 98]. OpenGrid utiliza las opiniones y comentarios de cientos de navegantes *web* sobre las páginas para clasificar los documentos de la Web. Para ello los creadores de las páginas *web* introducen en el enlace información acerca de la categoría y ranking del documento. De esta forma, juntando toda la información referente a esa página, se puede obtener de forma distribuida una categorización y ranking bastante aproximado.

<A ref.=http://www.algunapágina.foo/ cat="/News/Computers" rank ="80%"> Buenas noticias de computadores </A>

Evidentemente, OpenGrid solo es una propuesta, ya que necesita que los creadores de las páginas *web* se pongan de acuerdo para incluir esta información en los enlaces que introducen en sus páginas.

Los directorios más conocidos son Yahoo[3] y el Open Directory Project DMOZ[4].

Buscadores.

Los buscadores o motores de búsqueda son sistemas que indexan los documentos de Internet sin seguir una estructura jerárquica como hacen los directorios. Este tipo de sistemas poseen unos programas especializados en recorrer la *web* de forma automática denominados *crawlers* (también llamados *robots*, *spiders*, *wanderers*, *walkers* o *knowbots*), que indexan los documentos que no contiene su base de datos. Normalmente este tipo de sistemas cubre un mayor número de documentos que los directorios debido al proceso de automatización de indexación. Además, suelen estar mejor actualizados que los directorios debido a que cada cierto tiempo se comprueba si el documento referenciado no ha sufrido modificaciones. La forma de buscar documentos en este tipo de sistemas consiste en realizar una consulta introduciendo un conjunto de términos relacionados con lo que el usuario busca.

El mayor inconveniente que presenta este tipo de herramientas son los denominados problemas de Precisión y *Recall.* Lawrence y Giles [Lawrence, 99], en su trabajo de evaluación de los buscadores, identificaron 5 problemas principales que con frecuencia presentan:

- *La cobertura de los buscadores decrece*: Los sistemas de indexación no son capaces de contemplar el rápido crecimiento de la Web, quedándose una gran parte de ella fuera de los índices de los principales buscadores, y por tanto, no es accesible para los usuarios que utilicen los buscadores.
- *Acceso desigual*: Existen tendencias a la hora de indexar las páginas debido al rastreo de los buscadores en busca de enlaces a otras páginas que indexar. Por ejemplo, será más probable que se indexen páginas que reciben muchos enlaces de otras páginas (sitios populares). Así mismo, es más probable que se indexen sitios comerciales en vez de sitios educacionales.
- *Enlaces rotos*: los *crawlers* verifican cada cierto tiempo si la página que indexan no se ha movido. Pero esta comprobación, debido al elevado número de documentos que indexan, requiere de un cierto tiempo, durante el cual pueden producirse inconsistencias debido a que una página se cambie de sitio o que desaparezca.
- *Baja utilización de meta datos*: muchos buscadores utilizan los metadatos definidos en la página web como fuertes indicadores a la hora de indexar. Sin embargo, muy pocas páginas utilizan los meta-tags de HTML que cubren esta función como los tags "Keywords" y

[3] http://www.yahoo.com
[4] http://dmoz.org

"description" (solo un 34% según Lawrence y Giles). Existen otros mecanismos que incorporan un conjunto de meta-etiquetas estándar en las páginas web que ayudan a determinar distintos parámetros del documento, como por ejemplo Dublín Core, pero su utilización es todavía muy baja.

- *Distribución de la información*: hay una gran variedad de información en la Web y además su distribución es desigual. Por ejemplo, existe un mayor número de sitios de carácter comercial frente a los de carácter educativo o científico.

Las características que los usuarios mejor valoran de este tipo de herramientas son las siguientes:

- Fácil de utilizar
- Carga y repuesta rápidas
- Fiabilidad y precisión de los resultados
- Información organizada y actualizada

Operadores de búsqueda.

La facilidad de uso es uno de los principales factores a la hora de utilizar un buscador. Sin embargo, este tipo de sistemas suele incorporar una serie de opciones avanzadas que permiten especificar distintos filtros aplicados a los resultados. Estas características se suelen implementar mediante formularios de búsqueda avanzados donde se indican varias opciones, o mediante operadores de búsqueda. Los operadores de búsqueda clásicos son:

- ***Operadores lógicos***: Este tipo de operadores son los operadores clásicos boléanos. Para simplificar, y debido a la dificultad que implica la utilización de estos operadores para usuarios noveles, se suelen indicar pantallas de opciones avanzadas con las opciones 'incluir todos los términos' en vez del operador AND, 'incluir alguno de los términos' en lugar del operador OR y 'excluir el termino' en vez del operador NOT. Suelen utilizarse los siguientes:
  - AND, "+" o "&": indica que la página debe contener obligatoriamente los términos que están unidos por este operador.
  - OR: devuelve los documentos que contienen al menos uno de los dos términos que une el operador.
  - XOR: este operador es menos común, e indica al buscador que devuelva los documentos que tienen uno de los términos, pero que no tienen los dos.
  - NOT, "-" o "!": este operador a diferencia de los anteriores es unario, se aplica a un solo término y no implica dos términos como los operadores binarios anteriores. Excluye los documentos que tienen el término al que se refiere el operador.

- ***Operadores posicionales o de proximidad***: afectan a la posición de los términos en el documento y las relaciones de las palabras de la consulta atendiendo normalmente a criterios de proximidad u orden. Los más usuales son:
  - NEAR, "~", "[]": Se sitúa entre dos términos de la consulta para indicar que recupere los documentos que contengan ambos términos, pero que no estén separadas por un número determinado de palabras. Este número oscila entre 25 palabras o 100 caracteres, aunque a veces este número es configurable.
  - FAR: es el operador contrario a NEAR y recupera documentos en las que debe haber una distancia mínima entre los términos.

- ADJ: Se utiliza aplicado a dos términos y recupera solo los documentos que poseen los dos términos y además están juntos en el documento. El orden no se tiene en cuenta.
- FOLLOWED BY: Es un operador parecido a NEAR pero define muy claramente el cual debe ser el orden de los términos.
- BEFORE: Funcionamiento parecido al operador AND, pero teniendo en cuenta el orden de aparición en el documento.

Existen distintas variantes o modificaciones que se les pueden aplicar a estos operadores dependiendo de las características propias del lenguaje de consulta. Por ejemplo, existen modificadores de orden para los operadores ADJ, NEAR y FAR, que consiste en añadir delante del operador la letra O de orden (OADJ, ONEAR y OFAR), de esta forma si se utiliza por ejemplo coches OADJ carreras sólo recuperará los documentos referidos a coches de carreras y no a carreras de coches. Otro modificar que afecta a NEAR y FAR es el de la distancia entre palabras, que se puede indicar mediante el parámetro "/", por ejemplo NEAR/3 quiere decir que la máxima diferencia en palabras entre los términos es 3 palabras. También se puede utilizar esta característica con ADJ, para indicar el número exacto de palabras que debe haber entre los dos términos. No obstante, existen algunos otros operadores dependiendo del buscador, como por ejemplo A WITHIN 10 BEFORE B que indica que entre los términos A y B no debe haber más de 24 caracteres. Otro operador sería WITH o SENT que indica que dos términos deben aparecer en la misma sentencia, etc.

- ***Operadores de exactitud***: este tipo de operadores se utiliza para indicar que lo que se busca es literal, y normalmente suele implementarse mediante la utilización de las comillas '"'. Sin embargo, existen otros operadores como PHRASE que se utiliza de forma similar. Este tipo de operadores son muy útiles a la hora de buscar información concreta como por ejemplo la búsqueda de artículos por título, si se conoce.

- ***Operadores de truncamiento***: se aplican una serie de caracteres especiales en la consulta que se utilizan como comodines. Este tipo de operadores son muy útiles cuando queremos recuperar no sólo los documentos con un término concreto, sino también queremos recuperar los documentos que posean términos con variaciones morfológicas de ese texto. Normalmente se suelen utilizar como operadores los caracteres "*", "?" o "$".

- ***Operadores de campo***: Muchos buscadores permiten especificar mediante palabras clave concreta distintos comportamientos o filtros a la hora de realizar la búsqueda. Por ejemplo, se permite indicar los términos del título (normalmente mediante una etiqueta del estilo title), el dominio, si tiene enlaces a una web concreta, con restricciones de tamaño, fecha, tipo, etc.

Básicamente estos son los operadores principales. La sintaxis del operador depende de la implementación concreta del buscador. Así mismo, hay que nombrar también otros operadores importante como son los paréntesis, utilizados para agrupar términos y operadores, y otro curioso como es el operador tesauro "@" que reemplaza el término por alguno similar extraído de un tesauro (un sinónimo).

Arquitectura de los buscadores.

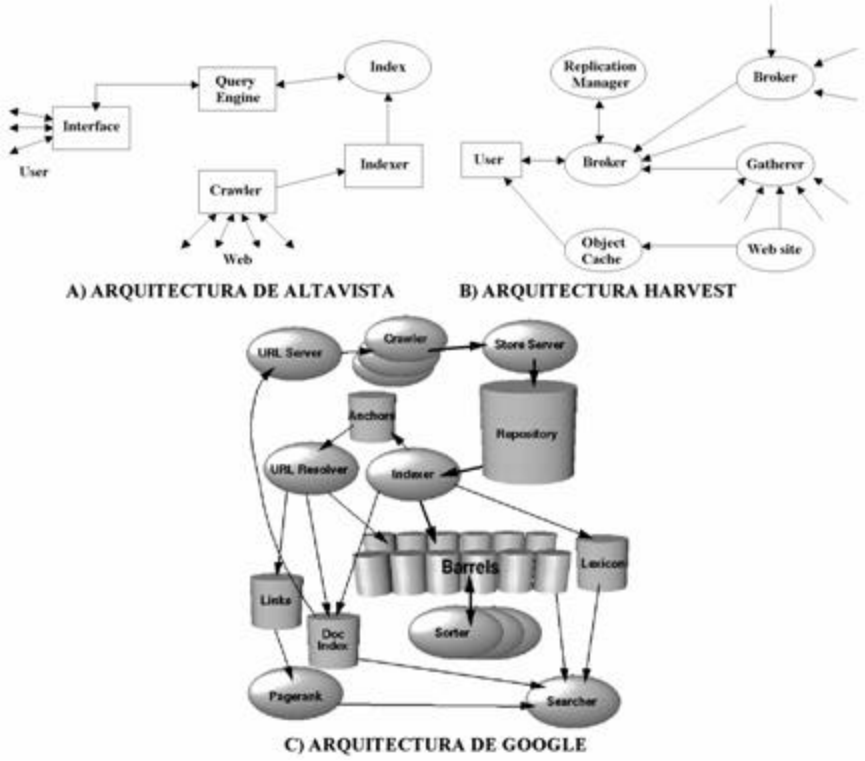
Se pueden distinguir dos tipos de buscadores: de propósito general y de propósito especial. Los buscadores de propósito general abarcan páginas de toda la web sin centrarse en ningún dominio específico. Los buscadores de este tipo más conocidos son *Google*, *Altavista*, *Excite*, *Lycos* y *HotBot*. Por el contrario, los buscadores de propósito especial se especializan en documentos

pertenecientes a un dominio concreto como puede ser una temática determinada o los documentos de una organización. Algunos ejemplos de este tipo de buscadores son *Citeseer*[5] que se centra en artículos de investigación y *Medical World Search*[6] que se centra en información sobre medicina.

Las distintas herramientas utilizadas para buscar en la Web presentan diferentes arquitecturas dependiendo de las características propias. Sin embargo, suelen existir un conjunto de componentes básicos. No obstante, las implementaciones de la mayoría de buscadores comerciales no están disponibles al público, a no ser que sean de carácter experimental. Afortunadamente, existen algunas excepciones que a continuación se describen.

Arquitectura de *Altavista.*

Un buen representante de la arquitectura típica de un buscador en la *Web* es la arquitectura de *Altavista* (véase Figura). Baeza-Yates [Baeza-Yates, 99] describe la arquitectura de *Altavista* como una arquitectura centralizada. Se le aplica este calificativo debido a que utiliza un proceso de rastreo e indexación de la web centralizado. Este rastreo de la *web* en busca de páginas web lo realiza el componente denominado *Crawler*, el cual se encarga de enviar peticiones a los servidores Web remotos para obtener las páginas Web. Una vez que las tiene las examina en busca de nuevos enlaces que recorre. El resultado del proceso de rastreo es un listado de las direcciones de las páginas *web* que el *Crawler* ha encontrado. Este listado se le pasa al indexador que es el encargado de almacenar en el índice los términos relevantes del documento.



[5] http://citeseer.ist.psu.edu
[6] http://www.mwsearch.com

Arquitectura de varios buscadores Web: Altavista, Harvest y Google (Fuente: [Lam, 01])

Altavista se considera formado por dos partes:

1. Una encargada de interactuar con el usuario. Su función es recibir las peticiones de los usuarios y de resolverlas. Esta formada por la Interfaz de Usuario y el motor de consultas.
2. Otra parte encargada del proceso de rastreo e indexación de los documentos Web. Esta parte la componen el indexador y el crawler.

Esta arquitectura presenta un par de problemas. El primero es consecuencia de la utilización de un único componente encargado de recoger las peticiones de los usuarios, lo que provoca la saturación de las líneas de comunicación, así como una sobrecarga en los servidores. El segundo problema radica la centralización del proceso de rastreo e indexación. Este planteamiento tiene problemas para manejar el crecimiento de la Web, debido al enorme volumen de datos que tienen que manejar el indexador y el *crawler*.

Arquitectura *Harvest*.

La arquitectura *Harvest* [Bowman, 94] utiliza una arquitectura distribuida para recoger los datos y almacenarlos, que es más eficiente que la arquitectura de *Altavista*. El principal inconveniente es que *Harvest* requiere la coordinación de varios servidores Web. Para tratar de resolver los inconvenientes de *Altavista*, debido a la saturación provocada por los crawlers, esta arquitectura introduce dos elementos principales: recogedores (*gatherers*) e intermediarios (*brokers*) como puede verse en la figura 2.2. El recogedor se encarga de recopilar y extraer información de indexado de uno o más servidores Web de forma periódica. Los intermediarios aportan el mecanismo de indexado y la interfaz de consulta a los datos recopilados. Los recogedores e intermediarios se comunican entre sí para intercambiar información de indexado, así como para balancear la carga en el tráfico de la red [Baeza-Yates, 99].

Arquitectura de *Google*.

El nombre de Google viene de la palabra googol, que significa 10^{100}. Las principales razones del éxito actual de Google son su enorme base de datos (la actual hoy en día) que indexa millones de documentos de la web, su algoritmo de evaluación basado en la estructura de la Web (*PageRank*) y su eficiencia al estar principalmente implementado en C/C++ sobre plataformas *Solaris* o *Linux*.

Google utiliza varios *Crawler* distribuidos para descargar las páginas *Web*. La lista de páginas que hay que recuperar se la proporciona el servidor URL. Las páginas *web* recuperadas se pasan al Servidor de almacenamiento, que se encarga de comprimir y almacenar las páginas recibidas de los Crawlers en un repositorio.

Una vez almacenadas las páginas en el repositorio, tienen que ser indexadas. Esta labor la realiza el indexador, que obtiene los documentos del repositorio descomprimiéndolos y analizándolos. Extraer de cada documento las ocurrencias de cada palabra y construye un registro de hits, que en definitiva son las ocurrencias de cada palabra. Cada registro de *hits* contiene la palabra, la posición en el documento, una aproximación del tamaño de la fuente y la capitalización. A continuación el indexador distribuye el registro entre un conjunto de “barriles”

y crea un índice parcial ordenado. Además, el indexador extrae los enlaces de cada página junto con el texto asociado y los almacena en el fichero de anclas.

A continuación el componente URL Resolver lee este fichero de anclas y extrae los enlaces, convirtiendo los relativos a enlaces absolutos. El texto que describe el enlace es asociado con el documento y almacenado en los barriles. Además, genera una base de datos de enlaces compuesta por dos identificadores de dos documentos (origen y destino del enlace). Esta base de datos es la fuente utilizada por el algoritmo *PageRank*.

Por último los clasificadores (*sorters*) generan un índice invertido a partir de los documentos de los barriles, así como también se construye un nuevo lexicón que contemple los nuevos documentos.

La resolución de las búsquedas utiliza el componente *searcher*, que se ejecuta en un servidor Web. Para la evaluación y obtención de los documentos, se utilizan el índice invertido, los barriles, el lexicón y el resultado proporcionado por el algoritmo *PageRank*.

Meta-Buscadores.

Los meta-buscadores son buscadores que no disponen de una base de datos propia que contenga la indexación de los documentos. Proporcionan una interfaz unificada para consultas a diferentes buscadores. Por tanto, simplemente se limitan a recibir las peticiones de los usuarios y enviarlas a otros buscadores. Los resultados que reciben deben ser sometidos a un proceso de clasificación para reunir en un solo listado los documentos devueltos por multitud de buscadores. Suelen ser más lentos que los buscadores debido a que siguen un proceso más complejo y elaborados. El problema de los meta-buscadores consiste en combinar las listas devueltas por otros buscadores de forma que se optimice el rendimiento.

Sin embargo, este tipo de sistemas mejoran algunos de los problemas presentes en los buscadores tradicionales, como el problema del *Recall*, aunque sin embargo todavía sufren el problema de la Precisión [Kerschberg, 01]. Según Kerschberg, la forma de solucionar el problema de la Precisión se aborda utilizando 4 mecanismos principalmente: métodos basados en el contenido, colaborativos, de conocimiento del dominio y basados en ontología. Los métodos basados en el contenido tratan de obtener una representación de las preferencias del usuario lo más concretas posibles para posteriormente mejorar la evaluación de las páginas devueltas basándose en el contenido del documento y las preferencias del usuario. Dentro de esta categoría se pueden nombrar a WebWatcher [Armstrong, 95], WAWA [Shavlik, 98] y WebSail [Chen, 00]. El método colaborativo se basa en la similitud entre los usuarios para determinar la relevancia de la información. Cabe destacar Phoaks [Terveen, 97] y Siteseer [Bollacker, 00]. El método basado en el conocimiento del dominio utiliza la ayuda del usuario y del conocimiento del dominio de la búsqueda para proporcionar una mayor relevancia. Por último, el método basado en ontología establece una jerarquía entre conceptos que permite concretar la búsqueda y mejorarla. De este tipo merece la pena nombrar a WebSifter II [Kerschberg, 01], que utiliza una representación en árbol denominada WSTT (Weighted Semantic Taxonomy-Tree) para representar las intenciones de búsqueda de los usuarios. También entran en esta categoría OntoSeek [Guarino, 99], On2Broker [Fensel, 99] y WebKB [Martin, 00].

Las principales ventajas de los meta-buscadores expuestas por Meng son cuatro:

- **Incrementa la cobertura de la búsqueda en la Web**. Debido a la enorme cantidad de documentos que contiene Internet, es imposible que un solo buscador indexe la totalidad de la Web. Por tanto, mediante la combinación de distintos buscadores se consigue cubrir un mayor número de documentos en las búsquedas.
- **Soluciona la escalabilidad de la búsqueda en la Web**.
- **Facilita la invocación de múltiples buscadores**. Permite mediante la utilización de una sola consulta obtener los documentos más relevantes indexados por múltiples buscadores, lo que evita al usuario buscar en cada uno de ellos.
- **Mejora la efectividad de la recuperación**. Al poder consultas a buscadores de propósito especial, permite obtener de ellos un conjunto más relevantes de documentos, sin sufrir la desviación típica que produce el elevado número de documentos que indexan los buscadores de propósito general.

Así mismo, según Aslam [Aslam, 01], también presentan las siguientes ventajas potenciales:

- Mejora el factor recall: al obtener los resultados de múltiples buscadores puede mejorar el número de documentos relevantes recuperados (el factor recall).
- Mejora la precisión: diferentes algoritmos de recuperación recuperan muchos documentos relevantes iguales, pero diferentes documentos irrelevantes. Basándose en este fenómeno, en caso de ser cierta esta teoría, cualquier algoritmo que prime los documentos que aparecen en las primeras posiciones en resultados de distintos buscadores obtendrá una mejora en la recuperación. Este fenómeno se denomina "efecto coro".
- Consistencia: los buscadores actuales responden con frecuencia de forma muy distinta ante la misma consulta transcurrido un tiempo. Si se utilizan distintas fuentes para obtener los resultados, es de esperar que la variabilidad se vea reducida favorecida por los buscadores que proporcionan resultados más estables.
- Arquitectura modular: las técnicas utilizadas en los meta-buscadores pueden descomponerse en módulos pequeños y más especializados que pueden realizarse en paralelo de forma colaborativa.

No todo son ventajas en la utilización de los meta-buscadores. Las principales desventajas comentadas por Meng son:

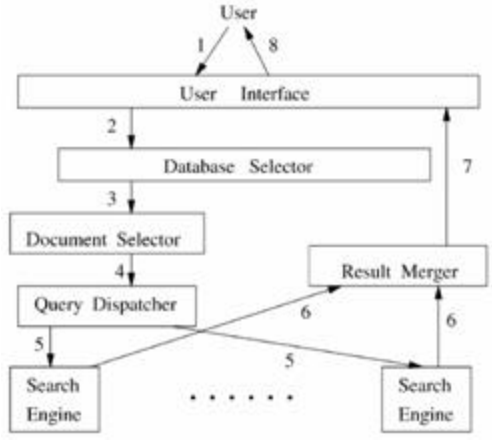
- La selección de la base de datos: este problema está asociado a la selección del buscador que recibirá la consulta. Se trata de seleccionar los buscadores que devuelvan buenos resultados ante una consulta concreta. Por ejemplo, la consulta sobre "fútbol" realizada a un buscador especializado en literatura científica no tendría demasiado sentido. Para tratar de solucionar este inconveniente, Meng propone la utilización de medidas que indiquen la utilidad de cada base de datos con respecto a una consulta dada. Clasifica estos mecanismos en 3 categorías: métodos de representación amplia (*rough representative approaches*), métodos de representación estadística (*Statistical representative approaches*) y métodos basados en el aprendizaje (*Learning-based approaches*).
- La selección de documentos: una vez seleccionado el origen de los documentos el problema consiste en determinar el número apropiado de documentos que hay que solicitar. Si se reciben demasiados documentos el coste computacional para determinar los mejores documentos y el coste de comunicación para obtenerlos puede ser excesivo. Meng también establece una serie de mecanismos que tratan de solucionar este problema divididos en 4 categorías: decisión del usuario (decide el usuario), asignación por peso (se obtienen mayor número de documentos del que se considere el mejor buscador), métodos basados en el aprendizaje (se basa en el pasado para determinar el número de documentos

de cada buscador) y la recuperación garantizada (trata de garantizar la recuperación de todos los documentos potencialmente útiles).

