Memoria de Investigación

Uso de Información Semántica para la mejora de la Recuperación de Información en la Web

Autor:
JESÚS GONZÁLEZ MARTÍ

Dirigida por:
JOSÉ ANTONIO TROYANO JIMÉNEZ

Abril 2009
Sevilla

A Elena,
por su compañía
y apoyo incondicional.

Agradecimientos

En primer lugar quiero agradecer a mi tutor José Antonio Troyano su apoyo en este proyecto. Han pasado más de seis años desde que me ofreció una beca en IBM para iniciarme en el apasionante mundo del Procesamiento del Lenguaje Natural. Hoy día, este conocimiento es parte fundamental de mi vida profesional.

Quiero agradecer también a todos los compañeros y profesores de los cursos de doctorado por sus charlas impartidas, así como al resto de profesores que han contribuido a que hoy día quiera seguir formándome como ingeniero informático.

Resulta imprescindible nombrar a mis compañeros de INDISYS y del grupo de investigación JULIETTA ya que entre todos han colaborado directa o indirectamente en que me interesara por la Recuperación de Información y la Búsqueda Semántica en la Web.

También quiero mostrar mi agradecimiento a todos los desarrolladores e investigadores que contribuyen día a día con su trabajo y esfuerzo a la mejora de la tecnología, especialmente la aplicada a Internet y al Procesamiento del Lenguaje Natural.

Por último, debo agradecer enormemente a mi familia y amigos simplemente por estar ahí y ser como son.

Resumen

Los recientes avances en la tecnología orientada al desarrollo web han dado lugar a nuevas aplicaciones que permiten involucrar a los usuarios en la creación de contenidos. Cada vez más, dichas aplicaciones se diseñan teniendo en cuenta las directrices de la web semántica. Sin embargo, las técnicas de recuperación de información actuales no son suficientes para dotar a los buscadores de la inteligencia y versatilidad necesarias para aprovechar tales características. En la presente memoria de investigación se expone el estado del arte de las técnicas basadas en el procesamiento del lenguaje natural para desarrollar motores de búsqueda semántica.

Abstract

Recent progress in web development technologies have resulted in applications which allow users to be involved in content creation. These applications are more increasingly being designed with the semantic web principles in mind. Nevertheless, current information retrieval systems do not provide search engines with the intelligence and versatility required by the abovementioned frameworks. This work surveys state-of-the-art Natural Language Processing techniques in the development of semantic search engines.

Prólogo

Desde la aparición de internet, los contenidos de la web han ido creciendo a velocidad de vértigo. Del mismo modo, la tecnología para el desarrollo web ha evolucionado desde las primeras páginas estáticas en simple `<HTML>` hasta el uso de comunicaciones AJAX, objetos Flash, portlets, etc... Hoy día se está llevando a cabo la revolución conocida como Web 2.0 en la que se anima a los propios usuarios a generar contenidos. Al mismo tiempo, se tiende al impulso de la web semántica en la que los contenidos se autodefinen. De este modo se permite la interacción automática con terceras aplicaciones sin intervención humana.

Todas estas nuevas características no son explotables sin las técnicas adecuadas. En el caso de la recuperación de información en internet estamos acostumbrados a usar buscadores comerciales que emplean palabras clave. Sin embargo, estos buscadores no están capacitados para sacar todo el rendimiento de las nuevas características de la web. Actualmente, los buscadores no comprenden que buscar por «coche rojo» es prácticamente idéntico que buscar por «automóvil colorado». O que «comida a domicilio» podría dar como resultados los mejores restaurantes de reparto de alrededor con información adicional sobre los mismos como: vídeos, fotos, opiniones, etc...

En la presente memoria de investigación se pretende dar una visión global del estado del arte actual en cuanto a los estudios que se están llevando a cabo sobre la denominada *búsqueda semántica*. Entendemos por tal, a la búsqueda de documentos en la que se utilicen técnicas de procesamiento del lenguaje natural para extraer la semántica de los contenidos y no realizar un simple «matching» por palabras clave. De aquí el título de la memoria «*Uso de Información Semántica para la mejora de la Recuperación de Información en la Web*». Finalmente, para la exposición del trabajo de investigación realizado se ha dividido la memoria en los siguientes capítulos:

Capítulo 1: Introducción y Antecedentes

En el primer capítulo introduciremos ampliamente los conceptos de «Procesamiento del Lenguaje Natural (PLN)» y una de sus aplicaciones: la «Recuperación de Información (RI)». Será sobre esta última disciplina sobre la que se desarrolle el resto de la memoria.

Capítulo 2: Hipótesis y Objetivos

Introducidas las materias de base, en el segundo capítulo se exponen las hipótesis e ideas que permitirían mejorar la búsqueda web tradicional. Así como una serie de objetivos a cumplir en el desarrollo de esta memoria.

Capítulo 3: Estado del Arte

En este capítulo se estudia en profundidad la evolución producida en la web hasta alcanzar la idea de la «*Web Semántica*» y se define el objetivo primordial: la «*Búsqueda Semántica*». Posteriormente se enuncian las distintas líneas de investigación y se finaliza con un exhausto informe sobre el contexto investigador en el que se recogen entre otros los congresos y grupos de investigación más importantes.

Capítulo 4: Conclusiones

Como en todo trabajo de investigación, se añade un capítulo de conclusiones con los resultados e impresiones de esta memoria.

Apéndice 4: Currículum Investigador

Se finaliza la obra con un apéndice dedicado al currículum investigador del autor de esta memoria.

Índice general

Capítulo 1

Introducción y Antecedentes

En este primer capítulo se realiza una introducción al «Procesamiento del Lenguaje Natural (PLN)». Veremos los orígenes de esta disciplina y se expondrán los problemas reales a los que hace frente. Seguidamente se enuncian las distintas aplicaciones del PLN prestando especial interés a la «Recuperación de Información (RI)». Será esta última actividad la que centre el foco de atención en el resto del capítulo, puesto que todo el estudio sobre búsqueda semántica de la presente memoria de investigación se basa en dicha disciplina.

1.1. Procesamiento del Lenguaje Natural (PLN)

El lenguaje posibilita la comunicación entre seres humanos, ya sea por medio oral o escrito. La forma de intercambiar mensajes entre personas ha ido evolucionado a lo largo del tiempo; hemos pasado de hablar cara a cara, escribirnos cartas, llamarnos por teléfono a finalmente chatear por internet. Quizás en un futuro nos comuniquemos incluso telepáticamente. A los lenguajes que utilizamos para estos propósitos se les denomina formalmente «*lenguajes naturales*».

Paralelamente, el desarrollo tecnológico ha permitido mejorar nuestra calidad de vida; la radio, el automóvil, la televisión, el mando a distancia, los ordenadores, los teléfonos móviles, los GPS, las PDAs, etc... Sin embargo, podríamos decir que existe un contrasentido en cuanto a la evolución de las formas de comunicación entre los humanos y entre los humanos y las nuevas tecnologías. Si preferimos el lenguaje natural para comunicarnos entre nosotros, ¿por qué no queremos *hablar a las máquinas*?

Para una respuesta completa habría que tener en cuenta elementos psicológicos y socioeconómicos, no obstante, es posible encontrar una razón mucho más simple atendiendo únicamente a factores tecnológicos: habilitar a las máquinas para procesar nuestro lenguaje natural es una tarea extremadamente compleja. Abarca campos tan amplios y diversos como el reconocimiento de voz o el diálogo inteligente. Podría decirse que el PLN todavía se encuentra en desarrollo para conseguir una aceptación

plena, sin embargo, y a pesar de las dificultades, realizamos ciertas tareas cotidianas sin ser conscientes de que son aplicaciones directas del PLN. De hecho, hemos llegado a un punto en el que estamos acostumbrados a:

- Realizar *búsquedas por palabras* en internet con una herramienta tipo Google.
- Redactar un documento usando el *corrector ortográfico*.
- Viajar con un GPS el cual emplea *la síntesis de voz* para dar instrucciones.
- Sistemas de *reconocimiento de voz* para personas discapacitadas.
- Escribir un *sms* con *texto predictivo*.

1.1.1. Definición e historia

Existen distintas definiciones para el PLN; en conjunto vienen a decir lo siguiente: El «Procesamiento del Lenguaje Natural» es una disciplina que se encuentra a medio camino entre la *inteligencia artificial* y la *lingüística computacional*. Su objetivo es que las máquinas puedan procesar el lenguaje humano de una manera semejante a como lo haría una persona. Podríamos definir por tanto el PLN como el «**conjunto de técnicas que permiten que un ordenador entienda el lenguaje natural como una forma más de comunicación**».

Historia

El origen del PLN se sitúa a finales de los años 40 con el intento de traducción automática de textos a varios idiomas. Desafortunadamente, se produce un fracaso estrepitoso en los resultados: por un lado, se subestimó la dificultad que entrañaría dicha tarea. Por otro lado, la potencia de los ordenadores de la época no era suficiente para realizar la computación de los algoritmos de traducción de forma eficiente. Ante tales adversidades se produjo una pérdida de confianza en el potencial del PLN.

Hubo que esperar hasta finales de los 60 y principios de lo 70, cuando se produjo un renacimiento del PLN gracias al desarrollo de interfaces de texto para consultas sobre bases de datos. Estas herramientas gozaron de una gran acogida, especialmente por parte de la comunidad científica investigadora. Finalmente, y gracias en parte al salto tecnológico en la potencia de los ordenadores, en la década de los 80 resurge la investigación sobre traducción automática, acompañada de nuevas líneas de investigación como la síntesis y el reconocimiento de voz o el diálogo inteligente.

El interés social por conseguir que las máquinas —hablen y entiendan— como los seres humanos, recibe un inesperado apoyo por parte del cine de ciencia ficción. Hace

ya más de treinta años que se estrenaron dos de las películas que mejor representan el ánimo de conseguir *máquinas parlantes*: «2001: Una Odisea en el Espacio» (1968) y «La Guerra de las Galaxias» (1977). En el film de Kubrick aparece el superordenador *HAL 9000* (fig. 1.1), el cual, más allá de su *inteligencia* que le conduce a prescindir de los humanos, comprende e interpreta a la perfección el lenguaje natural y es capaz de mantener un diálogo inteligente y fluido con el resto de protagonistas.