- Fusión de resultados: el problema consiste en fusionar los resultados de distintos buscadores con sus propias características y formas de evaluación en un único listado ordenado por relevancia. Además, existe la posibilidad de que halla documentos repetidos devueltos por distintos buscadores. Se pueden clasificar las técnicas utilizadas para resolver este problema en ajuste por similitud local (se basa en las características del buscador o la similitud devuelta por este) y estimación por similitud global (se evalúa o estima la similitud de cada documento recuperado con la consulta original).

Los metabuscadores tienen una serie de particularidades, al tener que reunir en un solo listado los documentos provenientes de múltiples fuentes, con sus propios criterios de evaluación. Para lograrlo, Gravano [Gravano, 88] atribuye 3 tareas principales a los metabuscadores, no presentes ni en los buscadores tradicionales ni en los directorios. Estas tareas principales son:

- Selección de la base de datos: consiste en elegir los buscadores a los que se les enviará la consulta del usuario.
- Traducción de la consulta: debido a que cada buscador posee un lenguaje de consulta característico, es necesario adaptar cada consulta al lenguaje de consulta del buscador destino.
- Combinación de los resultados: se trata de obtener un único listado de resultados.



Arquitectura de componentes software de un metabuscador (Fuente: [Meng, 02])

Existen multitud de arquitecturas de meta-buscadores propuestas, como la realizada por Li [Li, 01], Kerschberg [Kerschberg, 01] y Glover [Glover, 99], entre otras. Suelen descomponerse en una serie de módulos más o menos específicos. Meng describe una arquitectura de referencia formada por 5 componentes (Figura):

- Interfaz de usuario: Se encarga de obtener la consulta del usuario. En algunos casos puede proporcionar un sistema de refinamiento de la consulta interactivo, basado en la utilización de alguna estructura de conocimiento. Además, es la encargada de mostrar los resultados de la búsqueda.
- Selector del buscador: Trata de seleccionar los buscadores que mejor respuestas darán a la consulta del usuario. Intenta evitar un envío masivo de consultas a todos los buscadores que puede tener asociado un bajo rendimiento y un coste alto en tiempo.

- Selector de documentos: El objetivo es tratar de recuperar el mayor número de documentos relevantes, evitando recuperar documentos no relevantes. Si se recupera un número excesivo de documentos no relevantes influirá de forma negativa en la eficiencia de la búsqueda.
- Expedidor de consulta: Es el responsable de establecer la conexión con el buscador y pasarle la consulta (o consultas), así como obtener los resultados. El protocolo habitual que se suele utilizar en este proceso es http (HyperText Transfer Protocol) mediante la utilización de los métodos GET y POST. No obstante, existen buscadores que facilitan una interfaz de programación (API) para realizar consultas y que utilizan protocolos distintos (Google utiliza en su API el protocolo SOAP).
- Fusionador de resultados: Su función principal es combinar los resultados de los distintos buscadores en un único listado. Es imprescindible para la obtención de unos buenos resultados la utilización de algún criterio de evaluación para establecer un orden en el listado que muestra al usuario.

Algunos de los meta-buscadores más populares de la Web son Vivisimo[7], Mamma[8] y MetaCrawler[9]. En [Sherman, 04] puede verse un listado más exhaustivo de meta-buscadores.

Agentes inteligentes.

Un agente es una entidad software que recoge, filtra y procesa información contenida en la Web, realiza inferencias sobre dicha información e interactúa con el entorno sin necesidad de supervisión o control constante por parte del usuario. Estas tareas son realizadas en representación del usuario o de otro agente. Hay que distinguir los agentes inteligentes de los buscadores inteligentes. Estos últimos, incorpora información semántica en el proceso de búsqueda para mejorar los resultados de la búsqueda, normalmente la precisión, utilizando en la búsqueda los recursos previamente indexados. Sin embargo, un agente inteligente recorre la Web a través de los enlaces entre recursos (hiperdocumentos, ontologías, ...) en busca de aquella información que le sea solicitada, pudiendo además interactuar con el entorno para el cumplimiento de tareas encomendadas. Los agentes pueden realizar funciones de búsqueda, discriminación y selección. Así podemos distinguir los siguientes tipos de agentes [Hellmann, 95]:

- *Agentes vigilantes*: de forma autónoma buscan información específica y pueden utilizarse para elaborar versiones personalizadas de los periódicos según los intereses del lector. Ejemplos: Personal Journal (Dow Jones), JobCenter, Personal View (Ziff Davies).
- *Agentes ayudantes*: actúan sin intervención humana. Se suelen emplear para la gestión de red y para las funciones normales de mantenimiento. Ejemplo: LANAlert.
- *Agentes aprendices*: aprenden a ajustar sus prestaciones al modo de actuar de su usuario. Ejemplos: Firefly[10].
- *Agentes compradores*, capaces de comparar precios y determinar qué producto ofrece las mejores condiciones. Ejemplo: BargainFinder[11].
- *Agentes de recuperación de información*: buscan formas inteligentes de recoger información y son capaces de reducir la sobrecarga de información en la localización de

[7] http://vivisimo.com
[8] http://www.mamma.com
[9] http://www.metacrawler.com
[10] http://www.firefly.com
[11] http://bf.cstar.ac.com/bf/

documentos comprimiéndolos o resumiéndolos. Ejemplos: Architext, ConText, Autonomy[12].

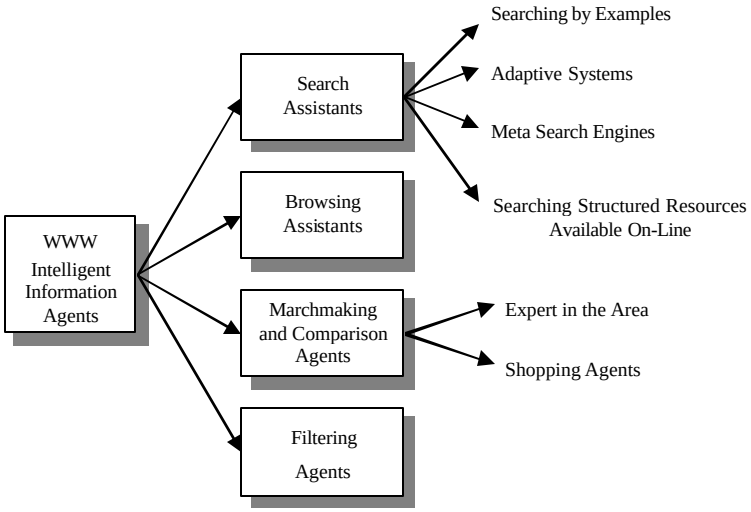
Algunas de las características deseables de los agentes son las siguientes:
- Poseer un nivel de inteligencia suficiente para aprender.
- Autonomía: La autonomía dependerá del grado de interactividad que se precise entre el usuario y el servidor.
- Movilidad: han de poder navegar por las redes y acceder a servidores.
- Modulares: Permite reutilizar el agente y reducir la complejidad de los problemas.
- Comunicación: Tienen que comunicarse con otros agentes para poder trabajar en entornos distribuidos.
- Fiables: Los usuarios sólo aceptarán a los agentes si éstos son de confianza.

Con objeto de facilitar la comunicación entre agentes, se han ideado herramientas para construir bases de conocimiento a gran escala que sean compartibles y reutilizables:
- KIF *(Knowledge Interchange Format)*. Es un lenguaje formal para el intercambio de conocimiento entre programas dispares (escritos por diferentes programadores, en diferentes momentos, en diferentes lenguajes, etc.).
- KQLM *(Knowledge Query and Manipulation Language)*. Es un formato de mensaje y un protocolo para manejar mensajes con el objetivo de soportar el conocimiento compartido entre agentes. Es además una interfaz de comunicación entre agentes; está enfocado a las operaciones que los agentes usan para comunicarse.

Muchos servicios de búsqueda de Internet están ya ofreciendo información personalizada. Por ejemplo My Yahoo!, que pregunta al usuario cuando se da de alta por sus temas de interés, aficiones, edad, sexo, en qué país y región vive, empresas cuya cotización en bolsa le interesa, y con toda esta información confecciona unas páginas que incluyen enlaces a noticias, otros servidores, y recursos disponibles en Internet que tratan de aquello que interesa al usuario.

Una buena clasificación establecida por Casasola [Casasola, 97] divide los agentes inteligentes según 4 dominios de aplicación:
- Asistentes en la navegación (Browsing Assistants): Este tipo de agentes monitorizan la actividad del usuario y recomiendan documentos facilitándole la búsqueda.
- Asistentes de búsqueda (Search Assistants): En esta categoría entran aquellos agentes que aportan nuevas características al proceso de búsqueda orientadas a la comunicación entre agentes.
- Agentes de emparejamiento y comparación (Marchmaking and Comparison Agents): Se encargan de monitorizar sitios que cambian con frecuencia para resumir su contenido y extraer relaciones entre sus componentes que permitan mantener información actualizada.
- Agentes de filtrado (Filtering Agents): Su función consiste en seleccionar de entre grandes volúmenes de información aquella más útil para el usuario.

[12] http://www.agentware.com



Clasificación de Agentes Inteligentes según Casasola [Casasola, 97]

Algoritmos de evaluación.

Los algoritmos de evaluación que utilizan muchos de los buscadores actuales se basan en la estructura de la web para determinar su relevancia. Estos algoritmos de evaluación se denominan algoritmos basados en enlace y son 3 principalmente: *PageRank* (utilizado en el buscador Google), HITS (Hypertext Induced Topic Selection) y SALSA. Los algoritmos basados en enlace se apoyan en la estructura de la Web, considerada como un grafo dirigido de páginas y enlaces: una página con muchos enlaces a ella se supone que es una página de alta calidad, especialmente si (circularmente) los enlaces vienen de páginas que son a su vez de alta calidad. Por tanto, se puede considerar a la web como un grafo dimitido G = (P, E) donde P son los nodos o páginas web y E los enlaces entre las páginas.

Este tipo de algoritmos sufren el "efecto de la contribución circular" [Wang, 04]. Este efecto se basa en el hecho de que las páginas se pueden enlazar unas a otras, de forma que se produzca un camino circular entre ellas. Por tanto, cada página estimula la evaluación de las que enlaza, y si existe un camino circular, entonces estimula su propia evaluación indirectamente. Para tratar de evitar este problema, Wang [Wang, 04] propone la aplicación del concepto de "distancia en la Web", de forma que se asignen pesos a los enlaces en función de la importancia de la página enlazada.

Además, también presentan el inconveniente de que son potencialmente vulnerables a ataques del tipo *link spamming* como demostró experimentalmente Lempel [Lempel, 01]. Un ejemplo de este tipo de desviaciones se produce en Google cuando se introduce el término "ladrones". Cuando se introduce esta consulta, el primer documento que aparece es la página de la sociedad general de autores. La explicación es que mucha gente se ha puesto de acuerdo para enlazarlas en sus páginas utilizando el término "ladrones", debido a la baja popularidad que tienen por el cobro de cánones en la adquisición de material informático.

El algoritmo *Pagerank* define un camino aleatorio con saltos aleatorios sobre la web (completa). Los estados del camino aleatorio son las páginas web, y la puntuación de cada página se define mediante sus valores de distribución estacionarios del camino aleatorio. Así, la puntuación

PageRank de una página se puede interpretar como *global*, evaluando la importancia de cada página independiente del tema [Lempel, 04].

Por otra parte, HITS y SALSA son específicos a un tema y se pueden considerar como algoritmos de evaluación *locales*. Estos dos algoritmos funcionan utilizando una pequeña porción de la Web donde los recursos correspondientes de un tema específico es probable que existan, analizando la estructura de enlaces de ese subgrafo Web y asignando a sus páginas puntuaciones hub y autoridad. Una página es una autoridad en un tema si contiene información valiosa y de alta calidad sobre ese tema. Una página es un "hub" sobre un tema si enlaza a buenas autoridades sobre el tema, si es por ejemplo una lista de recursos de calidad sobre ese tema.

PageRank

La puntuación PageRank de una página A (denotada como PR(A)) es la probabilidad de visitar A en un camino aleatorio que implique a toda la web, donde el conjunto de estados del camino aleatorio es el conjunto de páginas, y cada paso aleatorio es de uno de estos tipos:

1. Elegir una página Web aleatoriamente, y saltar a ella.
2. Desde un estado s dado, elegir aleatoriamente un enlace saliente de s y seguir ese enlace hasta la página destino.

Brin [Brin , 98] describe el cálculo del algoritmo PageRank de la siguiente forma:

Se asume que la página A tiene las páginas T1...Tn que apuntan a ella. El parámetro d es un factor que puede tomar valores comprendidos entre 0 y 1. Normalmente se establece d con el valor 0.85. Además C(A) se define como el número de enlaces que salen de la página A. El valor PageRank de una página A se determina como sigue:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

Hay que destacar que PageRank establece una distribución de probabilidad sobre las páginas web, de tal modo que la suma de todos los valores PageRank de las páginas web serán uno.

El valor PageRank o PR(A) se puede calcular utilizando un algoritmo iterativo. La idea sobre la que se basa PageRank es bastante intuitiva. Asume que si una página recibe bastantes enlaces provenientes de otras páginas, entonces se supone que esa página merece ser visitada. No obstante, también tiene en cuenta el hecho de que páginas muy importantes enlacen a otra página, lo que implica que es probable que esa página sea digna de ser visitada al ser enlazada por una página de calidad.

A grandes rasgos, se puede decir que el algoritmo PageRank mide la probabilidad de que un usuario visite una página web. El factor d es la probabilidad de que un visitante que navega en una página se aburra de ella y solicite otra.

HITS (Hypertext Induced Topic Selection)

HITS se basa en un modelo de la web que distingue hubs y autoridades. Cada página tiene asignado un valor "hub" y un valor "autoridad". El valor hub de la página H esta en función de los valores de autoridad de las páginas que enlaza H, el valor autoridad de la página A está en función de los valores Hub de las páginas que enlazan a A. Por tanto, según HITS cada página

tiene un par de puntuaciones: una puntuación hub (h) y una puntuación autoridad (a), basadas en los siguientes principios:

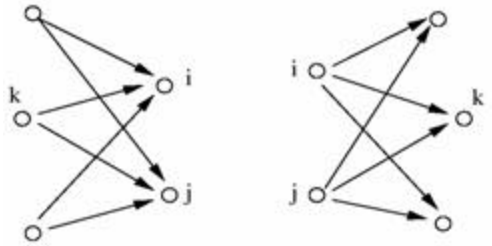
- La calidad de un hub se determina mediante la calidad de las autoridades que le enlazan.
- La calidad de una autoridad se determina mediante la calidad de los hubs a los que enlaza.

Por tanto, el algoritmo HITS establece que una página tiene un alto peso de "autoridad" si recibe enlaces de muchas páginas con un alto peso de "hub". Una páginas tiene un alto peso "hub" si enlaza con muchas páginas autoritativas. Dado un conjunto de n páginas web, el algoritmo HIT primero constituye una matriz de adyacencia A de dimensiones nxn, cuyo elemento (i,j) es 1 si la página 1 enlaza a la página j, y 0 en caso contrario. HITS se calcula mediante el cálculo iterativo de tres pasos:

1. Actualiza las puntuaciones de autoridad de cada página:
$$a^{t+1} = A^T \cdot h^t$$
2. Actualiza las puntuaciones de hub de cada página:
$$h^{t+1} = A \cdot a^{t+1}$$
3. Se normalizan las puntuaciones autoridad y hub.

Donde a y h son los vectores con los valores de autoridad y hub.

Las estructuras utilizadas para almacenar los efectos provocados por hubs y autoridades tienen relaciones con los fenómenos de *cocitación* y *coreferencia* en el estudio de las valoraciones basadas en los enlaces. Así, si dos páginas Web distintas pi y pj están *cocitadas* por muchas otras páginas Web pk (Figura 2.5), es probable que estén relacionadas en algún sentido. A su vez, si dos páginas Web distintas pi y pj *coreferencian* varias otras páginas web pk implica que pi y pj tienen ciertos aspectos en común.



A la izquierda las páginas Web p_i y p_j son cocitadas por la página web p_k. A la derecha las páginas web p_i y p_j coreferencian a la página p_k.

SALSA (Stochastic Approach for Link Structure Analysis)

SALSA también asigna dos puntuaciones a cada página: la puntuación hub y autoridad. Estas puntuaciones se basan en dos caminos aleatorios realizados en G, el camino autoridad y el camino hub. Intuitivamente, el camino autoridad sugiere que las páginas autoritarias deberían ser visibles (enlazadas) desde muchas páginas. Así, un camino aleatorio de este subgrafo visita aquellas páginas con alta probabilidad. Formalmente, el estado del camino autoridad son los nodos de G con al menos un enlace de entrada. Sea v un nodo, y q1, ..., qk los nodos que enlazan con v. Una transición desde v implica elegir un índice aleatorio i uniformemente sobre {1, 2, ..., k}, y seleccionar un nuevo estado desde los enlaces salientes de qi (de nuevo, aleatoriamente y uniformemente). Así, la transición implica atravesar dos enlaces Web, el

primero de ellos se atraviesa al revés (desde el destino al origen) y el segundo se atraviesa hacia delante. Si p denota la distribución estacionaria del camino aleatorio descrito anteriormente, cuando la distribución inicial es uniforme sobre todos los estados. La puntuación de cada página (=estado) v es p_v (las páginas que no tienen enlaces de entrada alcanzarán una puntuación 0).

Cabe destacar el efecto TKC (*Tightly-Knit Community*) [Lempel, 04] que remarca importantes diferencias entre el algoritmo HITS y SALSA. HITS favorece a los grupos de páginas que tienen muchas cocitaciones "internas", mientras SALSA prefiere las páginas con muchos enlaces de entrada. Una comunidad estrechamente tejida (*tightly-knit community*) es un conjunto de páginas pequeño pero sumamente interconectado. El efecto TKC se da cuando dichas colecciones de páginas (comunidades estrechamente tejidas) obtienen evaluaciones altas en los algoritmos basados en los enlaces, aunque esas páginas no sean autoridad en el tema, o solo conciernan a un aspecto de dicho tema.

Otros factores que intervienen en la evaluación

Además de la estructura de los enlaces, también se suelen tener en cuenta a la hora de evaluar una página otras características. Por ejemplo, Google tiene en cuenta el texto que acompaña a cada enlace, ya que se supone que da una descripción general o el nombre de la página a la que enlaza. Esto tiene varias ventajas ya que permite obtener una descripción bastante exacta de la página, además permite recuperar documentos que no estén basados en texto como por ejemplo imágenes, programas o bases de datos. Otros aspectos que se suelen tener en cuenta son el título de la página, el tamaño de la fuente empleada, etc.

Algoritmos de evaluación en los meta-buscadores

Los algoritmos de evaluación de los meta-buscadores se tienen que enfrentar a problemas diferentes a los buscadores tradicionales a la hora de evaluar. El principal problema se debe a la fusión de las listas de documentos evaluados que cada buscador devuelve. Las valoraciones producidas por diferentes buscadores normalmente no son comparables ya que se calculan frecuentemente utilizando alguna métrica en función de la distancia.

Problemas en la recuperación de información

Un sistema de recuperación de información abraza la creencia de que la información puede organizarse y representarse para su recuperación, y que las necesidades de información tienen alguna característica que se repite. Se asume que la representación de los documentos se realiza de forma textual. Sin embargo, no solo influye la representación de los documentos en el proceso de evaluación, sino que existe otro cúmulo de circunstancias que afectan de forma apreciable a la calidad de los resultados de una búsqueda.

Uno de estos factores que afecta al rendimiento en la recuperación de información es el usuario y su carencia de información para expresar lo que quiere. Además, en algunas situaciones la información relevante se reconoce solo cuando se encuentra y examina, no antes . Con bastante frecuencia, la búsqueda de los usuarios para cumplir sus necesidades de información se puede describir acertadamente mediante la frase:

> "No sé lo que estoy buscando, pero lo sabré cuando lo encuentre" [Bruza, 93]

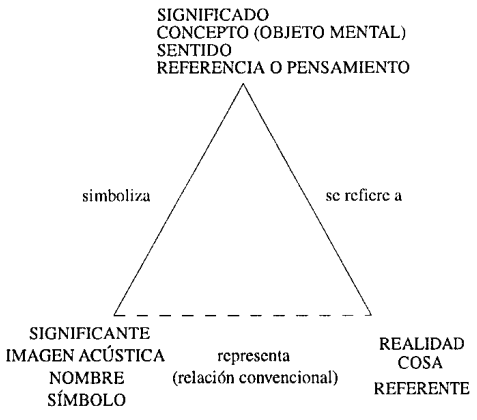
El lenguaje natural y la concepción y utilización particular del que hacen uso los distintos usuarios es uno de los principales problemas a la hora de buscar información. La representación de los documentos se realiza de forma textual, que es una representación escrita del lenguaje natural. Además, el usuario también debe expresar lo que busca de forma textual para que sea comprensible para el SRI. Por tanto, el problema derivado de la utilización particular del lenguaje natural se traslada también a su representación escrita, es decir, al texto. Un ejemplo de los problemas derivados de la variabilidad del lenguaje se produce cuando dos documentos tratan del mismo concepto pero utilizando dos palabras diferentes a la hora de nombrarlo, siendo por tanto indexados utilizando cada palabra concreta. Si un usuario busca información sobre este concepto, si no tiene presente que se puede describir el mismo concepto mediante dos palabras diferentes, será más complicado para él obtener ambos documentos. Esta situación ilustra que el usuario y sus conocimientos también juegan un papel importante en el proceso de búsqueda.

Lenguaje, semántica y contexto

En este proyecto se habla mucho de la relación semántica entre distintos términos. Para aclarar este aspecto de dicho trabajo, se procede a describir el concepto de semántica.