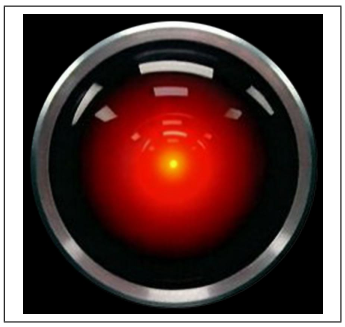

Figura 1.1: El amenazante ojo de HAL 9000

En la película de George Lucas la inteligencia es una característica común en los robots, aunque el dominio del lenguaje natural es exclusivo de los denominados *droides de protocolo*. Es el caso de *C-3PO* (fig. 1.2), que se presentaba ante su amo Luke Skywalker de la siguiente manera: —«Soy el androide de protocolo *C-3PO*. Relaciones cibernéticas humanas. Domino más de seis millones de formas de comunicación».[1]

Estas visiones futuristas despertaron un extraordinario interés en la sociedad, que se preguntaba si la inteligencia artificial, e implícitamente el procesamiento del lenguaje natural, alcanzarían algún día tal grado de sofisticación. Lamentablemente, hay que reconocer que el actual nivel de evolución del PLN dista mucho de lo que aún hoy se relata en la ciencia ficción. Podríamos decir que las máquinas ya pueden *hablar* con una aceptable inteligibilidad y *oír* con ciertas dificultades, pero aún les queda mucho camino por recorrer en el aspecto más importante: ***entender***.

1.1.2. Conocimiento del lenguaje

El procesamiento del lenguaje natural debe enfrentarse a multitud de problemas derivados de la inherente complejidad del propio lenguaje. Para un correcto entendimiento de todas las dificultades es necesario conocer las diferencias entre procesar un «lenguaje formal» y un «lenguaje natural». Como veremos a continuación, el gran

[1] Sin duda, todo un logro tecnológico de los investigadores en PLN de aquella galaxia tan lejana...

Figura 1.2: El parlanchín androide C-3PO

talón de Aquiles de los lenguajes naturales es la ambigüedad.

Lenguajes formales

Un «*lenguaje formal*» es aquel que viene definido por una gramática completa que no da lugar a ambigüedades, como es el caso del lenguaje matemático, la lógica booleana o los lenguajes de programación. En estos últimos, las instrucciones que deseamos ejecutar en la máquina quedan perfectamente definidas. No existen instrucciones ambiguas que el ordenador pueda interpretar de una forma u otra.

Para los lenguajes formales existen varios modelos que cubren todos los niveles: léxico, gramática, semántica, etc... Se puede decir que el grado de madurez de dichos modelos, así como de las herramientas existentes para trabajar con ellos es bastante alto, y por tanto, el abanico de nuevas vías de investigación es muy reducido.

Lenguajes naturales

Los «*lenguajes naturales*», como bien pueden ser el castellano, el inglés o incluso el latín, son formas de comunicación que no tienen reglas definidas en sus orígenes sino que han ido evolucionando a lo largo del tiempo. Actualmente muchos lingüistas y filólogos colaboran en la ardua tarea de encontrar dichas reglas. Por supuesto la dificultad varía de un idioma a otro, pero sin duda el gran problema común al que deben enfrentarse todos los modelos de procesamiento del lenguaje natural es la ya citada ***ambigüedad***.

En relación con los lenguajes formales, para los lenguajes naturales contamos con mucha más variedad en cuanto a modelos de procesamiento se refiere. Existen numerosas vías de investigación posibles y aún hoy siguen apareciendo nuevas tendencias y caminos por explorar. Para conseguir un alto grado de comprensión del lenguaje natural es necesario que los algoritmos en cuestión posean un completo conocimiento del idioma; desde saber qué caracteres delimitan una palabra hasta el contexto del diálogo. En general, el tipo de inteligencia que necesitan las herramientas de procesamiento del lenguaje natural abarcan los siguientes niveles:

- **Nivel fonológico:** Este nivel aplica especialmente a los sintetizadores de voz y a los reconocedores del habla. Estos sistemas deben conocer cuáles son las unidades fonéticas del idioma así como su correcta pronunciación. Podemos encontrar ambigüedad fonológica en los siguientes escenarios: Si únicamente escuchamos el fonema */ki/*, en castellano podría referirse a «*ki*» o a «*qui*». O el fenómeno contrario, en inglés la primera sílaba de «*woman*» y «*women*» son idénticas pero se pronuncian de forma distinta (*/uo/* y */ui/* respectivamente).

- **Nivel léxico:** El conocimiento léxico trata sobre qué palabras forman parte del lenguaje y cuales no. Además, debe proporcionar información sobre el papel que juega una palabra dentro de la oración. Así pues, una palabra puede encuadrarse en distintas categorías tales como como sustantivos, adjetivos, pronombres, etc... Como ejemplo de ambigüedad léxica, la palabra «para» puede ser una preposición o una forma verbal del verbo «parar». O la propia palabra «la» que puede ser un determinante, un pronombre o el sustantivo que define la nota musical que lleva su nombre.