La semántica estudia el significado de los signos lingüísticos y de sus combinaciones. Se puede distinguir entre semántica léxica, que se ocupa del sentido de las unidades del vocabulario, la semántica gramatical, cuyo cometido es estudiar las relaciones entre los elementos gramaticales que forman una oración y la semántica lingüística, que analiza el significado de los signos lingüísticos. Para este trabajo, la semántica lingüística es a la que se presta mayor atención, ya que las otras semánticas solo se fijan en la forma de las palabras y no en su significado.

Es evidente que cualquier signo lingüístico empleado hace referencia a una porción de la realidad, ya sea un objeto concreto o abstracto. Tal y como muestra el triángulo semántico o triángulo de Odgen-Richards [Odgen, 72], se puede apreciar que existe una relación entre las palabras y la realidad a través del sentido. Existen diferentes representaciones de este triángulo que utilizan en sus vértices diferentes palabras para representar la misma idea. En la figura se muestran algunas de las más empleadas.



Triángulo semántico

Mediante esta representación se puede apreciar que el significante de una palabra "simboliza" un significado, el cual "se refiere a" un elemento de la realidad o referente. La línea discontinua de la base del triángulo indica que no existe una relación directa entre el significante o palabra y el referente o elemento de la realidad: si no se conoce el significado de una palabra, no es posible establecer una relación que permita representar el referente perteneciente a la realidad mediante un símbolo o significante. Se puede definir por tanto el significante como la entidad física, típicamente en forma de imagen acústica, perceptible por los sentidos. El significado es un concepto, la idea que se representa en la mente al escuchar el significante. Por último, el referente es la realidad efectiva a la que se remite el signo.

Un aspecto relevante del lenguaje a tener en cuenta es el tipo de relación existente entre los significantes y el significado. Atendiendo al tipo de relación se pueden distinguir los siguientes casos:

- **Monosemia**: Se produce cuando la relación entre el significante y el significado es estrictamente de uno a uno y viceversa. Cuando se da esta situación se dice que el significante es monosémico. Este es el mejor caso computacionalmente hablando, ya que una vez conocido el significante se puede conocer sin lugar a dudas su significado. Pero esta situación dista mucho de ser la habitual en la lengua común, produciéndose en su lugar el siguiente tipo de relación.

- **Polivalencia semántica o significación múltiple**: Este fenómeno se presenta cuando la relación entre significante y significado es de uno a varios. En este caso, un mismo significante puede simbolizar distintos conceptos. La significación múltiple se puede presentar de dos formas distintas:

    - Homonimia: cuando dos significantes originalmente distintos en su forma fonética y que simbolizan distintos conceptos llegan a coincidir a través del tiempo en un mismo significante (con igual forma fonética). Por ejemplo /llama/ cuando significa "masa gaseosa en combustión" viene de *flamma*, si se refiere a un rumiante sudamericano la palabra se toma del quechua, y cuando es forma del verbo llamar, proviene del verbo latino *clamare*.

    - Polisemia: cuando un significante adquiere un nuevo significado a lo largo del tiempo. Por ejemplo /Java/ es una isla de Indonesia pero también el nombre de un lenguaje de programación.

- **Sinonimia**: Es el caso inverso a la significación múltiple. Se presenta cuando la relación entre significante y significado es de varios a uno. Es decir, se produce cuando existen varios significantes diferentes en su forma pero que hacen referencia a un "mismo" significado. Sin embargo, hoy se acepta normalmente que la sinonimia concebida como relación precisa, esto es, como relación de equivalencia entre dos expresiones, o como identidad de significado, no existe en la práctica [Fernández Lanza, 01]. Es decir, normalmente entre dos palabras consideradas como sinónimas existen ligeros matices que acentúan alguna característica del mismo concepto.

La existencia de significantes idénticos en la forma y diferente en el significado es un fenómeno común, al parecer ventajoso para la memoria, que se ve liberada de tener que retener una palabra diferente para cada concepto nuevo que se produzca. Así mismo, la existencia de sinonimia aporta variedad a la utilización del lenguaje, evitando la repetición excesiva de las

palabras e incorporando diferentes matices en función de la expresividad, énfasis o intención de una comunicación. Pero es necesario disponer de una serie de mecanismos que permitan desambiguar los casos de significación múltiple para obtener una comunicación efectiva. Este proceso de desambiguación se consigue teniendo en cuenta el resto de significantes que intervienen en una comunicación, y que influirán en cierta medida en la decisión de seleccionar alguno de los posibles significados.

Es decir, el significado de una misma palabra depende del contorno lingüístico que la envuelve y que determina su significación. Debido a este fenómeno, se pueden distinguir dos tipos de significado: referencial o contextual. Una palabra tiene un significado referencial cuando se "refiere" a su relación convencional con la realidad. Los significados referenciales son aquellos que se pueden encontrar en un diccionario. En cambio, el significado contextual es el que adquiere la palabra dentro de un contexto, cuando amplía, restringe y aún transforma el significado referencial.

Por tanto, el contexto es un elemento determinante en el acto comunicativo, y que viene determinado por los actos comunicativos anteriores y posteriores. Además, es el criterio utilizado para determinar uno de los posibles significados y permitir descartar el resto. Según Van Dijk [Van Dijk, 99], el contexto se define como el conjunto estructurado de todas las propiedades de una situación social que son posiblemente pertinentes para la producción, estructuras, interpretación y funciones del texto y la conversación. Es decir, que no solo intervienen factores lingüísticos, sino también sociales y culturales. De hecho, Claire Kramsch nombra cinco dimensiones que afectan al contexto (lingüística, situacional, interactiva, cultural e intertextual) [Kramsch, 96].

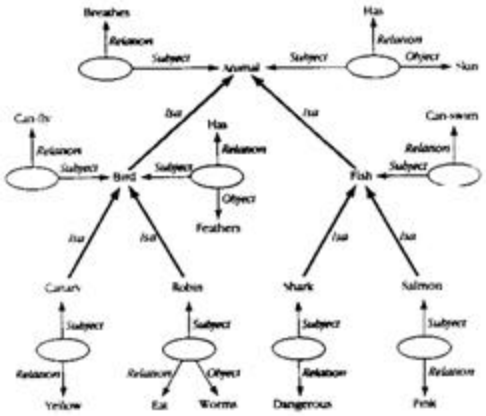
La computación de todos los elementos que afectan al contexto es una labor excesivamente compleja y que requiere de información previa que actualmente solo poseen las personas, adquirida a través de su propia experiencia. Sin embargo, si que existen buenos diccionarios o tesauros que permiten conocer las distintas acepciones de una misma palabra. Se poseen herramientas que permiten, en definitiva, conocer los significados referenciales de las palabras. El inconveniente de estos diccionarios es que la mayoría están diseñados para su utilización por personas, lo que implica que atienden a un único criterio de ordenación (típicamente el alfabético) y que carecen de una estructura que establezca un criterio que permita conocer las relaciones entre varias palabras sin atender a su forma (sustantivo, verbo, etc.), sino a su semántica. Afortunadamente, existen algunas excepciones como son por ejemplo WordNet.

Pero, ¿qué es lo que permite descartar o aceptar un determinado significado de una palabra con significación múltiple?. Una persona se decanta por un significado u otro en función de las posibles relaciones semánticas que se puedan establecer entre los significantes participantes de una comunicación. Por ejemplo, ante la expresión: "Coge el gato", en función del contexto o la situación "gato" se referirá a "un mamífero de la familia de los felinos" o bien a "un instrumento que se utiliza para levantar grandes pesos a poca altura". Si se modifica la expresión anterior aportando más información obtenemos: "Se ha pinchado el coche. Coge el gato". En la expresión anterior se puede apreciar que el significante "gato" se refiere al significado "un instrumento que se utiliza para levantar grandes pesos a poca altura". Esta afirmación se puede realizar debido a que todo el mundo sabe que el gato es un instrumento que se utiliza para arreglar pinchazos de los coches. Por tanto, basándose en este conocimiento previo y en el contexto, se puede determinar el sentido correcto de "gato". Se puede apreciar que existen tres factores básicos necesarios para poder realizar esta desambiguación correctamente: un contexto, algún tipo de relación semántica y la necesidad de un conocimiento previo que nos permita conocer esa relación semántica.

Para que exista alguna relación semántica entre varios significantes, es necesario que pertenezcan a algún campo semántico que contemple esa relación.. Fue J. Trier el primero que definió el *campo semántico* (aunque denominándolo "word field") en 1931, como el conjunto de elementos delimitados mutuamente sin sobreponerse, "como las piezas de mosaico" [Trier, 31]. Posteriormente Lyons [Lyons, 77] utiliza el término campo semántico (semantic field) para describir el mismo concepto. Así, en 1934, Trier afirma que el valor de una palabra sólo puede determinarse definiéndolo en relación con el valor de las palabras vecinas que contrastan. Sólo tiene sentido como parte del todo; pues hay significado sólo en el campo semántico [Marcos, 98].

Por tanto, se puede definir **campo semántico** como un conjunto de palabras que comparten un contenido común (un trozo de realidad al cual se refieren todas) y se lo reparten de tal modo que cada una de esas palabras se opone a las demás por rasgos propios. Cada uno de estos rasgos semánticos diferenciales se llama **sema** [Mmuruza, 04]. Un campo semántico puede estar formado por distintos tipos de palabras como sustantivos o verbos. Otro rasgo distintivo que se puede apreciar, es que suelen poseer una estructura interna con subcampos que comparten alguna característica común. Como ejemplo se nombrará el campo semántico de los colores formado por sustantivos como *rojo, verde, azul,* por adjetivos como *rojizo, verdoso, azulado,* y otras palabras no tan evidentemente relacionadas como *frío* y *cálido* entre otras. Dentro de dicho campo, cada palabra posee rasgos distintivos que lo oponen a los demás, como por ejemplo el **sema** distintivo de *frío* es «que produce efectos sedantes, como el azul o el verde» y de *cálido* «que predominan los matices dorados o rojizos».

Pero es extremadamente difícil construir alguna estructura que contemple la totalidad de las posibles relaciones existentes en la realidad. Existen algunos esquemas de representación formalizados que intentan representar el conocimiento para posteriormente poder obtener ese tipo de relaciones. Estos esquemas, cuyo primer impulsor fue Quillian [Quillian, 68] con su memoria semántica, se denominan **redes semánticas**, y han sido muy utilizados en la Inteligencia Artificial para representar las relaciones entre conceptos de una determinada área. Este problema de representación de la realidad afecta también al proceso de desambiguación, lo que repercute de forma negativa en los SRI.



Memoria semántica de Quillian (Fuente: [Shulman, 97])

Proceso de búsqueda

El inicio del proceso de búsqueda lo origina un **problema** que requiere información para poder resolverse. La carencia de esta información depende de la amplitud de conocimiento de cada usuario. Un usuario avezado en un tema concreto tendrá más claro que información solucionaría su problema y seguramente lo encontraría en un plazo de tiempo más corto. La aparición de un problema conlleva la demanda de información en el usuario para solucionarlo, y esta carencia de información origina lo que se denomina una **necesidad de información**.

Las personas buscan información basándose en su conocimiento previo, que es muy diferente de unas a otras personas. La necesidad de información se define como la representación implícita de un problema en la mente de los usuarios [Mizzaro, 96a]. Se diferencia del problema, ya que cada usuario percibe las cosas de diferente forma, y ante un mismo problema varios usuario pueden construir necesidades de información distintas. Las necesidades de información se pueden clasificar en necesidades verificativas, sobre temas conscientes e imprecisas o mal definidas [Ingwersen, 95]. La primera categoría se refiere a la situación en la que se buscan documentos con propiedades conocidas, por ejemplo se conoce el nombre del autor, el título, etc. En el segundo tipo se conoce el tema y es definible, pero menos exacto que en la primera categoría. En esta categoría una persona que busca información tiene algún nivel de comprensión de lo que busca. La tercera categoría son los casos en los que una persona desea encontrar nuevo conocimiento y conceptos en dominios que no le son familiares.

Una necesidad de información se puede satisfacer de distintas formas [Mizzaro, 96b]. Es decir, el concepto de necesidad de información tiene una naturaleza ambigua. Debido a esta característica, se han comentado distintos problemas cuyo motivo es la inexactitud de la necesidad de información, como el problema ASK [Belkin, 82] (Anomalous State of Knowledge), ISK (Incomplete State of Knowledge) y USK (Uncertain State of Knowledge) [Ingwersen, 92].

Los sistemas IR se basan en la idea de que las necesidades de información pueden describirse. La persona encargada de recuperar la información tiene que ser capaz de expresar la necesidad de información que demanda en forma de una **petición** (request). La petición es una representación de la necesidad de información del usuario en un lenguaje humano, normalmente lenguaje natural [Mizzaro, 96a].

El último paso consiste en indicar al SRI la necesidad de información en un lenguaje comprensible para él. El usuario debe formalizar su petición por medio de una **consulta** (query) cuya representación sea procesable por el SRI.

Proceso de construcción de las consultas

Evidentemente, la representación mental de la información que el usuario necesita para resolver su problema difiere con mucho de la información que recibe el SRI del usuario. Este proceso implica una adaptación de lo que el usuario cree que resolverá su problema a una expresión que represente lo que el usuario necesita encontrar.

Pero no basta con seguir este proceso para obtener la información que resuelva el problema. Si los resultados no satisfacen al usuario puede ser necesario repetir este proceso de forma cíclica. Durante cada ciclo el sistema recibe realimentación del usuario con nueva información, formalizada en forma de nuevas consultas. En este proceso, se pueden distinguir a grandes rasgos 4 fases [Hofstede, 96]:

1. Fase explorativa. El usuario reúne la información que pueda serle útil en el proceso de búsqueda.
2. Fase constructiva. Se aprovecha la información adquirida en la fase anterior para reformular una nueva consulta.
3. Fase de realimentación. Si los resultados de la consulta formulada en la fase 2 no son satisfactorios es necesario volver a realizar las fases 1 y 2 para refinar el resultado.
4. Fase de presentación. Se limita a la forma de representar los resultados.

Se pueden encontrar otras muchas descripciones del proceso de búsqueda en la literatura. Por ejemplo, el modelo propuesto por Kuhlthau [Kuhlthau, 88] que divide el proceso de búsqueda en siete etapas: comienzo, selección, exploración, formulación de la consulta, análisis de los resultados, recopilación de documentos y presentación de los resultados. En el marco del modelo de Kuhlthau se ha estudiado el comportamiento del usuario para la obtención de patrones de interacción [Stelmaszewska, 02].

Problema del vocabulario

El funcionamiento de muchos sistemas depende de que los usuarios introduzcan las palabras correctas. Los usuarios nuevos o esporádicos frecuentemente utilizan palabras incorrectas y fallan al lograr las acciones o información que quieren. Este es el problema del vocabulario [Furnas, 87].

Los usuarios utilizan una sorprendente cantidad de términos para referirse a conceptos similares o relativos al mismo tema. Por ejemplo, en los directorios donde los documentos se incluyen en categorías de forma manual, puede suceder que categorías similares pertenecientes a una misma

rama confundan al usuario y no encuentre en ella lo que esperaba. El problema es que la persona que asigna un documento a una categoría concreta puede dar más importancia a unas palabras concretas o tener una concepción diferente a la del usuario que realiza la búsqueda. Esta situación se deriva del problema del vocabulario, ya que cada usuario tiene una preferencia personal a la hora de utilizar unas u otras palabras. Según Furnas, en la elección espontánea de una palabra para objetos de cinco dominios, la probabilidad de que dos personas escojan el mismo término está por debajo de un 20% [Furnas, 87].

Los SRI también sufren este tipo de problema. Cuando un autor escribe un documento, utiliza un vocabulario específico y personal, que en muchos casos no coincidirá con el utilizado por otras personas. Puede darse el caso de que otra persona utilice otras palabras para describir lo mismo, con un sinónimo, un alias o una frase explicativa. Pues bien, los SRI cuando indexan un documento, lo hacen atendiendo exclusivamente a los términos que forman el documento (salvo excepciones como el modelo FIS-CRM). Cuando un usuario realiza una consulta elegirá un conjunto de términos que para él son representativos del concepto que busca. Si los términos utilizados en la consulta son distintos a los que forman el documento, es muy probable que ese documento no sea recuperado o que lo haga con un grado de relevancia muy bajo.

Otro aspecto que agudiza el problema del vocabulario es la distribución geográfica de los usuarios, ya que los conceptos o ideas y sus vocabularios asociados pueden evolucionar o cambiar a lo largo del tiempo [Chen, 94]. Es decir, este problema se acentúa entre comunidades que hablan la misma lengua, pero dispersas geográficamente. Por ejemplo, para un argentino la palabra “manejar” se refiere a la palabra “conducir” para un español. Además, el problema del vocabulario también se produce cuando se traducen palabras de un idioma a otro común, dando como resultado traducciones distintas. Por ejemplo, “fuzzy logic” se puede encontrar traducido como “lógica difusa” o “lógica borrosa”.

Existen distintas propuestas para tratar de solucionar este problema. Casi todas ellas se basan en la construcción de estructuras de conocimiento que contemplen las relaciones entre los distintos términos para ser tenidos posteriormente en cuenta. Cabe destacar el denominado Unlimited Aliasing [Furnas, 87], que consiste en asociar a cada objeto una lista de alias y el espacio concepto [Chen, 94], que es un algoritmo para crear diccionarios para almacenar la riqueza del vocabulario y la similitud de los términos. Otras técnicas que intentan paliar este problema son los mecanismos de expansión de consulta y más recientemente, el modelo FIS-CRM, que contempla los términos relacionados conceptualmente, aunque no presentes en el documento, a la hora de indexar.

El usuario

Otro aspecto muy importante que afecta a la recuperación de información es el usuario. Cada usuario es distinto de los demás en sus motivaciones, suposiciones, conocimientos y experiencia. Este cúmulo de circunstancias afecta a la forma en la que cada usuario utiliza un SRI. Los principales problemas que afectan al usuario en su interacción con el SRI [Lam, 01] son la forma en la que especifica su consulta y la forma en la que interpreta la respuesta proporcionada por el SRI.

Los usuarios no suelen aprovechar al máximo las posibilidades que ofrecen las herramientas de búsqueda. Diferentes estudios [Jansen, 98][Silverstein, 99][Cacheda, 01] de los buscadores de Internet muestran las deficiencias más comunes a la hora de utilizar estas herramientas.

La principal deficiencia es el bajo número de términos que los usuarios utilizan en sus cadenas de búsqueda. La media de utilización va desde 1.63 [Cacheda, 01] hasta 2.35 [Jansen, 98] [Silverstein, 99] términos por consulta. La utilización de pocos términos repercute en una menor precisión, debido a la desviación que puede causar la polisemia y a la inexistencia de otros términos que permitan concretar la búsqueda.

Otro factor que afecta a la obtención de buenos resultados es la poca realimentación que los sistemas de búsqueda reciben del usuario. Para estudiar este fenómenos, se establece el concepto de sesión, que es el periodo de tiempo durante el cual el usuario realiza consultas al sistema de forma continuada. Cacheda asume que una sesión de usuario no dura más de 30 minutos, de hecho lo normal es entre 9 y 30 minutos [Cacheda, 01]. El resultado es que el número medio de consultas por sesión es muy bajo, entre 1.75 y 2.8.

Además, el número medio de pantallas que los usuarios visitan es muy bajo (entre 1.39 y 2.21). Si se considera que la mayoría de los usuarios solamente comprueban una o dos pantallas del resultado, se puede determinar que los usuarios buscan de una forma muy limitada: sencillamente abriendo los documentos con el título y descripción que mejor se ajusta a sus necesidades [Cacheda, 01]. La forma en la que se proporciona la retroalimentación suele ser modificando alguno de los términos de la consulta, aunque también se suele hacer cambiando totalmente la consulta, añadiendo términos, eliminando términos o modificando los operadores de búsqueda.

	Términos por consulta	*Consultas por sesión*	*% usuario que solo examina 1ª pantalla*
Jansen	2.35	2.8	57%
Silverstein	2.35	2.02	85%

Comparativa de los estudios de uso de los buscadores

Por último, la utilización de los operadores de búsqueda es muy baja. El estudio de Jansen afirma que solo un 6% de los usuarios utilizan los operadores boleanos y un 7% los operadores '+' y '-'. Según Cacheda, el operador boleano AND es el más utilizado, seguido del operador "" y los paréntesis.

Todas las estadísticas aquí ofrecidas son orientativas, fruto de diversos trabajos de investigación, pero demuestran las carencias de los usuarios a la hora de utilizar las herramientas de búsqueda (los buscadores de Internet en este caso). Por tanto, es de esperar que si un usuario no utiliza de la mejor forma posible las posibilidades de los buscadores, estos no puedan proporcionar los mejores resultados. No obstante, es posible encontrar diversos sistemas o herramientas que tratan de paliar este inconveniente utilizando distintas técnicas, destacando la expansión de consulta interactiva [Efthamidis, 96], que ayudan al usuario en el proceso de búsqueda. Spink [Spink, 01] concluye que la interacción de los usuarios con los motores de búsqueda en la web son cortas y limitadas y propone para ajustar estos factores y el comportamiento humano la necesidad de una nueva generación de herramientas de búsqueda Web que trabajen con la gente para ayudarles a proseguir con la búsqueda electrónica para resolver sus problemas de información.

Las características del SRI

Los SRI poseen una interfaz de búsqueda que permite indicar al sistema lo que el usuario quiere buscar. Normalmente el usuario indicará sus necesidades de búsqueda utilizando un conjunto de palabras, que se llamará cadena de búsqueda, combinadas utilizando algunos operadores de búsqueda. Pero la cadena de búsqueda introducida por el usuario es sometida a transformaciones por los SRI en función de sus características particulares. El resultado de estas transformaciones es una expresión denominada términos de búsqueda, y suele consistir en un subconjunto de la cadena de búsqueda.

Las transformaciones que sufre la cadena de búsqueda a menudo propician cambios en los resultados esperados por el usuario. Por tanto, un adecuado conocimiento de las características particulares del SRI hace consciente al usuario de estas transformaciones. Esta situación provoca una desviación en las estrategias de búsqueda del usuario para intentar adecuar a sus intenciones de búsqueda los resultados. Algunas de las transformaciones de la cadena de búsqueda más comunes en las herramientas de búsqueda de Internet actuales son las siguientes [Muramatsu, 01]:

- ***Aplicación de un operador boleano de búsqueda por defecto***: aunque los SRI actuales no suelen basarse en el modelo boleano, si que permiten la utilización de operadores boleanos en la cadena de búsqueda. La utilización de los mismos influye de forma importante en los resultados de la búsqueda. Por tanto, la utilización de alguno de estos operadores por defecto cuando el usuario no indica ninguno puede afectar seriamente al resultado de la búsqueda. Se suele considerar el operador OR por defecto (como el buscador Excite), pero los resultados serán totalmente diferentes si el operador por defecto es AND (Northern Light).
- ***Eliminación de palabras prohibidas***: a la hora de indexar los documentos muchos SRI suelen ignorar las palabras prohibidas, por tanto, tampoco tendrá mucho sentido permitir la utilización de este tipo de palabras en la consulta del usuario, ya que no será posible recuperarlas. Google elimina las palabras prohibidas, por lo que la búsqueda “to be or no to be” en Google no devuelve ningún resultado. Por el contrario, actualmente existen formas de evitar este comportamiento utilizando el operador “, que permite indicar exactamente lo que se busca.
- ***Expansión de los términos con sufijos***: algunos SRI consideran, a la hora de recuperar documentos, como términos relevantes los que compartan la raíz con alguno de los términos de búsqueda, como por ejemplo Yahoo.
- ***Influencia del orden de los términos en el resultado de la búsqueda***: Algunos algoritmos de ranking prestan especial atención a la proximidad y orden de los términos de búsqueda. Evidentemente, si este aspecto se tiene en cuenta, el resultado será diferente cuando los mismos términos se utilizan en diferente orden. Esta situación se da en Google, donde por ejemplo se obtendrán diferentes resultados para la consulta “boat fire” y “fire boat”.

Solucionar este problema causado por el desconocimiento del usuario de las transformaciones a las que se somete la cadena de búsqueda es complicado. Normalmente, los usuarios suelen utilizar un único SRI, que hace que con el uso se conozcan mejor sus características y por tanto el usuario adapte su forma de buscar. Sin embargo, existen algunos trabajos, como por ejemplo las consultas transparentes [Muramatsu, 01], que informan al usuario de las transformaciones que sufrirá su cadena de búsqueda y le permita conocer como una consulta particular es procesada e interpretada.

La naturaleza de Internet

Hoy en día la mayoría de los SRI tienen a Internet como espacio de búsqueda. Este motivo influye en los SRI, ya que Internet posee una naturaleza particular y poco apropiada para su tratamiento por parte de los SRI tradicionales. Algunos de los principales problemas que presentan los documentos de la Web a la hora de buscar información son los siguientes [Baeza-Yates, 00]:

- *Distribuidos*: la naturaleza distribuida de la web influye en la confiabilidad de las conexiones y el ancho de banda disponible.
- *Volátiles*: el volumen de información crece, así como su contenido, estimándose que el 40% de la web cambia cada mes [Baeza-Yates, 00].
- *Dinámicos*: muchos documentos se suelen construir utilizando el contenido de bases de datos.
- *Sin estructura*: La web está formada por datos semi-estructurados sin una estructura constante.
- *Redundantes*: Hay muchas páginas duplicadas y se estima en un 30% el número de sitios replicados.
- *Tipos heterogéneos*: Cada vez hay más formatos de representación de los documentos, y por tanto, es necesario conocerlo para poder estudiar su contenido e indexarlos correctamente. Además, hay documentos en distintos lenguajes y que utilizan distintos alfabetos.
- *Calidad heterogénea*: Al no haber una autoridad encargada de velar por la corrección de los documentos de Internet, es posible que existan documentos con información falsa, con errores ortográficos, mal redactada, etc.

Las características comentadas hacen imposible para un SRI indexar la totalidad de la información que contiene Internet. Este tipo de sistemas serán incapaces de indexar los documentos generados de forma dinámica, como por ejemplo las páginas asp o php. Los SRI solo son capaces de encontrar las páginas que previamente han indexado, por tanto, los buscadores no serán capaces de encontrar todas las páginas posibles.

Los buscadores de Internet, mediante sus crawlers, comienzan indexando un conjunto de páginas (de las que tienen su URL), y van introduciendo en colas los enlaces que encuentran en las páginas que analizan para posteriormente indexarlas. Pero si una página no es enlazada por ninguna otra (o por una página no encontrada por el buscador) no será posible para el buscador encontrarla, y por tanto no podrá indexarla. También puede darse el caso de que existan islas de páginas enlazadas las unas con las otras, pero que no reciben enlaces fuera de su isla de páginas. Si un buscador no conoce ninguna página que pertenezca a esa isla, no será capaz de encontrar el resto.

SEO (Search Engine Optimization)

La amplia utilización de los motores de búsqueda hace indispensable para el éxito o fracaso de un negocio poder aparecer entre las primeras páginas de los buscadores para conseguir un mayor número potencial de clientes. Las empresas que aparecen en las primeras posiciones de una consulta relacionada con su dominio de negocio tienen una mayor posibilidad de adquirir nuevos clientes. Este importante fenómeno ha favorecido la aparición de técnicas y sistemas que intentan mejorar la posición del ranking de un buscador de una páginas web.

SEO se define como el proceso de intentar maximizar la exposición de un sitio en varios motores de búsqueda y directorios, mediante la utilización de palabras claves y frases específicas. Este proceso también se denomina posicionamiento web. La mecánica consiste principalmente en realizar cambios al sitio (título de la página, desarrollo de un contenido rico en palabras clave importantes, utilización de META-datos) con la finalidad de hacerlo más atractivo para los motores de búsqueda [Denning, 03].

En la construcción de la página optimizada juegan factores como la localización de las palabras clave, su frecuencia, las meta-etiquetas, el tamaño del texto, etc. Algunas de las principales técnicas utilizadas son la repetición de palabras clave al comienzo de la página. Esto, unido a la especial atención que suelen prestar algunos buscadores al comienzo de la página a la hora de indexar, permite asociar fuertemente esa página con los términos repetidos. Pero esta técnica es poco atractiva para los usuarios, por lo que suelen incorporarse en etiquetas ocultas o meta-etiquetas.

Algunos de los factores económicos que favorecen la utilización de las técnicas de SEO son [Vertexera, 04]:

- Los usuarios utilizan los buscadores y la búsqueda de productos más frecuentemente que cualquier otra actividad de Internet, excepto el correo electrónico.
- Una buena posición en el ranking de un buscador es 2 o 3 veces más efectivo a la hora de generar ventas que la utilización de banners.
- El 42% de la gente que realizó una compra on-line llegó al sitio a través de un buscador.
- El método promocional de sitios Web mejor calificado por los Webmasters es el posicionamiento en los motores de búsqueda.

Aunque la técnica de SEO no se puede considerar un problema en sí misma, si que es una forma de modificar el cálculo de la relevancia de una página en un buscador. Para poder mejorar la posición es necesario un proceso de prueba y error durante el que se comprueban las modificaciones introducidas. Evidentemente, para la modificación de la posición de una página en un buscador es necesario conocer el algoritmo que utiliza para evaluar cada página. Por ejemplo, Google se basa en la popularidad de una página y en los términos que se utilizan a la hora de enlazarla, entre otros factores. Pues es posible que una comunidad de usuarios se ponga de acuerdo para asociar a una página con un término concreto, aunque no tenga nada que ver con esa página. Este es el caso de la página de la sociedad general de autores, que es el primer documento devuelto por Google cuando se introduce la consulta “ladrones” (Abril 2004).

Así mismo, los buscadores también proporcionan servicios de pago que permiten a los sitios aparecer en posiciones relevantes. Este tipo de servicios se denomina enlaces esponsorizados. Algunos ejemplos de sistemas de SEO son Vertexera, Spider Hunter[13] o Spider Food[14].

[13] http://www.spiderhunter.com
[14] http://www.spider-food.net

3. Grupos relevantes sobre el tema:

Grupos internacionales que trabajan en el desarrollo de SRID:

1. El grupo de RI del Dr. D.H. Kraft del Dpto. de Computer Science (Louisiana State University). El Dr. Kraft es el editor de una de las más prestigiosas revistas de SRID, *J. of American Society for Information Science and Technology* .
2. El grupo de RI del Dr. R.R. Yager del IONA College (NY). El Dr. Yager es el editor de una importante revista sobre el desarrollo de sistemas de información inteligentes, *Int. J. of Intelligent Systems*.
3. El grupo de RI del Dr. Miyamoto del Inst. of Engineering Mechanics and Systems (University of Tsukuba, Japón).
4. El grupo de RI del Instituto de Tecnologías de la Información y Sistemas Multimedia del Consejo Nacional de Investigación Italiano (http://www.itim.mi.cnr.it), donde destacan las Drs. G. Bordogna y G. Pasi.
5. Soft Computing Laboratory (Dpto. de Matemáticas e Informática, Univ. De Salerno, Italia), donde destaca los Drs. Vincenzo Loia y Antonio Di Nola, editores de la revista *Soft Computing*.
6. El grupo BISC (Berkeley Initiative in Soft Computing) del Electrical Engineering and Computer Sciences Department (Universidad de Berkeley) que estudia el uso de técnicas de SC (lógica borrosa, AG, redes neuronales) en Internet.

Otros grupos dedicados a RI con técnicas de Inteligencia Artificial y otras técnicas de SC son:

1. Information Retrieval Group (Dpto. of Computer Science, Univ. Of the Strathclyde, Glasgow (UK)) en el que destaca el Dr. Fabio Crestani con sus trabajos sobre modelado de incertidumbre con técnicas de SC en IR.
2. Centro de Investigación de la Web (Dpto. Ciencias de la Computación, Escuela de Ingeniería, Universidad de Chile) en el que destaca el Dr. Ricardo Baeza Yates que estudia temas de RI, minería de datos en la Web.
3. Grupo de RI (Dpto. of Computer Science, Federal University of Minas Gerais, Brasil) donde destaca el Dr. Berthier Ribeiro-Neto que estudia la RI con redes bayesianas y bibliotecas digitales.
4. Information Retrieval Group (Dpto. of Computer Science, Univ. Of Glasgow (UK)) dirigido por el Dr. Van Rijsbergen que ha desarrollado modelos de RI basados en lógicas, teoría de probabilidad y lingüística computacional.
5. Center for Intelligent Information Retrieval (CIIR) (Dpto. of Computer Science, Univ. Of Massachusetts). Centro líder en RI en el desarrollo de sistemas para el acceso eficiente a grandes bases de datos, textuales y multimedia.

En cuanto a los grupos nacionales que trabajan en el desarrollo de SRID usando técnicas de Inteligencia Artificial y SC, la mayoría pertenecen a la Red Temática Nacional sobre sistemas de acceso a la información en la Web basados en Soft Computing:

1. Laboratorio de Inteligencia Artificial (Dpto. de Computación, Universidad de A Coruña), dedicado al uso de la lógica en RI y donde destaca el Dr. Alvaro Barreiro.
2. Grupo de Sistemas Inteligentes, (Dpto. de Inteligencia Artificial, Univ. Politécnica de Madrid), dedicado al estudio de sistemas multi-agentes y donde destaca la Dra. Ana García Serrano.
3. Grupo de Ingeniería del software, (Dpto. de Leng. y Ciencias de la Computación, Universidad de Málaga) dedicado al estudio de datos estructurados y RI en la Web, y donde destaca el Dr. José Mª Troya.

4. Grupo de Compiladores y Lenguajes, (Dpto. de Computación, Universidades de A Coruña y Vigo), dedicado al desarrollo de SRID con lógicas y donde destaca el Dr. Manuel Vilares.
5. Grupo de RI y Bibliotecas digitales, (Dpto. de Informática de la Univ. de Valladolid), dedicado al desarrollo de SRID con documentos estructurados, sistemas colaborativos de RI y bibliotecas digitales, donde destaca el Dr. Jesús Vega.
6. Grupo de Tratamiento de la Incertidumbre en I.A (Dpto. de Ciencias de la Comp. e I.A, Universidad de Granada), dedicado al desarrollo de SRID con redes de creencia, donde destacan los Drs. Juan Huete y Luis M. De Campos.
7. Grupo de Procesamiento del Lenguaje y Sistemas de Información (Dpto. de Sist. Informáticos y Computación, Univ. de Alicante), que estudia la RI multilingüe en la web, donde destacan los Drs. Manuel Palomar y Antonio Ferrández.
8. Laboratorio de Sistemas Interactivos (Dpto. de Informática, Universidad Carlos III), dedicado al desarrollo de sistemas hipermedia y de acceso a la información y donde destaca la Dra. Paloma Díaz.
9. Programación lógica e ingeniería del software (Dpto. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia), dedicado al desarrollo de software orientado a la Web, destacando la Dra. Lidia Moreno.
10. Procesamiento de Lenguaje Natural (Dpto. de Lenguajes y Sistemas Informáticos, UNED), dedicado al desarrollo de SRID multi-lenguaje con técnicas de procesamiento de lenguaje natural, destacando la Dra. Felisa Verdejo.
11. Bibliotecas digitales (Dpto. de Lenguajes y Sist. Informáticos, Univ. De Sevilla), dedicado a RI y lenguajes SGML, XML, donde destaca el Dr. Jesús Torres.
12. Grupo de lenguaje natural de la UPC (Dpto. Lenguajes y Sist. Informáticos, Univ. Politécnica de Cataluña), que estudia los SRID multi-lenguaje con técnicas de procesamiento de lenguaje natural, y donde destaca el Dr. Horacio Rodríguez.
13. Grupo de estructura de datos y lingüística computacional (Dpto. de Informática y Sistemas, Univ. Las Palmas de Gran Canarias), dedicado al estudio de estructuras de datos para la RI, destacando el Dr. Octavio Santana.
14. Grupo de procesamiento del lenguaje (Dpto. de Informática, Univ. De Jaen), donde destaca el Dr. Alfonso Ureña.

4. Bibliografía.

- Abásolo, C. and Gómez, M. (2002): A framework for meta-search based on numerical aggregation operators. In Procceedings of the Congrés Català d'Intel.ligència Artificial, CCIA 2002.
- Allan, J., Callan, J.P., Croft, W.B., Ballesteros, L., Byrd, D., Swan, R. & Xu, J. (1997). *Conference TREC*, 169-206.
- Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Essex, UK: Addison-Wesley.
- Bolc, L., Kowalski, A. & Kozlowska, M. (1985). A Natural Language Information Retrieval System with Extensions Towards fuzzy reasoning. *Int. J. of Man-Machine Studies*, 23, 335-367.
- Bordogna, G. & Pasi, G. (1993). A fuzzy linguistic approach generalizing boolean information retrieval: A model and its evaluation. *J. of the American Society for Information Science*, 44, 70-82.
- Bordogna, G. & Pasi, G. (1995). Linguistic aggregation operators of selection criteria in fuzzy information retrieval, *Int. J. of Intelligent Systems*, 10, 233-248.
- Bordogna, G., Carrara, P. Pasi, G. (1995). Fuzzy approaches to extend boolean information retrieval, P. Bosc, J. Kacprzyk Eds. Fuzziness in database management-systems, Germany: Physica-Verlag, 231-274.
- Brezeale, D. (1999). The Organization of Internet Web Pages Using Wordnet And Self-Organizing Maps. Masters thesis, University of Texas at Arlington, August 1999.
- Bruza, P.; McArthur, R. and Dennis S. (2000): Interactive Internet search: Keyword, directory and query reformulation mechanisms compared. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 280 – 287.
- Callan, J.P., Croft, W.B. & Stephen, M. (1992). The INQUERY Retrieval System. DEXA 1992, 78-83.
- Chen, H. (1995). Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. *J. of the American Society for Information Science*, 46(3), 194-216.
- Chen, H., Shankaranarayanan, G., She, L. & Iyer, A. (1998). A Machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms, and Simulated Annealing. *J. of the American Society for Information Science*, 49(8), 693-705.
- Choi, D. (2001). Integration of document index with perception index and its application to fuzzy query on the Internet, Proceedings of the BISC Int. Workshop on Fuzzy Logic and the Internet, 68-72.
- Cooley, R., Mobasher, B., Srivastaba, J. (1997). Grouping web page references into transactions for mining world wide web browsing patterns, Technical report TR 97-021, University of Minnesota, Minneapolis.

- Cordón, O., Herrera-Viedma, E., Moya, F., Luque, M. & Zarco, C. (2002a). Uso de un Esquema de Nichos en un Algoritmo GA-P para el Aprendizaje Automático de Consultas Booleanas Extendidas. *Actas del Primer Congreso Español de Algoritmos Evolutivos y Bioinspirados (AEB'02)*, 60-66.
- Cordón, O., Herrera-Viedma, E., Moya, F., Luque, M. & Zarco, C. (2002b). An Inductive Query by Example Technique for Extended Boolean Queries Based on Simulated-Annealing Programming. *Proc. of the Seventh International Society for Knowledge Organization Conference (ISKO'02)*, pp. 429-436.
- Crestani, F. & Pasi G. (2000). *Soft Computing in Information Retrieval: Techniques and Applications*. Physica-Verlag (A Springer-Verlag Company), New York.
- Delgado, M., Martin-Bautista, M.J., Sánchez, D., Serrano, J.M., Vila, M.A. (2003). Association rules and fuzzy associations rules to find new query terms, Proc. of the Third Conference of the EUSFLAT, 49-53.
- Deerwester, Scott, Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990). Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6), 391-407.
- Drucker, H., Shahrary, B. & Gibbon, D.C. (2002). Support Vector Machines: Relevance Feedback and Information Retrieval. *Information Processing & Management*, 38, 305-323.
- Efthimiadis, E. N. (1996). Query Expansion. ARIST, v31, 121-187.
- Ellis, D. (1996). *Progress & Problems in Information Retrieval*. London: Library Association Publishing.
- Fernandez, S. (2001). A contribution to the automatic processing of the synonymy using Prolog, PhD Thesis, University of Santiago de Compostela, Spain.
- Fernández, S., Sobrino, A.(2000). Hacia un tratamiento computacional de la sinonimia. In Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), Nº 26, September 2000, 89-95.
- Glover, E. J.; Lawrence, S.; Birming- ham, W. P. & Giles C. L. (1999): Architecture of a metasearch engine that supports user information needs. In Eighth International Conference on Information and Knowledge Management (CIKM'99) ACM, 210 216.
- Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J. (1998). Indexing with WordNet synsets can improve retrieval, Proc. of the COLING/ACL Work. on usage of WordNet in natural language processing systems.
- Gordon, M.D. (1991). User-Based Document Clustering by Redescribing Subject Descriptions with a Genetic Algorithm. *J. of the American Society for Information Science*, 42(5), 311-322.
- Harman, D. K. (1992). Relevance feedback and other query modification techniques. En: Frakes, W.B.; Baeza-Yates, R. (Eds.). *Information Retrieval: Data Structures and Algorithms*. NJ: Prentice Hall, 241-263.
- Harman, D. K. (1998). "Text REtrieval Conference -TREC's-: Providing a Test-Bed for Information Retrieval Systems". *Bulletin of the American Society for Information Science*, Apr/May, 11-13.

- Herrera, F., Herrera-Viedma, E. & Martínez, L. (2000). A Fusion Approach for Managing Multi-Granularity Linguistic Term Sets in Decision Making. *Fuzzy Sets and Systems*, 114, 43-58.
- Herrera, F., Lozano, M. & Verdegay, J.L. (1998). Tackling Real-Coded Genetic Algorithms: Operators and tools for the Behaviour Analysis. *Artificial Intelligence Review*, 12, 265-319.
- Herrera-Viedma, E. (2000a). An Information Retrieval System with Ordinal Linguistic Weighted Queries Based on Two Weighting Semantics. *Proc. Of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Bases Systems* (IPMU'2000), Vol. I, 454-461.
- Herrera-Viedma, E. (2000b). Modeling the Query Subsystem of a Linguistic IRS for Expressing Qualitative and Quantitative Restrictions on Query Terms. *Actas del X Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF´2000)*, 181-186.
- Herrera-Viedma, E. (2001a). Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach. *Journal of the American Society for Information Science and Technology*, 52(6), 460-475.
- Herrera-Viedma, E. (2001b). An information retrieval system with ordinal linguistic weighted queries based on two weighting elements. *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, 9, 77-88.
- Herrera-Viedma, E., Cordón, O., Herrera, J.C. & Luque, M. (2002). An IRS Based on Multi-granular Linguistic Information. *Proc. Of the 7th International Conference of the International Society for Knowledge Organization (ISKO'02)*, 372-378.
- Herrera-Viedma, R., Pasi, G. (2003). Fuzzy approaches to access information on the Web: recent developments and research trends, Proc. of the Third Conference of the EUSFLAT, 25-31.
- Holland, J.H. (1975). *Adaption in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press.
- Horng, J-T. & Yeh, C-C. (2000). Applying Genetic Algorithms to Query Optimiz ation in Document Retrieval. *Information Processing & Management*, 36, 737-759.
- Howard, L. & D´Angelo, D. (1995). The GA-P: a Genetic Algorithm and GeneticProgramming Hybrid. *IEEE Expert* , 11-15.
- Ide, E. (1971). New Experiments in Relevance Feedback. In: Salton, G. (Ed.) *The SMART Retrieval System*. Englewood Cliffs, N.J.: Prentice-Hall, 337-54.
- Ingwersen, P. (1996). Cognitive perspective of information retrieval interaction: elements of a cognitive IR theory. *J. of Documentation*, 52 (1), 3-50.
- Kerschberg, L.; Kim, W. and Scime, A. (2001): A Semantic Taxonomy-Based Personalizable Meta-Search Agent. Second International Conference on Web Information Systems Engineering (WISE'01).
- King-Ip, L., Ravikumar, K. (2001). A similarity-based soft clustering algorithm for documents, Proc. of the Seventh Int. Conf. on Database Sys. for Advanced Applications.