- **Nivel sintáctico:** El conocimiento sintáctico proporciona la cualidad de identificar si una oración tiene cabida o no dentro del lenguaje. Este determinismo se realiza en base a la gramática del propio lenguaje. Gracias al análisis sintáctico se obtienen las relaciones existentes entre las distintas palabras de una oración; qué adjetivo acompaña a que sustantivo o a qué objeto se refiere un pronombre. Existen multitud de casos de ambigüedad sintáctica. Hay un ejemplo muy conocido en la literatura científica que admite distintos árboles de sintaxis: «Juan vio a María en el parque con el telescopio.»

- **Nivel semántico:** En este nivel se trata el significado individual de cada palabra y de cómo combinarlos para determinar el sentido de la oración. Suele ser la etapa más compleja de cualquier herramienta de PLN puesto que hay que definir sin ambigüedades cada uno de los posibles significados. En otras palabras, en esta fase deben quedar resueltas todas las posibles ambigüedades provenientes de etapas anteriores. Las ambigüedades más comunes de este nivel

vienen dadas por las palabras polisémicas: «banco», «sierra», «red», etc...

- **Nivel pragmático:** En el nivel pragmático es donde se necesita un conocimiento previo a la oración que se está analizando para resolver ciertas referencias. El lenguaje natural está repleto de pronombres y adverbios que por sí solos carecen de significado. Los humanos resolvemos este problema inconscientemente, pero una máquina necesita tener almacenada dicha información. Un ejemplo de la complejidad de este nivel es la resolución de la anáfora: —«¿Viste a Javier con su nuevo coche ayer?» —«Si, lo vi». En la segunda frase, el pronombre «*lo*» necesita de más información para resolver su significado.

- **Nivel contextual:** Hay quienes consideran este nivel como parte del conocimiento pragmático. La información de contexto es el estado del mundo que rodea a la situación del diálogo: lugar, fecha, etc... Además, existe un conocimiento sobre nuestro interlocutor que nos influye a la hora de elaborar nuestras oraciones: edad, género o nivel cultural. Por tanto, hay una serie de factores con una gran influencia en el desarrollo del diálogo y que de un modo u otro deberían ser tenidos en cuenta por la máquina. En el ejemplo anterior, para que un ordenador entienda el significado del adverbio temporal «*ayer*», necesita conocer la fecha de *hoy*. Lamentablemente queda mucho camino por recorrer en las herramientas de PLN para conseguir este tipo de flexibilidad contextual.

1.1.3. Aplicaciones del PLN

Tras esta breve introducción, describiremos las aplicaciones más relevantes que requieren técnicas de procesamiento del lenguaje natural. Algunas ya han sido mencionadas por su popularidad como la traducción automática, el reconocimiento de voz, la corrección ortográfica o la recuperación de información.

Traducción automática

La traducción automática o MT (del inglés *"Machine Translation"*), consiste como su propio nombre indica, en convertir un texto de un idioma a otro por medio de un ordenador y sin intervención humana. Como ya se mencionó en la introducción, es la tarea que da origen al procesamiento del lenguaje natural como tal. Se trata por tanto, de la disciplina que más ha contribuido en el desarrollo de la lingüística computacional. Es seguramente también una de las aplicaciones informáticas que mayores recursos ha recibido por parte de las administraciones públicas ya que es muy grande el interés de distintos gobiernos en que esta tecnología se termine de desarrollar.

La aspiración de obtener artilugios mecánicos que sirvan para superar las barreras lingüísticas viene de antiguo. En el siglo XVII se habla de la utilización de diccionarios mecánicos, basados en códigos numéricos universales con el objeto de de crear una «*lengua universal*», no ambigua, basada en principios lógicos y símbolos icónicos que permitiese comunicarse a toda la humanidad. Este empeño precede por bastante tiempo a la propia existencia del ordenador. Según La Biblia (Libro del Génesis, Cap. 11. Antiguo Testamento) fue en «La Torre de Babel» (ver fig. 1.3) donde Dios creó la confusión entre los hombres para evitar que construyeran la torre que les haría alcanzar el cielo. De nuevo en la Época Contemporánea, desde el momento en que un ordenador estuvo disponible en la década de los 40, la traducción automática pasó a convertirse inmediatamente en una de las tareas estrella de la investigación y la computación.