- Kiryakov, A.K., Simov, K.I. (1999). Ontologically supported semantic matching, Proceedings of "NODALIDA'99: Nordic Conference on Computational Linguistics", Trondheim.
- Korfhage, R. R. (1997). *Information Storage and Retrieval.* New York: Wiley Computer Publishing.
- Kraft, D.H., Bordogna, G. & Pasi, G.(1994). An extended fuzzy linguistic approach to generalize Boolean information retrieval, *Information Sciences*, 2, 119–134.
- Kraft, D.H., Petry, F.E., Buckles, B.P., & Sadasivan, T. (1997). Genetic Algorithms for Query Optimization in Information Retrieval: Relevance Feedback. En: Sanchez, E., Shibata, T., Zadeh, L.A., (Eds.). *Genetic Algorithms and Fuzzy Logic Systems. Soft Computing Perspectives*, 155-173.
- Krovetz, R.: Viewing morphology as an inference process, in Proceedings of ACM SIGIR Conference, 16, 1993, pp. 191-202
- Lafourcade, M., Prince, V. (2001). Relative Synonymy and conceptual vectors, Proceedings the Sixth Natural Language Processing Pacific Rim Symposium, Japan, (202), 127-134.
- Leacock, C., Chodorow, M. (1998), Combining local context and Wordnet similarity for word sense disambiguation, In WordNet, an Electronic Lexical Database, MIT Press, Cambridge Ma, 285-303.
- Kowalski, G.: Information retrieval systems: Theory and Implementation. Kluwer Academic Publishers. 1997.
- Li, Z.; Wang, Y. and Oria, V. (2001): A New Architecture for Web Meta-Search Engines, in Proceedings of the 2001 Americas Conference on Information Systems, Boston, MA, 2001.
- López-Pujalte, C., Guerrero, V.P & Moya, F. (2002a). A Test of Genetic Algorithms in Relevance Feedback. *Information Processing and Management* , 38 (6), 793-805.
- López-Pujalte, C., Guerrero, V.P & Moya, F. (2002b). Order-Based Fitness Functions for Genetic Algorithms Applied to Relevance Feedback. *Journal of the American Society for Information Science and Technology*. Por aparecer.
- López-Pujalte, C., Guerrero, V.P. & Moya, F. (2002c). Funciones de Adaptación Óptimas para Implementar la Retroalimentación por Relevancia en los Sitemas de Recuperación de Información Actuales. *Actas del I Congreso Español de Algoritmos Evolutivos y Bioinspirados* , 67-72.
- López-Pujalte, C.; Guerrero, V.P. & Moya, F. (2002d). Evaluation of the Application of Genetic Algorithms to Relevance Feedback. Proc. of 7th International ISKO Conference, 422-428.
- López-Pujalte, C., Guerrero, V.P & Moya, F. (2002e). Genetic Algorithms in Relevance Feedback: A Second Test and New Contributions. *Information Processing and Management.* Por aparecer.
- Loupy, C., El-Bèze, M. (2002). Managing synonymy and polysemy in a document retrieval system using WordNet, Proceedings of the LREC2002: Workshop on Linguistic Knowledge Acquisition and Representation.

- Martín-Bautista, M.J.; Vila, M.A. & Larsen, H.L. (1999). A Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent. *J. of the American Society for Information Science*, 50(9), 760-771.
- Martin-Bautista, M.J., Vila, M., Kraft, D., Chen, J. (2001). User profiles and fuzzy logic in web retrieval, Proc. of the BISC Int. Workshop on Fuzzy Logic and the Internet, 19-24.
- Miller, G. (ed.) (1990). WORDNET: An Online Lexical Database. International Journal of Lexicography, 3(4).
- Miller, G.A. (1995). WordNet: A lexical database for English, Communications of the ACM 11, 39-41.
- Miyamoto, S., Miyake, T. Nakayama, K. (1983). Generation of pseudothesaurus for information retrieval based on co-occurrences and fuzzy set operations, IEEE Transactions on Systems, Man and Cybernetics, Vol 13, 1, 62-70.
- Miyamoto, S. (1990). Information retrieval based on fuzzy associations, Fuzzy Sets and Systems, 38 (2), 191-205.
- Miyamoto, S. (1990). *Fuzzy Sets in Information Retrieval and Cluster Analysis*. Kluwer Academic Publishers.
- Ohgaya, R., Takagi, T., Fukano, K., Taniguchi, K. (2002). Conceptual fuzzy sets- based navigation system for Yahoo!, Proc. of the 2002 NAFIPS annual meeting, 274-279.
- Olivas, J.A., Garcés, P.J., Romero, F.P. (2003). An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents, Int. Journal of Approximate Reasoning 34, 201-219.
- Pasi, G. (2002). Flexible information retrieval: some research trends, Mathware and Soft Computing 9, 107-121.
- Perkovitz, M., Etzioni, O. (2000). Towards adaptive web sites: Conceptual framework and case study, Artificial Intelligence 118, 245-275.
- Porter, M. F. (1980). An Algorithm for Suffix Stripping. Program, 14(3): 130-137.
- Ramakrishnan, G., Prithviraj, BP., Deepa E. et al, (2004). Soft word disambiguation, Second Global WordNet Conference, 2004.
- Ricarte, I., Gomide, F. (2001). A reference model for intelligent information search, Proc. of the BISC International Workshop on Fuzzy Logic and the Internet, 80-85.
- Robertson, A.M. & Willet, P. (1994). Generation of Equifrequent Groups of Words Using a Genetic Algorithm. *J. of Documentation*, 50(3), 213-232.
- Robertson, A.M. & Willet, P. (1996). An Upperbound to the Performance of Ranked-Output Searching: Optimal Weighting of Query Terms Using a Genetic Algorithm. *J. of Documentation*, 52(4), 405-420.
- Rocchio, J.J. (1971) Relevance Feedback in Information Retrieval. En: Salton, G. (Ed.) *The SMART Retrieval System: Experiments in Automatic Processing*. Englewood Cliffs, NJ: Prentice-Hall, 313-323.
- Romero, F. P.; Garcés, P.; Olivas, J. A. (2002). Improving Internet Intelligent Agents using Fuzzy Logic and Data Mining Techniques. Proc. of the International Conference on Artificial intelligence IC-AI'02, 225-230.

- Salton, G., Wang, A., Yang, C.S.A. (1975). Vector space model for automatic indexing, Communic ations of the ACM 18, 613-620.
- Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *J. of the American Society for Information Scienc e*, 41(4), 288-297.
- Salton, G. & McGill, M.H. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.
- Salton, G. (1989). *Automatic Text Processing - The Transformation, Analysis and Retrieval of Information by Computer.* Addison Wesley Publishing Company.
- Sánchez, L., Couso, I. & Corrales, J.A. (2001). Comparing GP Operators with SA Search to Evolve Fuzzy Rule Classifiers. *Information Sciences*, 136(1-4), 175-192.
- Sánchez, E., Miyano, H. & Brachet, J.P. (1995). Optimization of Fuzzy Queries with Genetic Algorithms. Application to a Data Base of Patents in Biomedical Engineering. *Proc. Sixth IFSA World Congress*, Vol. II, 293-296.
- Serrano-Guerrero, J.; Olivas, J. A. (2004). Discovery of conceptual relations for ontology construction in GUMSe. 2004 Annual Meeting of the North American Fuzzy Information Processing Society Proceedings - NAFIPS, IEEE, Banff, Alberta, Canada, 647 - 651.
- Smith, M.P. & Smith, M., (1997). The Use of Genetic Programming to Build Boolean Queries for Text Retrieval Through Relevance Feedback. *J of Information Science*, 23:6,423-431.
- Spark Jones, K. A.: Information Retrieval and Artificial Intelligence, Artificial Intelligence, 114, 1999, pp. 257-281.
- Spink, A. (2002). A user-centered approach to evaluating human interaction with web search engines: an exploratory study. *Information Processing & Management*, 38 401-426.
- Sun, J., Shaban, K., Poddre, S., Karry, F., Basir, O., Kamel, M. (2003). Fuzzy Semantic measurement for synonymy and its application in an automatic question-answering system, IEEE Int. Conference on National Language Processing and Knowledge Engineering, Beijing.
- Tang, Y., Zhang, Y. (2001). Personalized library search agents using data mining techniques, Proceedings of the 1st BISC Int. Workshop on Fuzzy Logic and the Internet, 119-124.
- Takagi, T., Tajima, M. (2001). Proposal of a search engine based on conceptual matching of text notes, Proc. of the BISC Int. Workshop on Fuzzy Logic and the Internet, 53-58
- Van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd ed. London: Butterworths.
- Voorhees, E. (1998). Using WordNet for text retrieval, In WordNet: an electronic lexical database, MIT Press.
- Vrajitoru, V. (1998). Crossover Improvement for the Genetic Algorithm in Information Retrieval. Information Proces sing & Management, 34(4), 405-415.
- Whaley, J.M. (1999). An application of word sense disambiguation to information retrieval, Dartmouth College Computer Science Technical Report PCS-TR99-352.
- Widyantoro, D., Yen, J. (2001). Incorporating fuzzy ontology of term relations in a search engine, Proceedings of the BISC Int. Workshop on Fuzzy Logic and the Internet, 155-160.

- Yager, R.R. (1988). On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making. *IEEE Trans. on Systems Man and Cybernetics*, 18(1), 183-190.
- Yang, J.J. & Korfhage, R.R. (1994). Query Modification Using Genetic Algorithms in Vector Space Models. *Int. J. of Expert Systems*, 7(2), 165-191.
- Xie, H. (2002). Patterns between interactive intentions and information-seeking strategies. *Information Processing & Management*, 38(1), 55-77.
- Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, 8, 338-353.
- Zadeh, L. A. (1975). The Concept of a Linguistic Variable and its Applications to Approximate Reasoning. Part I, *Information Sciences*, 8, 199-249. Part II, *Information Sciences*, 8, 301-357. Part III, *Information Sciences*, 9, 43-80.
- Zadeh, L. A. (1987). Fuzzy Sets and Applications (Selected Papers, edited by R. R. Yager, S. Ovchinnikov, R. M. Tong, H. T. Nguyen), John Wiley, Nueva York.
- Zadeh, L. A. (2003). From search engines to Question-Answering System: The need for new tools, E. Menasalvas, J. Segovia, P.S. Szczepaniak (Eds.): Proceedings of the Atlantic Web Intelligence Conference - AWIC'2003. Lecture Notes in Computer Science (LNCS), Springer.
- Zamir, O., Etzioni, O. (1999). Grouper: A dynamic clustering interface to web search results, Proceedings of the WWW8.

OTRAS REFERENCIAS (ESTADO DEL ARTE).

[Aaslberg, 92] Aalbersberg, I. J.: Incremental relevance feedback, *in Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen*, 1992, pp. 11-22.

[Abiteboul, 97] Abiteboul, S.; Quass, D.; McHugh, J.; Widom, J.; Wiener, J. L.: The lorel query language for semiestructured data *International Journal on Digital Libraries*, 1(1), 1997, pp. 68-88.

[Agirre, 95] Agirre, E.; Rigau, G.: A Proposal for Word Sense Disambiguation using Conceptual Distance, *in Proceedings of the First International Conference on Recent Advances in Natural Language Processing*, Velingrad, 1995.

[Agirre, 96] Agirre, E.; Rigau, G.: Word Sense Disambiguation Using Conceptual Density, *in Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*, Copenhagen, Denmark, 1996, pp. 16-22.

[Agirre, 00]Agirre, E.; Ansa, O.; Hovy, E.; Martínez, D.: Enriching very large ontologies using the WWW, *in Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI), 2000.*

[Agirre, 03] Agirre, E. and Lopez de Lacalle, O.: Clustering WordNet Word Senses, *in Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP '03)*, Borovets, Bulgary. 2003.

[Agrawal, 98] Agrawal, R.; Chakrabarti, S.; Dom, B.; Raghavan, P.: Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies, *VLDB Journal*, 1998.

[Armstrong, 95] Armstrong, R.; Freitag, D.; Joachims, T.; Mitchell, T.: WebWatcher: A Learning Apprentice for the World Wide Web, *in Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed environments*, 1995.

[Aslam, 01] Aslam, J. A.; Montague, M.: Models for Metasearch, *in Proceedings of 24th Internationsl ACM SIGIR*, New Orleans, Louisiana, USA, 2001, pp. 276-283.

[Baeza-Yates, 99] Baeza-Yates, R.; Ribeiro-Neto, B.: Modern Information Retrieval, *ACM Press*, 1999.

[Baeza-Yates, 00] Baeza Yates, R.: Desenredando La Madeja, *NOVATICA*, May-jun 2000, 25 aniversario, 2000, pp. 72-77.

[Banerjee, 02] S. Banerjee and T. Pedersen. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, *in Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, 2002.

[Belkin, 82] Belkin, N. J.; Oddy, R. N.; Brooks, H. M.: ASK for information retrieval: Part I. Background and theory, *Journal of Documentation*, 38(2), 1982, pp. 61-71.

[Belkin, 82] Belkin, N. J.; Oddy, R. N.; Brooks, H. M.: ASK for information retrieval: Part II. Results of a design study, *Journal of Documentation*, 38(3), 1982, pp. 145-164.

[Berners-Lee, 01] Berners-Lee, T.; Hendler, J.; Lassila, O.: The Semantic Web, Scientific American, 2001.

[Berzal, 02] Berzal, F.; Martín-Bautista, M. J.; Vila, M. A.; Blanco, I. J.: La lógica difusa en el proceso de recuperación de información. *ESTYLF 2002*. León, 2002.

[Billerbeck, 03] Billerbeck, B.; Scholer, F.; Williams, H. E.; Zobel, J.: Query Expansion using Associated Queries, *in Proceedings of the CIKM International Conference on Information and Knowledge Management*, O. Frieder, J. Hammer, S. Quershi, and L. Seligman (eds), New Orleans, Louisiana, 2003, pp. 2-9.

[Billsus, 96] Billsus, D.; Pazzani, M.: Revising User Profiles: The Search for Interesting Web Sites, *in Proceedings of the Third International Workshop on Multistrategy Learning (MSL '96)*, AAAI Press, 1996.

[Blair, 90] Blair, D.C.: Language and representation in information retrieval, Amsterdam, *Elsevier Science Publishers*, 1990.

[Blanco, 03] Blanco Gómez, M.: Estudio de buscadores, 2003.

[Bollacker, 00] Bollacker, K. D.: Discovering Relevant Scientific Literature on the Web, *IEEE Intelligent Systems*, Vol. 15, No. 2, 2000, pp. 42-47.

[Bordogna, 96] Bordogna, G.; Pasi, G.: Controlling Retrieval through a User-Adaptive Representation of documents, *International Journal of Approximate Reasoning* 12, 1996, pp. 317-399.

[Bowman, 94] Bowman, C. M.; Danzig, P. B.; Hardy, D. R.; Manber, U.; Schwartz, M. F.: The Harvest information discovery and access system, *in Proceedings 2nd International WWW Conference*, 1994, pp. 763-771.

[Boy, 94] Boy, G.: Interface Agents for Handling Fuzzy Descriptors in Information Retrieval, *Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'94*, Paris, Francia, 1994.

[Brezeale, 99] Brezeale, D.: The Organization of Internet Web Pages Using WordNet and Self-Organizing Maps, *Master of Science in Computer Science and Engineering*, University of Texas at Arlington, 1999.

[Brickley & Guha, 02] Brickley, D.; Guha, R. V.: Resource Description Framework (RDF) Schema Specification 1.0: W3C Working Draft. 2002.

[Brin, 98] Brin, S.; Page, .: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *in Proceedings of 7th WWW Conference*, 1998.

[Brown, 94] Brown, C. M.; Danzig, P. B.; Hardy, D.; Manber, U.; Schwartz, M. F.: The Harvest information discovery and access system, *in Proceedings of the Second International World Wide Web Conference*, 1994, pp. 763-771.

[Bruza, 93] Bruza, P. D.: Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia, *PhD thesis*, University of Nijmegen, Nijmegen, The Netherlands, 1993.

[Bruza, 00] Bruza, P.; McArthur, R.; Dennis, S.: Interactive Internet search: Keyword, directory and query reformulation mechanisms compared, *Special Interest Group on Information Retrieval (SIGIR)*, 2000.

[Budanitsky, 01] Budanitsky, A.; Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, *in Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, June 2001.

[Buono, 02] Buono, P.; Costabile, M. F.; Guida, S.; Piccinno, A.; Tesoro, G.: Integrating User Data and Collaborative Filtering in a Web Recommendation System, *in "Hypermedia: Openness, Structural Awareness, and Adaptivity"*,

[Cacheda, 01] Cacheda, F.; Viña, A.: Understanding how people use search engines: a statical analysis for e Business, 2001.

[Cacheda, 01] Cacheda, F.; Viña, A.: Simulación para la Evaluación de Sistemas de Recuperación de Información en el WWW, *Actas del Primer Congreso Iberoamericano de Telematica (CITA 2001)*, Cartagena de Indias, Colombia, 2001.

[Cañas, 03] Cañas, A. J.; Valerio, A.; Lalinde-Pulido, J.; Carvalho, M.; Arguedas, M.: Using WordNet for Word Sense Disambiguation to Support Concept Map Construction, *SPIRE 2003*, Manaus, Brazil, 2003.

[Casasola, 97] Casasola, E.; Gauch, S.: Intelligent Information Agent for the World Wide Web, *Information and Telecommunication Technology Center*, Technical Report: IITTC-FY97-11100-1, 1997.

[Chakrabarti, 99] Chakrabarti, S.; Dom, B.; Gibson, D.; Kleinberg, J.; Kumar, S.; Raghavan, P.; Rajagopalan, S.; Tmkins, A.: Mining the link structure of the world wide web. *IEEE Computer*, 32(8), 1999, pp. 60-67.

[Chklovski, 03] Chklovski, T. and Mihalcea, R. (): Exploiting Agreement and disagreement of Human Annotators for Word Sense Disambiguation, *in Proceedings of Recent Advances In NLP (RANLP 2003)*, September 2003.

[Chen, 94] Chen, H.: Collaborative Systems: Solving the Vocabulary Problem, *Computer*, Vol. 27, No. 5, 1994, pp. 58-66.

[Chen, 00] Chen, Z.: WebSail: from on-line learning to Web search, *in Proceedings of the First International Confererence on Web Information Systems Engineering*, Vol. 1, 2000, pp. 206-213.

[Chidlovskii, 00] Chidlovskii, B.; Glance, N. S.; Grasso, M. A.: Collaborative Re-Ranking of Search Results, *in Proceedings of AAAI-2000 Workshop on AI for Web Search*, 2000.

[Claypool, 01] Claypool, M.; Le, P.; Wased, M.; Brown, D.: Implicit Interest Indicators, *in Proceedings of the 6 th International Conference on Intelligent User Interfaces (IUI '01)*, USA, 2001, pp. 33-40.

[Cleverdon, 66] Cleverdon, C.W.; Mills, J.; Keen, M.: Factors Determining the Performance of Indexing Systems, Volume I - Design, Volume II - Test Results, *ASLIB Cranfield Project*, Cranfield, 1966.

[Collins, 75] Collins, A. M.; Loftus, E. F.: A spreading activation theory of semantic processing, *Psychological Review*, 82(6), 1975, pp. 407-428.

[Cooper, 73] Cooper, W.S.: On selecting a Measure of Retrieval Effectiveness, Journal of the American Society for Information Science, Vol. 24, 1973. pp. 87-92.

[Cornella, 98] Cornella, A.: La importancia de la "Relevancia" en Información, *Extra!-Net*, 1998. http://intranet.logiconline.org.ve/Techinfo/relevancia.html (Abril 2004).

[Creighton, 04] http://www.hsl.creighton.edu/hsl/Searching/Recall-Precision.html (Abril, 2004).

[Crestani, 98] Crestani, F.; Lalmas, M.; Campbell, I.; van Risbergen, C. J.: Is this document relevant? ...probably. A survey of probabilistic models in information retrieval. *ACM Computing Surveys*, 30(4), 1998, pp. 528-552.

[Crestani, 99] Crestani, F.; Pasi, G.: Soft Information Retrieval: Applications of Fuzzy Set Theory and Neural Networks, in: N. Kasabov and R. Kozma, editors, *Neuro-Fuzzy Techniques for Intelligent Information Systems*, Physica Verlag (Springer Verlag), Heidelberg, Germany, 1999, pp. 287-315.

[Delgado, 98] Delgado Domínguez, A. M.: Mecanismos de Recuperación de Información en la WWW, *Memoria de Investigación para la obtención de la Suficiencia Investigadora*, Doctorado en Informática, 1998.

[Dillon, 80] Dillon, M.; Desper, J.: The use of Automatic Relevance Feedback in Boolean Retrieval Systems, *Journal of Documentation*, 36(3), 1980, pp. 197-208.

[Dhillon, 01] Dhillon, I. S.: Co-clustering documents and words using Bipartite Spectral Graph Partitioning, *UT CS Technical Report # TR 2001-05*, 2001.

[Denning, 03] Denning, A.: SEO (Search Engine Optimization): A Case Study, The Beach Trail Cottages, *Internet Marketing Newsletter*, 2003.

[Dombi, 82] Dombi, J.: A general class of fuzzy operators, the De Morgan class of fuzzy operators and fuzziness measures induced by fuzzy operators, *Fuzzy Sets and Systems* **8**,1982, pp. 149-163.

[Doorenbos, 96] Doorenbos, R. B.; Etzioni, O.; Weld, D. S.: A scalable comparison-shopping agent for the world-wide web, *Technical Report 96-01-03*, University of Washington, Department of Computer Science and Engineering, 1996.

[Edmonds, 02] Edmonds, P.: SENSEVAL: The evaluation of word sense disambiguation Systems, *in the ELRA Newsletter*, Vol. 7 No. 3, 2002.

[Efthimiadis, 93] Efthimiadis, E. N.: A User-Centred Evaluation of Ranking Algorithms for Interactive Query Expansion, *SIGIR Forum* , 1993, pp. 146-159.

[Efthhimiadis, 96] Efthimiadis, E. N.: Query expansion, *in M. E. Williams (Ed), Annual Review of Information Science and Technology*, Vol. 31., 1996, pp. 121-187.

[Etzioni, 96] Etzioni, O.: The world wide web: Quagmire or gold mine. *Communications of the ACM* , 39(11), 1996, pp. 65-58.

[Fagin, 98] Fagin, R.: Fuzzy Queries in Multimedia Database Systems, *in proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Seattle, Washington, 1998.

[Fensel, 99] Fensel, D.: On2Broker: Semantic-Based Access to Information Sources at the WWW, *in Proceedings of the World Conference on the WWW and Internet (WebNet 99)*, Honolulu, Hawaii, USA, 1999, pp. 25-30.

[Fensel, 00] Fensel, D.; et al.: OIL in a Nutshell.