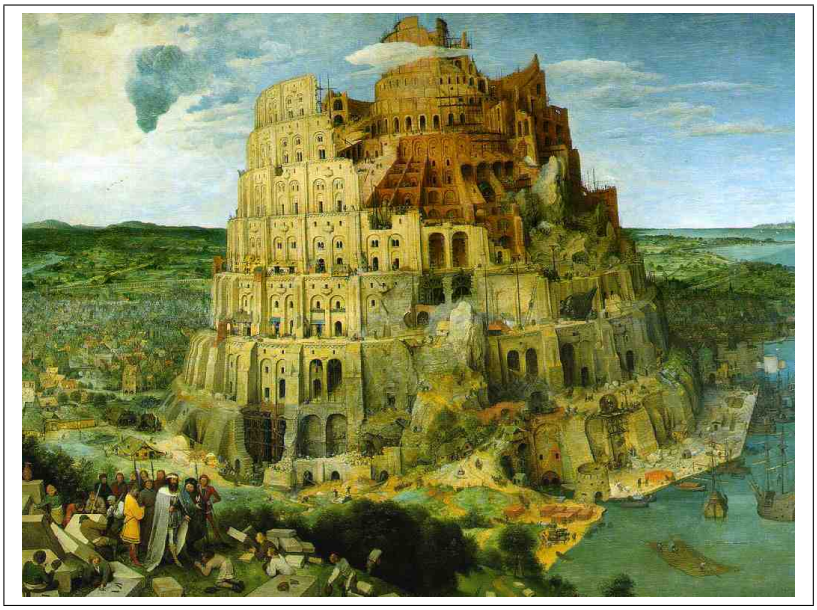

Figura 1.3: Ilustración de la Torre de Babel

El proceso que rige el funcionamiento de la traducción automática es el siguiente: dado un texto de entrada escrito en un idioma origen, se analiza, se procesa, y finalmente se genera el mismo texto en el idioma destino. Las herramientas actuales consiguen un alto grado de fiabilidad, no obstante, todavía no puede prescindirse totalmente de la posterior revisión humana. Es posible conseguir mejores resultados pero rescindiendo la traducción a dominios limitados. En general, podemos clasificar los sistemas de traducción automática de la siguiente forma:

- **Sistemas de traducción directa:** Podrían equipararse con grandes diccionarios. Generalmente realizan la traducción casi palabra por palabra, ya que la

información sintáctica que poseen es mínima. Por ello los resultados que ofrecen suelen ser bastante pobres.

- **Sistemas de transferencia:** Contienen grandes léxicos bilingües además de un amplio conocimiento sintáctico y semántico de las lenguas tratadas. Esto permite traducir palabras de una lengua a otra teniendo en cuenta el contexto morfológico, sintáctico y semántico de la frase. También pueden llevar a cabo la transferencia estructural, es decir, los cambios en el orden de elementos y en la estructura de la frase para adecuarse a cada lengua.

- **Sistemas «interlingua»:** Estos sistemas se basan en un lenguaje artificial conocido como «interlingua», el cual comparte todas las características y hace todas las distinciones entre todos los idiomas. Para realizar una traducción se usa un analizador para convertir el texto en idioma fuente al interlingua, y un generador que convierte el texto interlingua al idioma destino. Aunque teóricamente se trataría del mejor enfoque de los tres, en realidad estos sistemas están en fase de laboratorio o se utilizan para aplicaciones muy restringidas, debido a los problemas prácticos que presenta el diseño y la implementación de una «interlingua» eficaz.

- **Sistemas estadísticos:** La traducción automática estadística consiste, en esencia, en calcular las distintas probabilidades de que cierta cadena de texto en el lenguaje destino sea la traducción de la misma cadena en el lenguaje origen. Al traducir un texto completo, el objetivo es maximizar la probabilidad de todos los pares de cadenas origen-destino. Para el cálculo de dichas probabilidades se realiza una fase de entrenamiento previo con textos ya traducidos y supervisados por un humano.

Recuperación de información

La recuperación de información o IR (del inglés *"Information Retrieval"*), consiste en la búsqueda de cierta información o de la obtención de los documentos relevantes que aportan dicha información. Esta búsqueda podrá realizarse en distintas fuentes como pueden ser Internet, una base de datos, un catálogo de biblioteca, etc... Puesto que la recuperación de información es el principal objeto de estudio de esta memoria de investigación, dejaremos para el apartado 1.2 su estudio en profundidad.

Corrección ortográfica

La corrección ortográfica o *"Spell Checking"* es una de las aplicaciones del PLN de mayor uso cotidiano a pesar de que suele pasar desapercibida. Todos hemos manejado

procesadores de texto con capacidad de corrección ortográfica; la aplicación está continuamente analizando el texto y le avisa de los errores normalmente subrayando la palabra en rojo.

Los primeros correctores ortográficos eran muy simples: contaban con un corpus léxico del idioma a analizar y si una palabra no estaba en el mismo entonces era considerada como errónea. Con el tiempo, los algoritmos fueron evolucionado y aparecieron los primeros correctores ortográficos que mostraban sugerencias cercanas a la palabra mal escrita. Así pues, si cometíamos una «herrata», entre las sugerencias encontraríamos la palabra correcta: «errata». Los siguientes avances fueron encaminados a encontrar errores sintácticos y/o de concordancia de género y número. De este modo, un potente corrector automático[2] nos avisaría ante frases erróneas tales como «los texto» o «la niña bonito».

Actualmente, los correctores automáticos han dado el salto más allá de los procesadores de texto de tal forma que podemos encontrarlos en multitud de aplicaciones: los navegadores web ya incorporan un corrector para los contenidos introducidos por el usuario, los propios sitios web de edición de blogs y contenidos wiki y hasta los clientes de mensajería instantánea hacen uso de esta tecnología.

Generación automática de resúmenes

La generación automática de resúmenes (en inglés *automatic summarization*) consiste en crear una versión más reducida de un texto donde se condense la información más importante. Normalmente estas aplicaciones realizan un análisis iterativo del texto de entrada de forma que van remarcando las partes que se consideren relevantes.

Para conseguir resultados aceptables es necesario tener en cuenta diversos aspectos tales como el tipo del documento: noticia, ensayo, etc... El resumen final también vendrá determinado por el estilo de redacción del texto, su longitud o incluso el tipo de expresiones utilizadas. En general, hay dos formas de generar los resúmenes: extrayendo o abstrayendo. La primera variante consiste en extraer literalmente las partes más relevantes del documento. Sirva de ejemplo los pequeños resúmenes o *extractos* que nos muestran los buscadores donde marcan el contexto en el que aparecen las palabras que buscamos. Estos pequeños fragmentos del texto original se obtienen por extracción y en la jerga técnica se les conoce como «*snippets*». En cuanto a la *abstracción*, consiste en parafrasear el texto original en unas pocas líneas, es decir, realizar una explicación no literal para hacerlo más claro, conciso e inteligible.

[2] En este escenario, el adjetivo *ortográfico* deja de tener sentido.

Síntesis de voz

La síntesis de voz tiene como objetivo transformar, de forma automática, cualquier texto en una locución auditiva lo más semejante posible al habla de un humano. A un sistema capaz de producir dicha salida se le conoce como sintetizador de voz o en inglés *Text-To-Speech* (TTS). Los primeros esfuerzos por conseguir voz artificial datan de 1779, con la construcción de artilugios mecánicos a base de resonadores muy similares a los instrumentos de viento. Más tarde llegarían los sintetizadores electromecánicos. El más famoso que se recuerda es el *VODER* que se presentó en sociedad en 1939. Es en la década de los años 60 cuando empiezan a proliferar los primeros sintetizadores software, que finalmente darían lugar a los sintetizadores modernos.

Un sistema de TTS actual recibe como entrada una cadena de texto en un idioma determinado, la analiza y procesa y da como resultado una salida auditiva que contiene una locución verbal representando al texto de entrada. Durante la etapa de análisis se realizan multitud de operaciones como la normalización del texto, la pronunciación o el cálculo de la prosodia: velocidad, tono, duración, etc... A la hora de generar la salida auditiva, necesitan de un modelo de lenguaje así como de información fonética del mismo. En la figura 1.4 se muestra la arquitectura general de un sistema de *Text-To-Speech*.

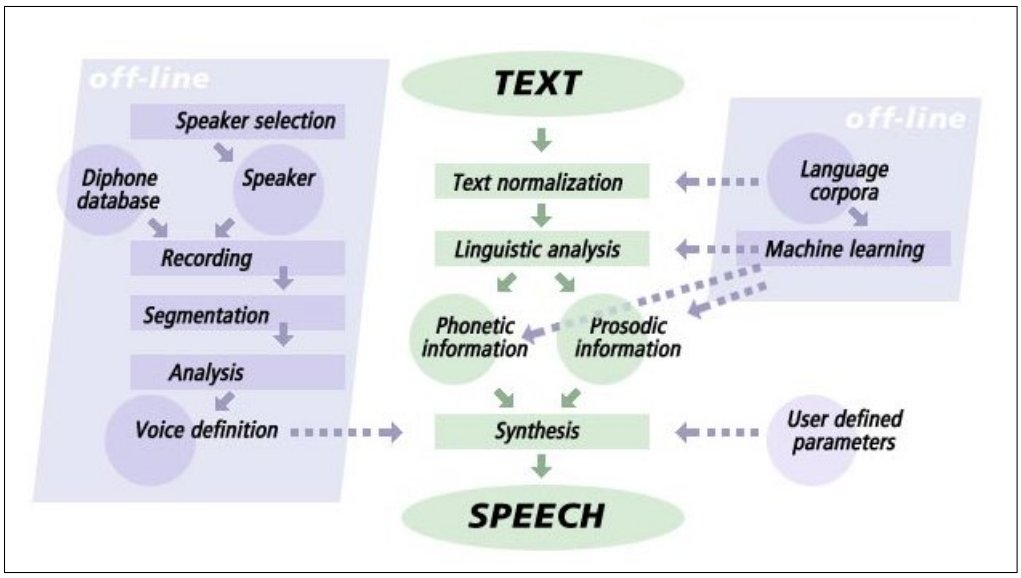

Figura 1.4: Arquitectura de un sistema de *Text-To-Speech*

Son muy utilizados en los sistemas de avisos automáticos de aeropuertos, estaciones de autobuses y ferrocarriles. También podemos encontrarlos en sistemas telefónicos de atención al cliente, dispositivos móviles y GPS, etc... Fuera del ámbito comercial,

también son empleados en aplicaciones educativas y en los sistemas para personas con discapacidades visuales.

Reconocimiento de voz

El reconocimiento de voz o ASR (del inglés *"Automatic Speech Recognition"*), puede definirse como el proceso inverso a la síntesis de voz. Un sistema de ASR recibe como entrada una locución de audio y genera como salida una cadena de texto que transcribe dicha locución. Su arquitectura básica puede verse en la figura 1.5.