[Fensel, 02] Fensel, D.; Hendler, J.; Lieberman, H.; Wahlster, W.: Creating the Semantic Web, In D. Fensel et al. (eds.), *Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential, MIT Press*, Boston, 2002

[Fernandez, 98] Fernandez, M; Suciu, D.: UNQL: A query Language for Web Sites. http://www.cs.huji.ac.il/course/2003/sdbi/1998/yarivi/sdbi/

[Fernández Lanza, 01] Fernández Lanza, S.: Una contribución al procesamiento automático de la sinonimia utilizando Prolog. *Tesis doctoral*, Universidad de Santiago de Compostela, 2001.

[Figuerola, 00] Figuerola, C. G.; Gómez, R.; López de San Román, E.: Stemming and n-grams in spanish: an evaluation of their impact on information retrieval, *Journal of Information Science*, 26(6), 2000, pp. 461-467.

[Figuerola, 01] Figuerola, C. G.; Gómez, R.; Zazo Rodríguez, A. F.; Alonso Berrocal, F. L.: Stemming in Spanish: A First Approach to its Impact on Information Retrieval, CLEF 2001 Cross-Language System Evaluation Campaign,Darmstadt, Germany, 2001.

[Finin, 94] Finin T.; Fritzson, R; McKay, D.; McEntire R.: KQML as an Agent Communication Language, in proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94), 1994.

[Fipa, 97] Foundation For Intelligent Physical Agents Fipa 97 Specification Part Fipa 97, Version 1.0 Part 2 Fipa, 1997.

[Fipa, 01] Foundation for Intelligent Physical Agents: FIPA ACL Message Representation in XML Specification, 2001.

http://www.fipa.org/specs/fipa00071/XC00071C.html

[Frakes, 92] Frakes, W.B.; Baeza-Yates, R.: Information Retrieval: data structures and algorithms, *Englewood Cliffs: Prentice Hall*, 1992.

[Frigui, 00] Frigui, H.; Nasraoui, O.: Simultaneous clustering and attribute discrimination, *in Proceedings of FUZZIEEE*, San Antonio, 2000, pp. 158-163.

[Fugmann, 85] Fugmann, R.: The five axiom theory of indexing and information supply, *Journal of the American Society forInformation Science*, 36(2), 1985, pp. 116-129.

[Furnas, 87] Furnas, G. W.; Landauer, T. K.; Gomez, L. M.; Dumais, S. T.: The Vocabulary Problem in Human-System Communication, *Communications of the ACM*, Vol 30, No. 11, 1987, pp. 964-971.

[Garofalakis, 99] Garofalakis, M.; Rastogi, R.; Seshadri, S.; Shim, K.: Data Miing and the Web: Past, Present and Future, *in Proceedings of the Second International Whorkshop on Web Information and Data Management*, Kansas City, USA, 1999.

[Gils, 03] Van Gils, B.; Proper, H. A.; Van Bommel, P.; Schabell, E. D.: Profile-based retrieval on the World Wide Web, in Proceedings de Bra, editor, *in Proceedings of the Conferentie Informatiewetenschap (INFWET2003)*, Eindhoven, The Netherlands, EU, 2003, pp. 91-98.

[Glover, 99] Glover, E. J.; Lawrence, S.; Virmingham, W. P.; Giles, C. L.: Architecture of a Metasearch Engine that Supports User Information Needs, *in Proceedings of the Eighth International Conference on Information Knowledge Management, (CIKM-99)*, 1999, pp. 210-216.

[Gomez, 02] M. Gomez ,J. M. Abasolo. Improving meta-search by using query-weighting and numerical aggregation operators, in Proc. Information Processing and Magement of Uncertainty conference, IPMU 2002.

[Google Api, 04] Google Web APIs References –

http://www.google.com/apis/reference.html

[Gravano, 88] Gravano, L.; Papakonstantinou, Y.: Mediating and Metasearching on the Internet, Data Engineering 21(2), 1988, pp. 28-36.

[Greenberg, 01] Greenberg, J.: Automatic query expansion via lexical-semantic relation-ships, Journal of the American Society for Information Science and Technology, 52(5), 2001, pp. 402-415.

[Grosso, 99] Grosso, W.: Knowledge Modeling Tools: Challenges for Protégé-2000 in the Coming Years, *Twelfth Workshop on Knowledge Acquisition, Modeling and Management*, Banff, Alberta, Canada, 1999.

[Gruber, 93a] Gruber, T. R.: A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220, 1993.

[Gruber, 93b] Gruber, T. R.: Toward principles for the design of ontologies used for knowledge sharing, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, The Netherlands, Kluwer Academic Publishers, 1993.

[Guarino, 99] Guarino, N.: OntoSeek: content-based access to the Web, IEEE Intelligent Systems, Vol. 14, No. 3, 1999, pp. 70-80.

[Hammond, 95] Hammond, K.; Burke, R.; Martin, C.; Lytinen, S.: FAQ Finder: A case-based approach to knowledge navigation, in Working Notes of the AAAI Spring Symposium: Information Gathering from Heterogeneous, Distributed Environments, Stanford University, AAAI Press, 1995, pp. 69-73.

[Harman, 86] Harman, D.: An experimental study of factors important in document ranking, in Proceedings of the Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, 1986, pp. 186-193.

[Harman, 91] Harman, D.: How effective is suffixing?, Journal of the American Society for Information Science (JASIS), Vol. 42, No. 1, 1991, pp. 7-15.

[Harman, 95] Harman, D.: The TREC Conferences, in R. Kuhlen y M. Rittberger (Eds.): Hypertext – Information Retrieval – Multimedia: Synergieeffekte Elektronischer Informations Systeme, in Proceedings of HIM '95,. Universitätsverlag Konstanz, 1995, pp. 9-28.

[Herrera-Viedma, 0?] Herrera-Viedma, E.; Herrera, F:; Martínez, L.; Herrera, J. C.; López, A. G.: Incorporating Filtering Techniques in a Fuzzy Linguistic Multi-Agent Model for Information Gathering on the Web.

[Hofstede, 96] Ter Hofstede, A. H. M.; Proper, H. A.; Van Der Weide, P.: Query Formulation as an Information Retrieval Problem, The Computer Journal, Vol. 39, pp. 255-274, 1996.

[Hong Ding, 02] Hong Ding, C.; Buyya, R.: Guided Google: A Meta Search Engine and its Implementation using the Google Distributed Web Services.

[Huang, 00] Huang, L.: A Survey On Web Information Retrieval Technologies. 2000.

[Hull, 96] Hull, D.; Grefenstette, G.: A detailed analysis of English stemming algorithms, XEROX Technical Report,Rank Xerox Research Centre, 1996.

[Ide, 71] Ide, E.: New experiments in relevance feedback, in G. Salton, editor, The SMART Retrieval System, Prentice Hall, 1971, pp. 337-354.

[Ide y Véronis, 98] Ide, N.; Véronis, J.: Word Sense Disambiguation: The State of the Art, Computational Linguistics, 24(1), 1998.

[Indyk, 98] Indyk, P.; Chakrabarti, S.; Dom, B.: Enhanced hypertext categorization using hyperlinks, in Proceedings of ACM SIGMOD, 1998.

[Ingwersen, 92] Ingwersen, P.: Information Retrieval Interaction, Taylor Graham, London, 1992.

[Ingwersen, 95] Ingwersen, P.; Willet, P.: An introduction to algorithmic and cognitive approaches for information retrieval. Libri, 45, 1995, pp. 160-177.

[Jansen, 98] Jansen, B.; Spink, A.; Bateman, J.; Saracevic, T.: Real Life Information Retrieval: A Study Of User Queries On The Web, in Proceedings SIGIR FORUM Spring 98, 1998.

[Jirapanthong, 00] Jirapanthong W. & Sunetnanta T. (2000):An XML-Based Multi-Agents Model for Information Retrieval on WWW. Proceeding of the 4th National Computer Science and Engineering Conference (NCSEC2000),

Queen Sirikit National Convention Center (Organized by Chulalongkorn University), Bangkok, Thailand, November 16-17.

[Johnson, 03] Johnson, F. C.; Griffiths, J. R.; Hartley, R. J.: Task dimensions of user evaluations of information retrieval systems, Information Research, Vol. 8 No. 4, 2003.

[Kerschberg, 01] Kerschberg, L.; Kim, W.; Scime, A.: WebSifter II: A Personalizable Meta-Search Agent based on Semantic Weighted Taxonomy Tree, International Conference on Internet Computing (1), 2001, pp. 14-20.

[Kilgarriff, 97] Kilgarriff, A.: I don't believe in word senses, *Computers and the Humanities 31 (2)*, 1997, pp 91-113.

[Kim, 03] Kim, H-J.; Lee, S-G.: Building topic hierarchy based on fuzzy relations, Neurocomputing 51, 2003, pp. 481-486.

[Kleinberg, 98] Kleinberg, L.: Authoritative sources in a hyperlinked environment, in Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[Kobayashi, 00] Kobayashi, M.; Takeda, K.: Information Retrieval on the Web, ACM Computing Surveys, 2000.

[Konopnicki, 95] Konopnicki, D.; Shmueli, O.: W3QS: A query system for the World Wide Web, in Proceeding of the 21th VLDB Conference, Zurich, 1995, pp. 56-65.

[Korzyk, 00] Korzyk A. D. (2000): Towards XML As A Secure Intelligent Agent Communication Language. 23rd National Information Systems Security Conference. Page(s): 371 – 387. October 2000

[Kosala, 00] Kosala, R.; Blockeel, H.: Web Mining Research: A Survey. ACM SIGKDD'00, Vol2(1), 2000.

[Kowalsky, 97] Kowalski, G.: Information retrieval systems: Theory and Implementation. Kluwer Academic Publishers. 1997.

[Kramsch, 96] Kramsch, C.: Context and Culture in Language Teaching. Oxford University Press. 1996.

[Krovetz, 92] Krovetz, R.; Croft, W. B.: Lexical ambiguity in information retrieval, ACM Transactions on Information Systems, 10(2), 1992, pp. 115-141.

[Krovetz, 93] Krovetz, R.: Viewing morphology as an inference process, in Proceedings of ACM SIGIR Conference, 16, 1993, pp. 191-202.

[Kucera, 67] Kucera, H.; Francis, W. N.: Computational Análisis of Present-Day American English, Brown University Press, Providente, 1967.

[Kuhlthau, 88] Kuhlthau, C.: Longitudinal case studies of the information search process of users in libraries, Library and Information Science Research, 10, 3, 1988, pp. 257-304.

[Kummamuru, 03] Kummamuru, K.; Dhawale, A.; Krishnapuram, R.: Fuzzy Co-clustering of Documents and Keywords, in Proccedings of The IEEE International Conference on Fuzzy Systems, 2003, pp. 772-777.

[Kwok, 96]C. Kwok and D. Weld. Planning to gather information, in Proc. 14th National Conference on AI, 1996

[Lakshmanan, 96] Lakshmanan, L.; Sadri, F.; Subramanian, I. N.: A declarative language for querying and restructuring the web, in Proceedings 6th International Workshop on Research Issues in Data engineering: Interoperability of Nontraditional database systems (RIDE-NDS'96), 1996.

[Lam, 01] Lam, S.: The Overview of Web Search Engines. 2001.

[Lancaster, 73] Lancaster, F.W., and Fayen, E.G. (1973). Information Retrieval On-Line Los Angeles, CA: Melville Publishing Co. Chapter 6.

[Langley, 99] Langley, P.: User modeling in adaptive interfaces, in Proceedings of the Seventh International Conference on User Modeling, 1999, pp. 367-370.

[Langville, 04] Langville, A. N.; Meyer, C. D.: Deeper Inside PageRank. Accepted by *Internet Mathematics*, February 2004.

[Lassila & Swick, 99] Lassila, O.; Swick, R.: Resource Description Framework (RDF) Model and Syntax Specification: W3C Recommendation, 1999.

http://www.w3.org/2000/01/rdf-schema#

[Latiri, 03] Latiri, C. Ch.; Ben Yahia, S.; Chevallet, J. P.; Jaoua, A.: Query expansion using fuzzy association rules between terms, Fourth International Conference Journées de l'Informatique Messine JIM'2003, Metz , France, 2003.

[Lawrence, 99] Lawrence, S.; Giles, C. L.: Accessibility of Information on the Web, Nature, Vol. 400, 1999, pp. 107-109.

[Lee, 97] Lee, J. H.: Analyses of multiple evidence combination, in N. J. Belkin, A. D. Narasimhalu, and P. Willett, editors, Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, Pennsylvania, USA, 1997, pp. 267-275.

[Lempel, 00] Lempel, R.; Moran, S.: The stochastic approach for link-structure analysis (salsa) and the TKC effect, in Proceedings 9th International World Wide Web Conference, 2000.

[Lempel, 04] Lempel, R.; Moran, S.: Rank Stability and Rank Similarity of Link-Based Web Ranking Algorithms in Authority ConnectedGraphs, to Information Retrieval, special issue on Advances in Mathematics/Formal Methods in Information Retrieval, 2004

[Lempel, 01] Lempel, R.; Moran, S.: SALSA: The Stochastic Approach for Link-Structure Analysis, in Proceedings ACM Transactions of Information Systems 19(2), 2001, pp. 131-160.

[Li, 95] Li, X.; Szpakowicz, S.; Matwin, S.: A WordNet-based Algorithm for Word Sense Disambiguation, in Proceedings of IJCAI-95. Montréal, Canada, 1995.

[Li, 01] Li, Z.; Wang, Y.; Oria, V.: A New Architecture for Web Meta-Search Engines, in Proceedings of Seventh Americas Conference on Information Systems, 2001, pp. 415-422.

[Lifantsev, 98] Lifantsev, M.: Opengrid. 1998. http://www.ecsl.cs.sunysb.edu/~maxim/OpenGRiD/ (Abril 2004)

[Liu, 04] Liu, F.; Yu, C.; Meng, W.: Personalized Web Search For Improving Retrieval Effectiveness, IEEE Transactions on Knowledge and Data Engineering, Vol.16, No.1, 2004, pp.28-40.

[Lopez-Pujalte, 02] C. Lopez-Pujalte, V. P. Guerrero and F. De Moya. Retroalimentación por Relevancia: nueva perspectiva desde la programación evolutiva. I Jornadas de Tratamiento y Recuperación de la Información (JOTRI-2002). Valencia, 4 y 5 de julio de 2002.

[Lovins, 68] Lovins, J.B.: Development of a stemming algorithm, Mechanical Translation and Computational Linguistics 11(1-2), 1968, pp. 22-31.

[Luke, 96] Luke, S.; Spector, L.; Rager, D.: Ontology-Based Knowledge Discovery on the World-Wide Web. 1996.

[Lyons, 77] Lyons, J.: Semantics. 2 vols. Cambridge: Cambridue University Press, 1977.

[Maarek, 96] Maarek, Y. S.; Ben Shaul, I. Z.: Automatically organizing bookmarks per content, in Proceedings of 5th International World Wide Web Conference, 1996.

[McGuinness, 00]. McGuinness, D.; Fikes, R.; Rice, J.; Wilder, S.: The Chimaera ontology environment, *in Proceedings of the 17th National Conference on Artificial Intelligence*, 2000.

[Madria, 99] Madria, S.; Bhowmick, S. S.; Ng, W. K.; Lim, E.-P.: Research issues in Web Data Mining, in Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWak '99, 1999, pp. 303-312.

[Maedche, 00] Maedche, A., Staab, S.: Semi-automatic Engineering of Ontologies from Text, in Proceedings of the 12th International Conference on Software Engineering and *Knowledge Engineering*, 2000

[Maes, 94] Maes, P.: Agents that reduce work and information overload, Communications of the ACM, Vol. 37, No.7, 1994, pp. 30-40.

[Marcos, 98] Marcos Marín, F. A.; Satorre Grau, F. J.; Viejo Sánchez, M. L.: Gramática española. Madrid: Editorial Síntesis, Letras Universitarias 20, 1998.

[Martin, 00] Martin, P.; Eklund, P. W.: Knowledge retrieval and the World Wide Web, IEEE Intelligent Systems, Vol. 14, No. 3, 2000, pp. 18-25.

[Martín-Bautista, 01] Martín-Bautista, M. J.; Larsen, H. L.; Sánchez, D.; Vila, M-A.: Information Filtering and User Profile Construction with Fuzzy Sets and Genetic Algorithms, *Information Processing & Management*, 2001.

[Martín-Bautista, 02] Martín-Bautista, M. J.; Kraft, D. H.; Vila, M. A.; Chen, J.; Cruz, J.: User Profiles and Fuzzy Logic for Web Retrieval Issues, *Journal of Soft Computing*, v. 6, n. 5, 2002, pp. 365-372.

[Martínez, 01] Martínez Méndez, F. J.: Aproximación general a la evaluación de la recuperación de información por medio de los motores de búsqueda en Internet. Scire, Vol. 7 (1), 2001. http://www.um.es/gtiweb/fjmm/ibersid2000.PDF (Abril 2004).

[Martínez, 03] Martínez Méndez, F. J.; Rodríguez Muñoz, J.V.: Síntesis y crítica de las evaluaciones de la efectividad de los motores de búsqueda en la web, Information Research, 8(2), No. 148, 2003. http://InformationR.net/ir/8-2/paper148.html (Abril 2004)

[Martínez, 04] Martínez Méndez, F. J.: Aspectos de la evaluación de los sistemas de recuperación de información: necesidad y utilidad, Anales Vol. 8, 2004. http://www.um.es/gtiweb/fjmm/anales2004.pdf (Abril 2004).

[Masand, 00] Masand, B.; Spiliopoulou, M.: Webkdd-99: Workshop on web usage analysis and user profiling. SIGKDD Explorations, 1(2), 2000.

[Meng, 02] Meng, W.; Yu, C.; Liu, K-L.: Building Efficient and Effective Metasearch Engines, in Proceedings of ACM Computing Surveys, Vol. 34, No. 1, 2002, p. 48-89.

[Merialdo, 97] Merialdo, P.; Atzeni, P.; Mecca, G.: Semistructured and strucutred data in the web: Going back and forth, in Proceedings of the Workshop on the Management of Semistructured Data (in conjunction with ACM SIGMOD), 1997.

[Mihalcea, 98] Mihalcea, R.; Moldovan, D. I.: Word Sense Disambiguation based on Semantic Density, in *Proceedings of COLING-ACL '98 Workshop on Usage of WordNet in* Natural Language Processing Systems, Montreal, Canada, 1998, pp.16-22.

[Mihalcea, 00] Mihalcea, R.; Moldovan, D.: Semantic Indexing using WordNet Senses, *in Proceedings of ACL Workshop on IR & NLP*, Hong Kong, 2000.

[Miller, 90] Miller, G.: Special issue, WordNet: An on-line lexical database, Intl. Journal of Lexicography 3(4), 1990.

[Miniño, 03] Miniño, R.: Nociones de semántica. Español para comunicadores.

http://www.pucmmsti.edu.do/materias/espanolcom/ (Abril 2004).

[Mizzaro, 96a] S. Mizzaro. How many relevances in IR?, In *Proceedings of the Workshop `Information Retrieval and Human Computer Interaction', GIST Technical Report GR96-2, Glasgow University*, Glasgow, UK, 1996, pp. 57-60.

[Mizzaro, 96b] Mizzaro, S.: On the Foundations of Information Retrieval, in Proceedings of the conference *AICA'96*, Rome, 1996.

[Mmuruza, 04] http://cursos.pnte.cfnavarra.es/mmuruza1/ (Abril 2004).

[Mobasher, 03] Mobasher, B.; Dai, H.; Luo, T.; Sun, Y.; Zhu, J.: Integrating Web Usage and Content Mining for More Effective Personalization, *Managing data mining technologies in organizations: techniques and applications*, 2003, pp. 239-249.

[Molinari, 96] Molinari, A.; Pasi, G.: A fuzzy representation of HTML documents for Information Retrieval Systems, *in Proceedings of IEEE Int. Conf. On Fuzzy Systems*, New Orleans, 1996.

[Muggleton, 94] S. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. J. Logic Programming, 19(20):629{679, 1994.

[Muramatsu, 01] Muramatsu, J.; Pratt, W.: Transparent Queries: Investigating Users' Mental Models of Search Engines, *in Proceedings of SIGIR 2001*, New Orleans, LA , 2001.

[Nambiar, 03] Nambiar, U.; Kambhampati, S.: Answering imprecise database queries: a novel approach, *in Proceedings of the fifth ACM international workshop on Web information and data management, WIDM'03*, New Orleans, Louisiana, US, 2003.

[Nayak, 02] Nayak, R.; Witt, R.; Tonev, A.: Data Mining and XML Documents, *International Conference on Internet Computing*, 2002, pp. 660-666

[Nie, 02] Nie, J.; Jin, F.: Integrating Logical Operators in Query Expansion in Vector Space Model, *ACM SIGIR Workshop on Formal/Mathemetical Methods for Information Retrieval (MF/IR 2002)*, 2002.

[Nikravesh, 03] Nikravesh, M.; Takagi, T.: Web Intelligence: Concept-Based Web Search, FALL, 2003.

http://www.eecs.berkeley.edu/~sguada/cs199-299/Fall2003/GroupC/Book3WebIntelFi nal.pdf

[Nikravesh, 01] Nikravesh, M.; Azvine, B.: FLINT: New Directions in Enhancing the Power of the Internet, *UC Berckley Electronics Research Laboratory*, Memorandum No. UCB/ERL M01/28. 2001.

[Nikravesh, 02] Nikravesh, M.; Loia, V.; Azvine, B.: Fuzzy logic and the Internet (FLINT): Internet, World Wide Web, and search engines, *in proceeding of Soft Computing*, 6. 2002, pp. 287-299.

[O'Day, 93] O'Day, V. L.; Jeffries, R.: Information artisans: patterns of result sharing by information searchers, *in Proceedings ACM COOCS'93*, 1993, pp. 98-107.

[Odgen, 72] Odgen, C. K.: The Meaning of meaning. A Study on the Influence of Language upon Thought and of the Science of Symbolism, London: Routledge and Kegan Paul, 1972.

[Oh, 01] Oh, C. H.; Honda, K.; Ichihashi, H.: Fuzzy clustering for categorical multivariate data, *in proceedings of IFSA/NAFIPS*, Vancouver, USA, 2001, pp. 2154-2159.

[Ohgaya, 03] Ohgaya, R.; Shimmura, A.; Takagi, T.; Aizawa, A.: Meiji University Web and Novelty Track Experiments at TREC 2003, *in proceedings of 12th Text Retrieval Conference TREC'03*. 2003.