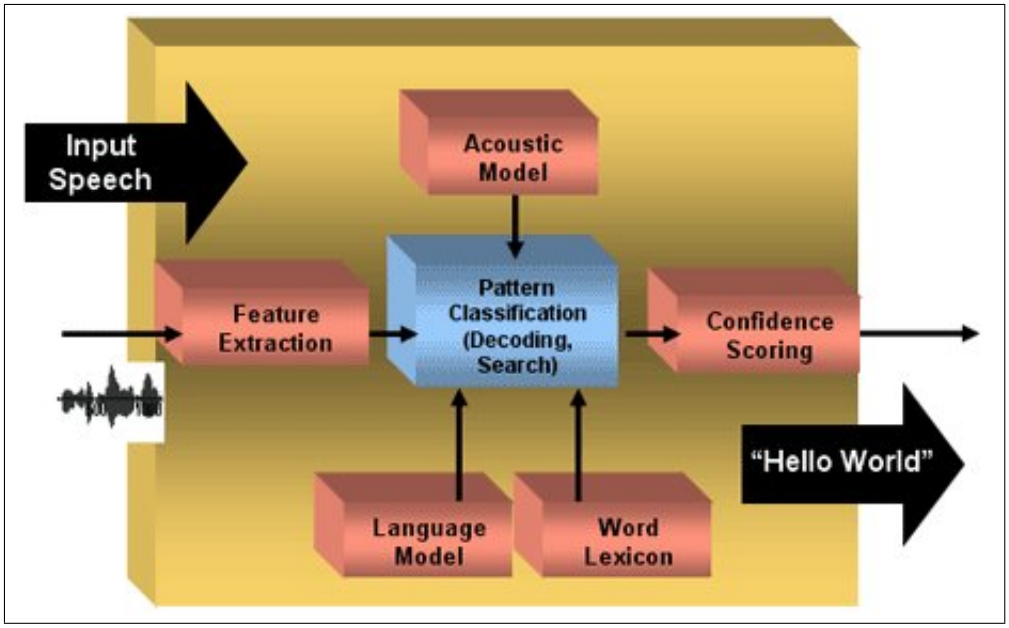

Figura 1.5: Arquitectura de un sistema de *ASR*

Los sistemas comerciales de reconocimiento de voz han estado disponibles desde el año 1990. Están siendo muy utilizados en aplicaciones telefónicas: agencias de viajes, reserva de billetes/entradas, atención al cliente, información, etc... En los últimos años los han incorporado también a los dispositivos móviles, entornos domóticos, teleasistencia e incluso en vehículos militares. Teniendo en cuenta que esta disciplina es relativamente joven respecto al resto, su tecnología todavía se encuentra en intenso desarrollo. Hoy día suelen emplearse sistemas de ASR con muy buenos resultados en dominios cerrados. Por el contrario, encuentran dificultades al enfrentarse a dominios abiertos, habla espontánea o ruido ambiental.

Generación de lenguaje natural

A la generación de lenguaje natural se le conoce por sus siglas en inglés NLG (*Natural Language Generation*). Esta tarea tiene como misión generar texto en lenguaje natural partiendo de información estructurada. El principal problema para conseguir buenos resultados consiste en transformar cierta información de carácter más o menos formal, en un texto rico en fenómenos lingüísticos típicos del lenguaje natural.

El problema de la generación de lenguaje natural suele enfocarse como una toma de decisiones con múltiples restricciones: hay que tener en cuenta el objeto de la comunicación, el contexto, el discurso pasado, etc... Normalmente su uso práctico está relacionado con otras tareas del PLN. Así pues, se utiliza con sistemas de extracción de información y tal como veremos a continuación, conforman una pieza fundamental en los sistemas de diálogo, que suelen emplearlos para comunicarse con el usuario bien a través de texto o mediante el sintetizador de voz.

Sistemas de diálogo

Los sistemas de diálogo persiguen el objetivo de que podamos conversar con un sistema informático con la misma naturalidad con la que conversaríamos con un humano. Entre los problemas o fenómenos que debe tratar el propio gestor del diálogo se encuentran: el contexto del discurso, el tipo de estrategía de diálgo (colaborativa, proactiva, ...), la elipsis, las contrucciones anafóricas, etc...

Desarrollar un sistema de diálogo robusto y eficaz puede considerarse una de las tareas más complejas dentro del PLN. Esta dificultad se debe fundamentalmente a dos factores. En primer lugar, es una de las tareas más jóvenes y por tanto con menos años de investigación y experiencia a sus espaldas. En segundo lugar, un gestor de diálogo por sí solo no tiene utilidad práctica; necesita integrarse con otros módulos típicos como el reconocedor de voz, el sistema de TTS o el de generación de lenguaje natural. En definitiva, el éxito de un sistema de diálogo no depende sólo del módulo que gestiona el diálogo en sí, sino de la perfecta armonía entre sus componentes. En la figura 1.6 puede verse una arquitectura de ejemplo de un sistema de diálogo donde el usuario interaccione mediante voz.