[Ohlms, 02] Ohlms, C.: The future of the Semantic Web. A Perspective on the Market Adoption of Semantic Web Technologies, 9. AIK- Symposium, 2002.

[Olivas, 03] Olivas, J.A.; Garcés, P.J.; Romero; F.P.: An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents, *Soft Computing Applications to Intelligent Information, Retrieval on the Internet*, Vol. 34, No. 2-3, 2003.

[Oostendorp, 94] Oostendorp, K. A.; Punch, W. F.; Wiggins, R. W.: A tool for individualizing the web, *in Proceedings 2nd International World wide Web Conference*, 1994.

[Page, 97] Page, L.; Brin, S.; Motowani, R.; Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web, *Stanford Digital Library working paper*, 1997.

[Paice, 90] Paice, C. D.: Another stemmer, *in Proceedings SIGIR 90*, 1990, pp. 56–61.

[Pazzani, 96] Pazzani, M.; Muramatsu, J.; Billsus, D.: Syskill & webert: Identifying interesting web sites, *in Proceedings AAAI Spring Symposium on Machine Learning in Information Access*, Portland, Oregon, 1996.

[Pazzani, 97] Pazzani, M.; Billsus, D.: Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning* 27, 1997, pp. 313–331.

[Peis, 03] Peis Redondo, E.; Hassan Montero, Y.; Herrera Viedma, E.; Herrera, J. C.: Ontologías, metadatos y agentes: recuperación "semántica" de la información, *II Jornadas de Tratamiento y Recuperacion de la Informacion (JOTRI 2003)*, 2003, pp.157-165.

[Porter, 80] Porter, M.: An algorithm for suffix stripping, *Program* 14(3), 1980, pp. 130-137.

[Proximity, 04] Google API Proximity Search (GAPS) –

http://www.staggernation.com/gaps/readme.html

[Quass, 95] Quass, D.; Rajaraman, A.; Sagiv, Y.; Ullman, J.; Widom, J.: Querying semistructured heterogeneous information, *International Conference on Deductive and Object Oriented Databases*, 1995.

[Quillian, 68] Quilliam, M. R.: Semantic Memory, Minsky, M. ed., 1968, *Semantic Information Processing*. Cambridge (MA), MIT Press, 1968, pp. 27-70.

[Resnik, 95] Resnik, P.: Disambiguating Noun Groupings with Respect to WordNet Senses, *in Third Workshop on Very Large Corpora. Association for Computational Linguistics*, 1995.

[Resnik, 97] Resnik, P.; Yarowsky, D.: A Perspective on Word Sense Disambiguation Methods and Their Evaluation, *in Proceedings of SIGLEX '97*, Washington, DC, 1997, pp. 79-86.

[Robertson, 76] Robertson, S E.; Spark Jones, K.: Relevance Weighting of Search Terms, *Journal of American Society for Information Science*, 27(3), 1976, pp. 129-146.

[Robertson, 90] Robertson, S. E.: On Term Selection for Query Expansion, *Journal of Documentation*, 45(4), 1990, pp. 359-364.

[Rocchio, 71] Rocchio, J. J.: Relevance Feedback in Information Retrieval, in Salton G. (ed.), The SMART Retrieval Storage and Retrieval System, N. J. Englewood Cliffs, *Prentice Hall*, Inc. 1971, pp. 313-323.

[Ruthven, 01] Ruthven, I.: Abduction, explanation and relevance feedback. *PhD Tesis*. Vol 1. 2001.

[Sakai, 01] Sakai, T.; Robertson, S. E.; Walker, S.: Flexible Pseudo-Relevance Feedback via Direct Mapping and Categorization of Search Request, *in Proceedings BCS-IRSG ECIR'01*, 2001, pp. 3-14.

[Salton, 83] Salton, G.; McGil, M. J.: Introduction to Modern Information Retrieval, *McGrawHill, Book Company*, New York, 1983.

[Salton, 83] Salton, G.; Fox, E. A.; Wu, H.: Extended Boolean information retrieval. *Communications of the ACM*, 26(11), 1983, pp. 1022-1036.

[Sanderson, 94] Sanderson, M.: Word Sense Disambiguation and Information Retrieval, *in proceedings of ACM-SIGIR*, 1994.

[Sanderson, 00] Sanderson, M.: Retrieving with good sense, *in Information Retrieval* Vol. 2 No. 1, 2000, pp. 49-69.

[Schäuble, 97] Schäuble, P.: Content-Based Information Retrieval from Large Text and Audio Databases, Section 1.6 Evaluation Issues, *Kluwer Academic Publishers*, 1997, pp. 22-29.

[Schütze, 95] Schütze, H.; Pedersen, J. O.: Information retrieval based on word senses, *in Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, USA, 1995, pp. 161-175.

[Selberg, 00] Selberg, E.; Etzioni, O.: On the instability of web search engines, *in Proceedings of Content-Based Multimedia Information Access (RIAO)*, Paris, France, 2000, pp. 223-235.

[Search, 04] Search Operators, http://www.ci.eugene.or.us/plweb/info/docs/operat.html (Abril 2004)

[Shavlik, 98] Shavlik, J.; Eliassi-Rad, T.: Building intelligent agents for web-based tasks: A theory-Refinement approach, *in Proceedings of the Conference on Automated Learning and Discovery: Workshop on Learning from Text and the Web*, Pittsburgh, PA, 1998.

[Shulman, 97] Shulman, H. G.: Memory and Cognition, 1997, http://www.psy.ohio-state.edu/psy312/semmem.html (Abril 2004).

[Sherman, 04] Sherman, C.: Metacrawlers and Metasearch Engines, 2004, http://searchenginewatch.com/links/article.php/2156241 (Abril 2004).

[Silverstein, 99] Silverstein, C.; Henzinger, M.; Marais, H.; Moricz, M.: Analysis of a Very Large Web Search Engine Query Log, *in Proceedings SIGIR FORUM Fall 99*, 1999.

[Slator, 87] Slator, B. M.; Wilks, Y. A.: Towards semantic structures forom dictionary entries, *in Proceedings of the 2nd Annual Rocky Mountain Conference on Artificial Intelligence*, Boulder, Colorado, 1987, pp. 85-96.

[Smeaton, 92] Smeaton, A. F.: Progress in the application of Natural Language Processing to Information Retrieval tasks, *The Computer Journal*, 35(3), 1992, pp. 268-278.

[Sneath, 73] Sneath, A.P.H.; Sokal, R. R.: Numerical Taxonomy – The Principles and Practice of Numerical Classification, *W. H. Freeman*, San Francisco, CA, 1973.

[Softnik, 04] Softnik Technologies: Google API Search Tool – http://www.searchenginelab.com/common/products/gapis/docs/

[Spark, 72] Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol. 28, No. 1, 1972, pp. 11-20.

[Spark Jones, 99] Spark Jones, K. A.: Information Retrieval and Artificial Intelligence, *Artificial Intelligence*, 114, 1999, pp. 257-281.

[Spiliopoulou, 96] Spiliopoulou, M.: Data mining for the web, in Principles of Data Mining and Knowledge Discovery, *Second European Symposium, PKDD '99*, 1999, pp. 588-589.

[Spink, 98] Spink, A.; Greisdorf, H.; Bateman, J.: From highly relevant to not relevant: examining different regions of relevance, *Information Processing and Management*, Vol. 34, No. 5, 1998, pp. 599-621.

[Spink, 01] Spink, A.; Wolfram, D.; Jansen, B. J.; Saracevic, T.: Searching the Web: The Public and Their Queries, *Journal of the American Society for Information Science and Technology*, 52(3), 2001, pp. 226–234.

[Stelmaszewska, 02] Stelmaszewska, H.; Blandford, A.: Patterns of interactions: user behaviour in response to search results, in A. Blandford & G. Buchanan (Eds.) *Proceedings JCDL Workshop on Usability*, 2002.

[Stenmark, 03] Stenmark, D.: Query Expansion Using an Intranet-Based Semantic Net, *in Proceedings of IRIS-26*, Porvoo, Finland, 2003.

[Sullivan, 03a] Sullivan, D.: Search Engine Sizes. *SearchengineWatch.com*, 2003. http://searchenginewatch.com/reports/article.php/2156481 (Abril 2004)

[Sullivan, 03b] D. Sullivan (2003): Search Engine Size Wars & Google's Supplemental Results. *SearchengineWatch.com*, Septiembre 2003.

http://searchenginewatch.com/searchday/article.php/3071371 (Abril 2004)

[Taghva, 98] Taghva, K.; Borsack, J.; Condit, A.: The Effectiveness of Thesauri-Aided Retrieval, *Technical Report 98-01, Information Science Research Institute*, University of Nevada, Las Vegas, 1998.

[Takagi, 95] Takagi, T.; Imura, A.; Ushida, H.; Yamaguchi, T.: Conceptual Fuzzy Sets as a Meaning Representation and theid Inductive Construction, *International Journal of Intelligent Systems*, V. 10, 1995, pp. 929-945.

[Takagi, 03] Takagi, T.: Concept-Based Information Retrieval and Search Engine. *BISC Distinguished Lecturer Series*, 2003.

[Terveen, 97] Terveen, L.; Hill, W.; Amento, B.; McDonald, D.; Creter, J.: Phoaks: a system for sharing recommendations, *Communications of the ACM*, Vol. 40, No. 3, 1997, pp. 59-62.

[Tomiyama, 03] Tomiyama, T.; Ohgaya, R.; Shinmura, A.; Kawabata, T.; Takagi, T.; Nikravesh, M.: Concept-Based Web Communities for Google™ Search Engine, *FALL*, 2003.

http://www.eecs.berkeley.edu/~sguada/cs199-299/Fall2003/GroupC/Fuzzy_IEEENikU SA(Corrected)_Final.doc.pdf (Abril 2004)

[Towell, 98] Towell, G.; Voorhees, E. M.: Disambiguating Highly Ambiguous Words, *Computational Linguistics*, Vol. 24, No. 1, 1998, pp. 125-145.

[Trier, 31] Trier, J.: Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachliches Feldes. *I. Von den Anfängen bis zum Beginn des* 13. Jahrhunderts. Heidelberg: Winter, 1931.

[Twidale, 96] Twidale, M. B.; Nichols, D. M.: Collaborative browsing and visualisation of the search process, *in Proceedings Aslib*, 48(7-8), 1996, pp. 177-182.

[Van Dijk, 99] Van Dijk, T.: Ideología. Una aproximación multiciplinaria. Barcelona, *Editorial Gedisa*, 1999.

[Van Rijsbergen, 77] Van Rijsbergen, C. J.: A Theoretical Basis for the Use of Co-Occurence Data in Information Retrieval, *Journal of Documentation*, 33, 1977, pp. 106-119.

[Van Rijsbergen, 79] Van Rijsbergen, C. J.: Information Retrieval, *2nd ed. Butterworth, London*, 1979.

[Van Setten, 00] Van Setten, M.; Moelaert-El Hadidy, F.: Collaborative Search and Retrieval: Collaboration in Information Retrieval, *GigaCE report, Telematica Instituut*, The Netherlands, 2000.

[Van Setten, 04] Van Setten, M.; Moelaert-El Hadidy, F.: Collaborative Search and Retrieval: Finding Information Together. *GigaCSCW software*.

https://doc.telin.nl/dscgi/ds.py/Get/File-8269/GigaCE-Collaborative_Search_and_Retrieval__Finding_Information_Together.pdf (Abril 2004).

[Velásquez, 04] Velásquez de la Cruz, J. M.: El lenguaje como vehículo comunicativo. *Universidad Abierta*.

http://www.universidadabierta.edu.mx/Biblio/V/Velazquez%20Jesus-Vehiculo.htm (Abril 2004).

[Vertexera, 04] Vertexera Inc.: Benefits of Search Engine Optimization. http://www.vertexera.com/wp/SEO.pdf (Abril 2004).

[Vogt, 99] Vogt, C. C.: Adaptive Combination of Evidence for Information Retrieval, *PhD thesis*, University of California, San Diego, 1999.

[Voorhees, 93] Voorhees, E. M.: Using WordNet to Disambiguate Word Sense for Text Retrieval, *in Proceedings ACM SIGIR'93*, Pittsburgh, 1993, pp. 171-180.

[Voorhees, 94a] Voorhees, E. M.: On expanding query vectors with lexically related word, *in Proceedings of the Second Text Retrieval Conference (TREC-2)*, D. K. Harman, ed., National Institute of Standards and Technology, (Gaithersburg, MD), 1994, pp. 223-231.

[Voorhees, 94b] Voorhees, E. M.: Query expansion using lexical-semantic relations, in W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 61-69.

[Wang, 04] Wang, Z.: Improved Link-Based Algorithms for Ranking Web Pages, *in Proceedings of The Fifth International Conference on Web-Age Information Management (WAIM 2004)*, Dalian, China, 2004.

[Weiss, 96] Weiss, R.; Velez, B.; Sheldon, M. A.; Namprempre, C.; Szilagyi, P.; Duda, A.; Gifford, D. K.: Hypursuit: a hierarchical network search engine that exploits content-link hypertext clustering, *in Hypertext096: The Seventh ACM Conference on Hypertext*, 1996.

[Wen, 01] Wen, J-R.; Nie, J-Y.; Zhang, H-J.: Clustering User Queries of a Search Engine, *in proceedings of the tenth international conference on World Wide Web table of contents*, Hong Kong , 2001, pp 162 - 168

[White, 01] White, R.; Jose, J. M.; Ruthven, I.: Query-Biased Web Page Summarisation: A Task-Oriented Evaluation. Poster Paper, *in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, New Orleans, USA, 2001.

[White, 02] White, R. ; Ruthven, I.; Jose, J. M.: The use of implicit evidence for relevance feedback in web retrieval, *in Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research (ECIR '02)*, Lecture Notes in Computer Science, Glasgow. 2002.

[Widyantoro, 01a] Widyantoro, D.; Yen, J.: Incorporating fuzzy ontology of term relations in a search engine, *in Proceedings of the BISC International Workshop on Fuzzy Logic and the Internet*, 2001, pp. 155-160.

[Widyantoro, 01b] Widyantoro, D. H.; Yen, J.: Using Fuzzy Ontology for Query Refinement in a Personalized Abstract Search Engine, *in Proceedings of Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, Vancouver, Canada, 2001

[Widyantoro, 01c] Widyantoro, D. H.; Yen, J.: A Fuzzy Ontology-based Abstract Search Engine and Its User Stuedies, In *the Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, Melbourne, Australia, 2001, pp. 1291-1294.

[Winship, 02] Winship, I.: World Wide Web searching tools - an evaluation, *Vine* No. 99, 2002, pp. 49-54.

http://gti1.edu.um.es:8080/javima/World-Wide-Web-searching-tools-an-evaluation.htm (Abril 2004).

[W3C, 98] W3C: The Query Language Position Paper of the XSL Working Group, *in Proceedings of the Query Language Workshop*, Massechusetts, 1998.

[W3C, 00] W3C: XSL Specification. Working Draft.

[XML, 00] Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation 6 October 2000.

http://www.w3.org/TR/REC-xml

[Yang, 94] Yang, Y.; Chute, C. G.: An example-based mapping method for text categorization and retrieval, *ACM Transaction on Information Systems (TOIS)*, 12(3), 1994, pp.252-277.

[Zadeh, 99] Zadeh, L. A.: From Computing with numbers to computing with words-from manipulation of measurements to manipulation of perceptions, *IEE Trans Circuit and Systems-I Fundamental Theory and Applications* 54(1), 1999, pp. 105-119.

[Zadeh, 01a] Zadeh, L.A.: The problem of deduction in an environment of imprecision, uncertainty, and partial truth, in: Nikravesh M, Azvine B (eds), FLINT 2001, New Directions in Enhancing the Power of the Internet, UC Berkeley Electronics Research Laboratory, 2001.

[Zadeh, 01b] Zadeh, L. A.: A New Direction in AI—Toward a Computational Theory of Perceptions. *AI Magazine* 22(1), 2001, pp. 73-84.

[Zeballos, 98] Zeballos, G.S.: Tools for efficient collaborative web browsing, in Churchill, E., Snowdon, D, Golovchinsky, G. (Eds.), *Proceedings of CSCW98 workshop on Collaborative and co-operative information seeking in digital information environment*, 1998.

[Zhang, 04] Zhang, Y-J.; Liu, Z-Q.: Refining Web Search Engine Results Using Incremental Clustering, International journal of intelligent systems, Volume N° 19, Fascicule N° 4, 2004.

5. Aportaciones y Propuestas.

- **Representación de documentos mediante el modelo FIS-CRM.**

El modelo FIS-CRM (*Fuzzy Interrelations and Synonymy based Concept Representation Model*) es un modelo de representación lógica del contenido de cualquier tipo de documentos, y se puede considerar como una extensión "borrosa" del modelo vectorial (*Vector Space Model – VSM–*) de representación de documentos. La diferencia fundamental de este modelo respecto al modelo VSM es que, mientras que en el modelo VSM los pesos de cada vector representan ocurrencias de términos, en el modelo FIS-CRM los pesos representan ocurrencias de conceptos. La primera gran utilidad que representa esta característica es la de abordar la comparación del contenido de diferentes documentos en base a los conceptos contenidos en los mismos. Y esta peculiaridad convierte a este modelo en una interesante herramienta en el campo de la Recuperación de Información.
El modelo FIS-CRM se sustenta en dos tipos de interrelaciones borrosas entre términos: la de sinonimia y la de generalidad, almacenadas en un diccionario de sinónimos y diferentes ontologías temáticas de términos respectivamente. De esta forma, se puede hablar del grado de sinonimia existente entre dos términos (que puede ser obtenido mediante expresiones del tipo Jaccard) y del grado de generalidad entre dos términos (obtenido a partir de la co-ocurrencia de términos en colecciones temáticas de documentos).
La forma de describir el papel que juegan estos dos tipos de interrelaciones en el modelo FIS-CRM se constata en sus dos premisas fundamentales:
Las apariciones u ocurrencias de una palabra en un documento se deben "repartir" entre los diferentes sinónimos de la palabra contenida, realizando el reparto en base al grado de sinonimia entre los sinónimos y la palabra contenida.
Si una palabra está contenida en un documento, las palabras que representan conceptos más generales que ésta deben tener un peso en el vector de dicho documento, que será proporcional a su grado de generalidad.
La construcción de los vectores de los documentos se realiza en dos fases: En primer lugar se parte de la representación vectorial estándar de los documentos (basada en ocurrencias de términos), y posteriormente se realiza un proceso de reajuste de los pesos de los elementos de los vectores, de forma que un término pueda tomar un peso en un vector aunque no aparezca en el documento, siempre y cuando el concepto al que represente dicho término subyaga en el mismo. En realidad, el proceso de reajuste se lleva a cabo en dos fases: reajuste por generalidad y reajuste por sinonimia.
El reajuste de pesos realizado en base a la interrelación de generalidad se basa en, partiendo de la aparición de un término 't' en un documento, otorgar un peso a los términos más generales a 't', que sea proporcional al grado de generalidad entre 't' y dichos términos.
El reajuste de pesos realizado en base a la interrelación de sinonimia se realiza mediante las siguientes expresiones:

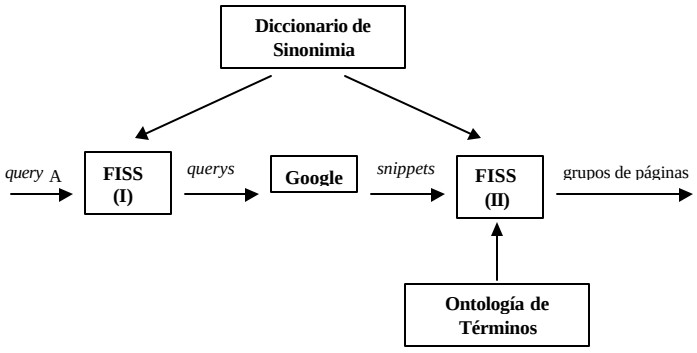
$$w'_i = N * \frac{1}{\sqrt{\sum_{i=1}^{n} m(t_i, C)^2}} * m(t_i, C) \qquad N = \sum_{i=1}^{m} w_i * m(t_i, C)$$

En estas expresiones w_i y w'_i representan el peso del término 'i' en el documento antes y después del reajuste respectivamente, C representa el concepto al cual convergen los

diferentes sinónimos encontrados en el documento y $m(t_i,C)$ representa el grado de pertenencia de cada uno de estos sinónimos al concepto en sí. A su vez, el grado de pertenencia de un término al concepto es obtenido a partir del grado de sinonimia entre ese término y el conjunto de sinónimos que conforman el concepto.
Es importante resaltar que en la última versión del modelo FIS-CRM las entradas del diccionario de sinónimos se realizan a nivel de término y acepción, por lo que para cada pareja término-acepción se dispone de un conjunto de sinónimos con sus correspondientes grados de sinonimia. Esta característica del diccionario es la que permita identificar la correcta acepción (*word sense disambiguation*) de los términos polisémicos (*débiles*), en contraposición de los términos con una sola acepción (*fuertes*), y representa otra de las novedades que incorpora la última versión del modelo FIS-CRM. Este proceso se basa en el estudio de la co-ocurrencia de sinónimos en el documento para así identificar del conjunto de sinónimos adecuado. Cuando en el documento no co-ocurren al menos dos sinónimos pertenecientes al mismo conjunto de sinónimos es necesario recurrir al contexto local del término polisémico para, utilizando la información almacenada en las ontologías temáticas, identificar el conjunto de sinónimos más adecuado para dicho término.

- **Integración del modelo FIS-CRM en sistemas de búsqueda.**

El modelo FIS-CRM fue inicialmente integrado en el metabuscador FISS (*Fuzzy Interelations and Synonymy Searcher*) con el fin de representar los *snippets* de las páginas Web recuperadas por un motor de búsqueda tradicional (Google en este caso). En este sistema, el objetivo propuesto fue el de poder comparar conceptualmente los *snippets* recuperados, de manera que los resultados ofrecidos al usuario se pudiesen organizar jerárquicamente en grupos atendiendo a los conceptos contenidos en los documentos.



Búsqueda en FISS.

El primer componente del metabuscador se encarga de generar nuevas consultas a partir de los sinónimos de la consulta introducida por el usuario. Las nuevas consultas tienen un grado de compatibilidad con la consulta original obtenido a partir de los correspondientes grados de sinonimia de los términos implicados. Posteriormente los *snnipets* devueltos por el motor se representan mediante el modelo FIS-CRM y finalmente, mediante un algoritmo de *soft-clustering*, se obtiene una jerarquía de grupos de enlaces a páginas web "conceptualmente" relacionadas entre sí.
Otra interesante aplicación del modelo FIS-CRM al campo de la recuperación de la información es mediante un sistema integral de búsqueda basado en coincidencias de conceptos. La principal diferencia de este sistema respecto del metabuscador FISS es que el modelo FIS-CRM está también integrado en el subsistema de recopilación e indexado, por lo que el conjunto de páginas web accesibles por el motor están representadas conceptualmente

mediante este modelo. Así, el motor de este sistema, utilizando un sistema estándar de *matching* entre consultas y documentos, es capaz de recuperar las páginas que contienen los conceptos solicitados por el usuario. En este proyecto se espera integrarlo de la misma forma en la plataforma SCAIWEB.

- **Gestión documental: Ejemplo FzMail: Una Herramienta para la gestión Inteligente del Correo Electrónico.**

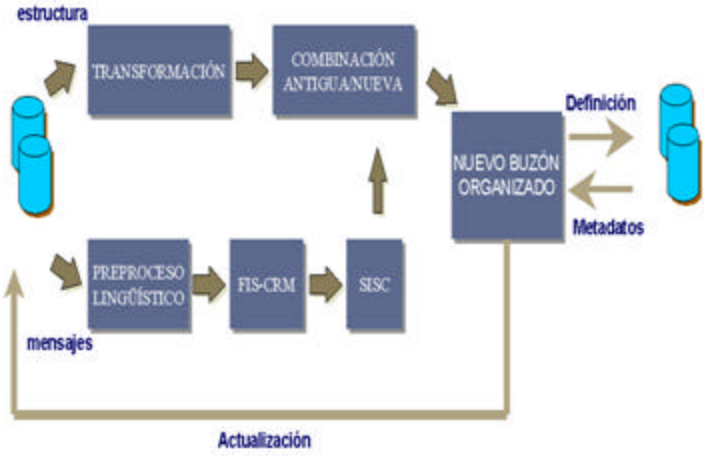
FzMail es una herramienta que se pretende desarrollar completamente e integrar en la plataforma SCAIWEB, que utiliza diversas técnicas de soft-computing con el propósito de gestionar de forma eficiente grandes volúmenes de correo electrónico. Los objetivos principales a perseguir son la obtención de una organización jerárquica y borrosa basada en los conceptos tratados en los mensajes, la clasificación automática de los correos entrantes de forma eficaz y facilitar las diferentes búsquedas que se realicen sobre el buzón de correo electrónico.
Para lograrlo se basa en tres pilares fundamentales: la representación conceptual de los mensajes, el *clustering* jerárquico-borroso y la representación del conocimiento obtenido mediante Prototipos Deformables Borrosos.
La organización a construir debe de tener la mayor calidad posible para que permita un aprovechamiento posterior óptimo por parte del usuario. Las características del resultado a obtener serán las siguientes:

1. Organización jerárquica de los mensajes en carpetas siguiendo criterios basados tanto en el contenido "conceptual" del mensaje como en los campos estructurados de éste. La representación conceptual de los mensajes permitirá la posterior búsqueda basada en conceptos.
2. Definición de las carpetas mediante términos/conceptos relevantes y Prototipos Deformables Borrosos.
3. Permitir que un mensaje pueda almacenarse según su contenido en más de un grupo de entre los que forman la jerarquía.
4. Cada mensaje debe de tener un grado de afinidad (o pertenencia) a cada grupo en el que esté clasificado. Los mensajes se ordenarán según este grado de afinidad.
5. Sencillo mecanismo de comparación entre documentos y la definición de cada una de las carpetas. Los resultados de este *"matching"* se utilizarán para almacenar el mensaje en una o varias carpetas.

Las fases de construcción esta organización son las siguientes:

1. Preproceso Lingüístico: Utilización de técnicas clásicas de Procesamiento del Lenguaje Natural (*stop words, steeming, ranking* de palabras) con el fin de tratar los términos existentes.
2. Representación Conceptual: Utilización de FIS-CRM para abstraer los conceptos inherentes en los mensajes basándose en las relaciones de sinonimia y ontologías borrosas existentes entre los términos.
3. *Clustering* Jerárquico-Borroso: Utilización de algoritmos de *clustering* para agrupar de forma jerárquica los mensajes conceptualmente similares, con la posibilidad de pertenecer un mensaje a varios grupos.

4. Post-Proceso y Representación del Conocimiento: Transformación de las agrupaciones del *clustering* para dar lugar a una estructura de carpetas y la representación de los grupos de documentos mediante Prototipos Deformables Borrosos.



Tratamiento de los mensajes en FzMail.

- **Desarrollo de un metabuscador basado en agentes: GUMSE**

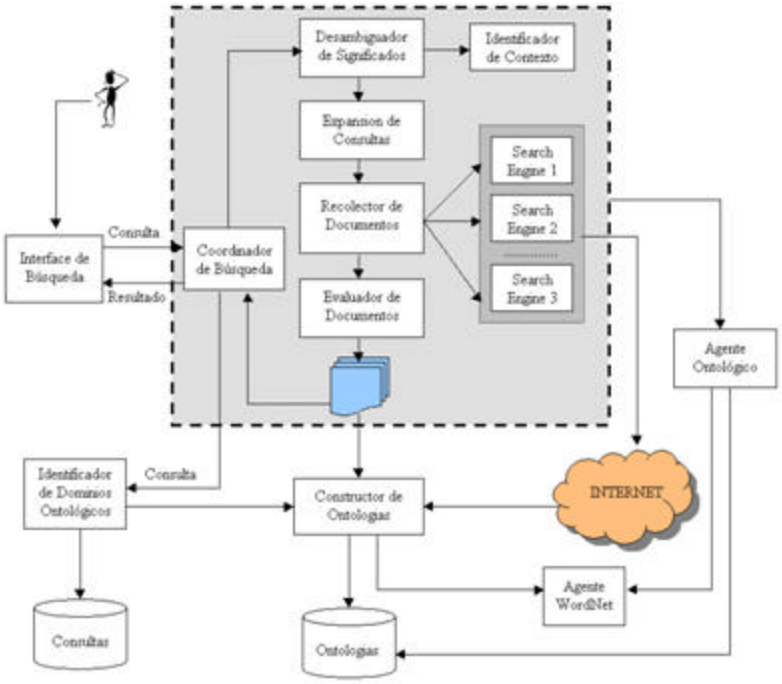
GUMSe (GUM Search) es un meta-buscador experimental que se pretende desarrollar para la plataforma SCAIWEB. Los meta-buscadores se presentan como una alternativa prometedora para tratar de paliar la baja precisión de los buscadores actuales. Las líneas de investigación actuales para mejorar la relevancia de los documentos devueltos por el proceso de búsqueda se basan en la incorporación de capacidades semánticas y deductivas mediante el uso de diccionarios, conceptos tales como la sinonimia y la antonimia, el uso de ontologías o la teoría de las percepciones. Uno de los principales problemas, a la hora de buscar, es la selección de los términos que formarán parte de la consulta. Existen distintos términos que se pueden elegir a la hora de referirse al mismo concepto, es un caso muy común que el autor de un documento utilice términos distintos a los que el usuario utiliza para tratar de recuperarlo. Este es el conocido problema del vocabulario. Hasta ahora era el usuario el encargado de atenuar este problema mediante el sucesivo refinamiento de la consulta. Este proceso conlleva una serie de iteraciones en las que el usuario debe construir nuevas consultas para obtener los resultados esperados. Para intentar reducir este proceso, y por tanto obtener unos resultados más satisfactorios, GUMSe combina varias técnicas para reducir el número de iteraciones, con la finalidad última de obtener un conjunto de documentos más relevantes. Las principales técnicas que se utilizan en GUMSe son 3 fundamentalmente: la expansión de consulta, la desambiguación del significado y una arquitectura basada en agentes. Dentro de la expansión de consulta se utiliza dependiendo de la situación la expansión automática o interactiva.

La **expansión de consulta automática** selecciona los términos que formarán parte de la consulta de forma transparente para el usuario. Pero antes de aplicar la expansión de consulta, es necesario conocer el significado adecuado al que el usuario se refiere cuando utiliza términos polisémicos, por lo que es preciso realizar una **desambiguación del significado**. En un principio, se utiliza WordNet como fuente de los términos que formarán parte de la expansión. WordNet es una herramienta muy utilizada para las aplicaciones del

procesamiento del lenguaje natural. Una de sus principales aplicaciones ha sido la recuperación y extracción de información, principalmente para la desambiguación automática de los significados de las palabras. En la **expansión de consulta interactiva** el objetivo es solicitar del usuario el significado más cercano a sus intenciones de búsqueda en aquellos casos en los que no se halla podido obtener de forma automática (o si el usuario decide utilizarlo). La información obtenida se puede utilizar para expandir la consulta añadiendo términos. Estos términos están relacionados mediante relaciones de sinonimia e hiperonimia con los introducidos por el usuario inicialmente. Además, este mecanismo interactivo permite al sistema aprender las relaciones entre términos y sus significados o acepciones. Esta característica aporta una valiosa información que se puede utilizar posteriormente para mejorar el sistema. Por ejemplo, es posible aprender nuevos significados, reconocer acrónimos o detectar significados que se pueden agrupar en uno solo. Además, permite mejorar la desambiguación automática basándose en la acumulación de la experiencia de los usuarios, ya que el sistema puede aprender las relaciones existentes entre los términos de una consulta y sus significados.

Habitualmente, las técnicas de expansión de consulta aumentan el factor *recall* sin que lleve asociado un aumento de la precisión. Debido a este inconveniente, no basta con un mecanismo de refinamiento de las consultas para obtener un conjunto de documentos más relevantes para el usuario. Se requiere un posterior procesamiento de los distintos resultados obtenidos que implica la integración y reajuste de la relevancia a la búsqueda del usuario. Por este motivo, se incorporan **mecanismos de evaluación** que permitan aumentar la precisión de la búsqueda, reduciendo la importancia de los documentos alejados conceptualmente. Una de las formas de abordar esta cuestión consiste en utilizar la información obtenida en la desambiguación del significado. Los términos relacionados con las acepciones no seleccionadas se utilizarán como indicadores negativos a la hora de determinar el grado de relevancia de un documento. De esta forma, conseguimos aprovechar las acepciones no seleccionadas para determinar la afinidad conceptual de los documentos con la consulta del usuario.

Por último, un aspecto determinante en el diseño de GUMSe es su **arquitectura basada en agentes** mostrada en la Figura 1. GUMSe está formado por una serie de agentes con distintas funciones que se comunican entre sí a través de una red. Se ha elegido este planteamiento porque permite la utilización de recursos distribuidos. Esta posibilidad aumenta la eficiencia y rendimiento al distribuir la carga de procesamiento entre diferentes computadoras y permitiendo la ejecución de tareas en paralelo.



Arquitectura de agentes de GUMSe.

Se pueden distinguir dos tipos de agentes dependiendo del tipo de participación en el proceso de recuperación. Si la participación es directa, se denominan "agentes de búsqueda", y son los que se encargan de satisfacer la consulta de los usuarios. El segundo tipo de agentes, denominados "agentes de soporte a la búsqueda", agrupa a los agentes que aportan una funcionalidad destinada a facilitar el acceso, construcción, gestión, o mejora de las estructuras de conocimiento que GUMSe utiliza para mejorar la búsqueda. La finalidad de estos últimos es identificar aquellos conceptos más demandados por los usuarios con el fin de construir ontologías que capten la semántica que subyace en torno a ellos. La información que poseen estas ontologías se utiliza posteriormente para refinar la consulta introducida por el usuario. El primer paso es identificar los dominios conceptuales más requeridos por los usuarios para mejorar o construir ontologías que traten de captar las relaciones semánticas y jerárquicas de los conceptos pertenecientes a ese dominio. Posteriormente se construyen ontologías utilizando los documentos relacionados con un concepto para obtener un conjunto de términos representativos. Evidentemente, el proceso de construcción de estas estructuras requiere el procesamiento de elevadas cantidades de documentos, lo que implica un proceso de recuperación, indexación y posterior tratamiento. Esta labor la realiza el Agente Constructor de Ontologías ayudado por el Agente Identificador de Dominios Ontológicos. Su labor es identificar los conceptos más buscados por los usuarios. Una vez conocido se procede a construir las ontologías que les den soporte. En este proceso se utilizan las consultas de los usuarios y los mejores documentos obtenidos por dichas consultas.

CÁLCULO DE DISTANCIAS BORROSAS ENTRE LOS TÉRMINOS DE UNA ONTOLOGÍA: Una ontología se define habitualmente como la conceptualización consensuada, formal y explícita de una realidad.

La representación de una ontología puede ser entendida como un grafo donde los nodos representan los conceptos o términos de la ontología y los arcos el tipo de relación existente entre los nodos que unen. Existen un gran número de relaciones entre términos: holonimia, hipernimia, meronimia, etc., que pueden ser obtenidas de diversas maneras: procesamiento de lenguaje natural, diccionarios electrónicos, etc., pero generalmente dotan de una gran complejidad a la ontología para poder ser usada en procesos automáticos. Por esta razón se realiza una simplificación del tipo de relaciones existentes entre nodos, distinguiéndose dos grandes tipos: relaciones físicas y relaciones semánticas.

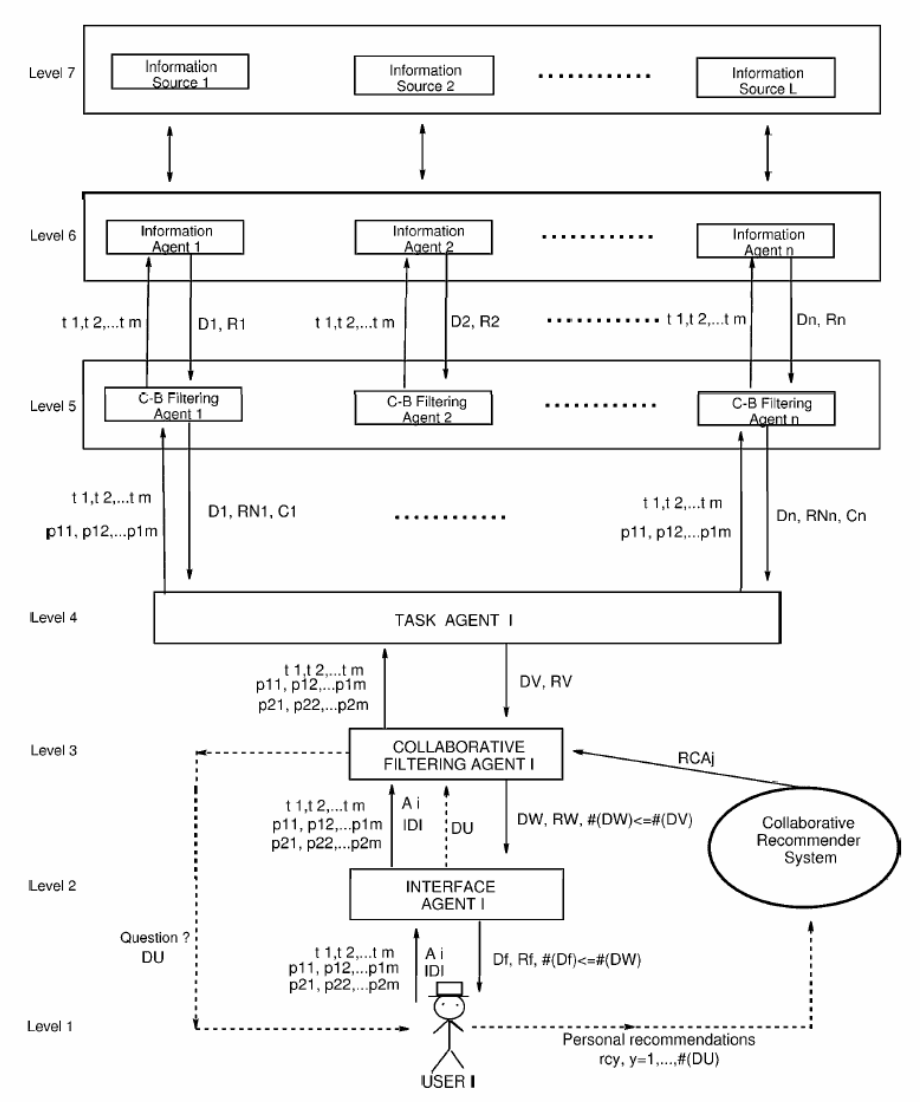
Las relaciones físicas engloban todas las relaciones anteriormente citadas: holonimia, meronimia, hiponimia, etc., relaciones *fuertes*, *evidentes* e *intuitivas* entre términos. Las relaciones semánticas se basan en la co-ocurrencia de términos y el concepto de distancia física. Así, la aparición reiterada de términos aislados o la co-ocurrencia de términos en el mismo documento fortalecen la idea de que el tema principal del texto versa entorno a ellos. Mientras, la distancia física esta basada en la idea de que los términos que más cercanos se encuentran en un documento tienen mayor probabilidad de estar fuertemente relacionados que aquellos términos que aun estando en el mismo documento se encuentran en párrafos distintos. Basándose en estas dos únicas relaciones se aplican distintos operadores borrosos para obtener distintas ontologías generadas de manera completamente automática.

El proceso de búsqueda se puede descomponer en una serie de fases. El objetivo consiste en refinar sucesivamente la carga semántica de la consulta para permitir expandir los resultados de la misma con documentos relacionados conceptualmente. Las 4 fases son las siguientes:

1. **Refinamiento del significado de la búsqueda**: Se identifica los significados adecuados de los términos polisémicos que intervienen en la consulta.
2. **Expansión de la consulta**: Se utiliza la información adquirida de la fase anterior para aportar una mayor carga semántica al conjunto de consultas. En esta etapa se construyen una serie de consultas adicionales a la original. En GUMSe, las nuevas consultas se pueden construir utilizando cuatro procedimientos distintos. El primero consiste en introducir términos que posean relaciones de sinonimia e hiponimia con los términos de la consulta del usuario. También se utilizan variaciones morfológicas de los términos de la consulta. El aspecto más novedoso consiste en la utilización de operadores avanzados de búsqueda en la consulta. Cada motor de búsqueda existente tiene sus propias características y operadores de búsqueda concretos. Este motivo ha propiciado el diseño de un lenguaje propio basado en XML utilizado en el proceso de expansión de consulta, que posteriormente se traduce al lenguaje de consulta específico de cada buscador. De esta forma se aísla la construcción de las nuevas consultas expandidas de un buscador concreto, para permitir en futuras mejoras del sistema incorporar otros buscadores a GUMSe. El último método de expansión consiste en introducir conceptos relacionados semánticamente. Junto a los métodos anteriores se utiliza un refinamiento negativo de la consulta. El objetivo es descartar documentos que posean términos relacionados con significados descartados, consiguiéndose de esta forma enfocar la búsqueda.
3. **Recolección de la web**: El siguiente paso consiste en lanzar las distintas consultas a una serie de motores de búsqueda para obtener una colección finita de páginas. Actualmente, solo se considera la utilización de Google y Altavista.
4. **Evaluación de los resultados** : Por último se procede a recuperar las páginas obtenidas y mostrarlas al usuario ordenadas bajo un único criterio de relevancia. Se deben obtener los resultados de las consultas realizadas a los diferentes buscadores y agruparlas en una única colección de documentos, eliminando aquellos documentos repetidos y

estableciendo un orden de aparición. Por motivos de eficiencia, se han considerado dos sistemas diferentes de evaluación de los documentos denominados evaluación rápida y evaluación exhaustiva. La primera es más eficiente y rápida que la segunda ya que no entra a examinar el contenido de los documentos que evalúa, sin embargo, la segunda modalidad de evaluación ofrece unos cálculos de relevancia más fiables.

- **Diseño de un mecanismo de recuperación de información y filtrado.**
- **Diseño de un mecanismo de minería de textos.**
- **Diseño de un mecanismo de ayuda a la toma de decisiones en base a perfiles de usuario:**
    - Modelado de las necesidades de información de los usuarios a través de lenguajes de consultas ponderados lingüísticos difusos y algoritmos genéticos.
    - Desarrollo de funciones de evaluación y operadores de agregación soft para la evolución de las consultas ponderadas.
    - Desarrollo de sistemas multi-agente para el acceso a la información Web basados en técnicas lingüísticas difusas, tal y como se describe en la figura:



- Desarrollo de técnicas de filtrado de información basadas en modelado lingüístico difuso.
- Aprendizaje de perfiles de usuario basados en algoritmos genéticos.

Short Bio: José A. Olivas

Born in 1964 in Lugo (Spain), received his M.S. degree in Philosophy in 1990 (University of Santiago de Compostela), Master on Knowledge Engineering of the Department of Artificial Intelligence, Polytechnic University of Madrid in 1992, and his Ph.D. in Computer Science in 2000 (University of Castilla–La Mancha). In 2001 was Postdoc Visiting Scholar at Lotfi Zadeh's BISC (Berkeley Initiative in Soft Computing), University of California-Berkeley, USA. His current main research interests are in the field of Soft Computing for Information Retrieval and Knowledge Engineering applications. He received the Environment Research Award 2002 from the Madrid Council (Spain) for his PhD. Thesis. PRINCIPAL EMPLOYMENT AND AFFILIATIONS: From 1997: Associate Professor of the Department of Computer Science, University of Castilla–La Mancha, Ciudad Real, Spain. From 1997: Professor of the Department of Computer Science, ICAI–Universidad Pontificia Comillas, Madrid, Spain. 1995-97: Head of the Department of Artificial Intelligence and Computer Science, University Antonio de Nebrija–UNNE, Madrid, Spain. From 1995: Collaborations with INSA (Aero Spatial Engineering and Services, NASA - Spain), Processing of forest fires data from satellites. 1992-1996: Head of the Computer Science Department of PPM studies center (Tres Cantos, MADRID): Consulting on Intelligent Systems to Enterprises such as SOUTHCO or ATT.

Address and Affiliation:
José A. Olivas,
Computer Science Dept. Esc. Superior de Informática, UCLM.
Paseo de la Universidad 4, 13071-Ciudad Real, Spain.
Tel. +34 926295300 Ext. 3730
Fax +34 926 295354
e-mail: joseangel.olivas@uclm.es