Los sistemas de diálogo han proliferado mucho en los últimos años. Cada vez es más fácil encontrarlos en forma de asistentes virtuales en distintos portales web con una misión orientativa de cara al usuario. Probablemente encontremos en los próximos años nuevos asistentes dedicados a tareas más complejas: gestionar una cita médica, comprar billetes de tren o avión, tienda virtual, etc... Para ilustrar el aspecto de estos asistentes virtuales, en la figura 1.7 aparece Laura, que nos facilita información sobre los contenidos de la web de la Cámara de Comercio de Sevilla.

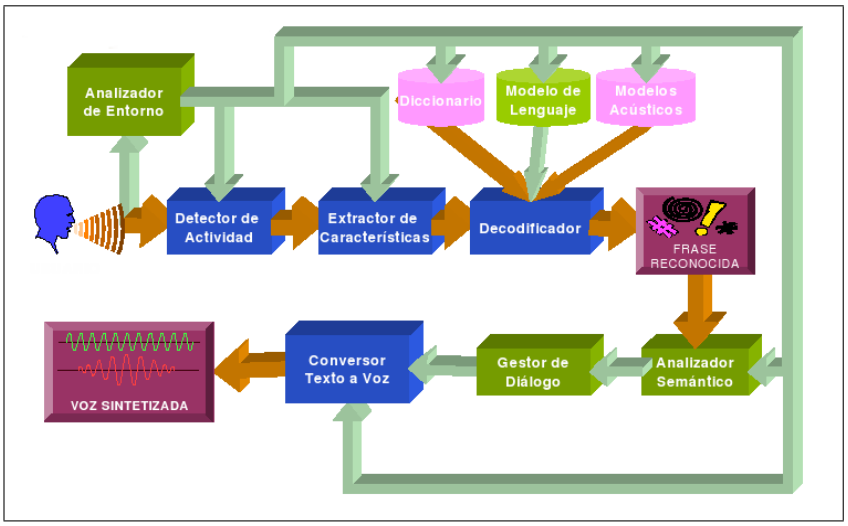

Figura 1.6: Arquitectura de un sistema de diálogo

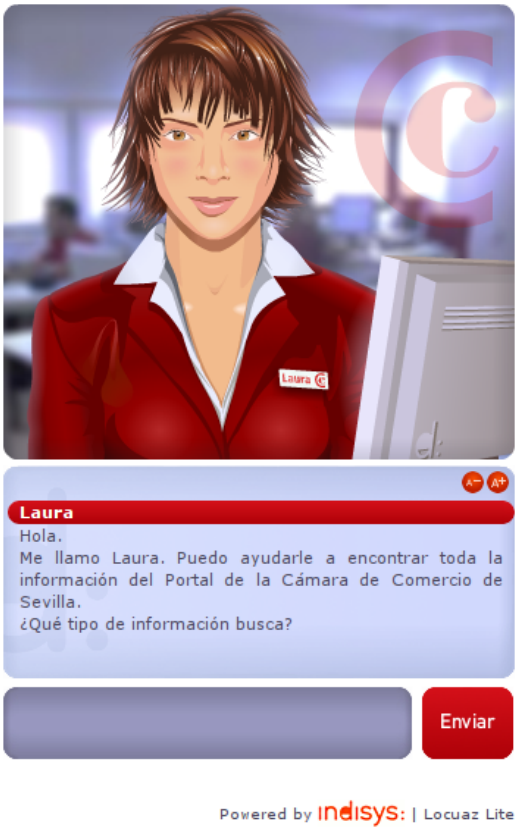

Figura 1.7: Laura, el asistente virtual de la Cámara de Comercio de Sevilla

1.2. Recuperación de Información (RI)

Una de las primeras aplicaciones prácticas de los sistemas de recuperación de información fueron los catálogos digitales, muy utilizados en organismos con gran cantidad de documentos como las administraciones públicas o las bibliotecas. Permitían a los usuarios realizar búsquedas rápidas sobre criterios típicos como título, autor o año de publicación. Estas aplicaciones obtuvieron una gran aceptación, sin embargo, los sistemas de RI obtendrían el impulso definitivo con el auge de Internet. Buscadores como *Yahoo!* o *Google* representan un gran avance, especialmente en el modo en que nos han facilitado el acceso a las distintas fuentes de conocimiento. Este hecho lo confirman diversos estudios que califican a los buscadores web como la primera opción a la que acudimos para nuestras necesidades de información.

Centrándonos en la web y desde el punto de vista práctico del usuario, Internet es una inmensa colección de documentos con cierta relación entre sí. Estas relaciones vienen determinadas por los vínculos o hiperenlaces, que no son más que referencias de un documento a otro. El «problema» es que no existe un *ente* gestor que jerarquice o supervise dichas relaciones, quedando establecidas a criterio de los autores. Este es uno de los motivos por los que preferimos utilizar un buscador y visitar sus resultados ya que confiamos en su neutralidad. Por tanto, uno de los grandes retos de los sistemas de recuperación de información en la web no es otro que superar la inherente subjetividad de los documentos existentes en Internet y devolver al usuario un conjunto de resultados completo y de alta precisión.

Otro aspecto a tener en cuenta es el carácter dinámico de la información existente en la red. Internet está sometido a un proceso de constante cambio y crecimiento, por lo que se necesitan sistemas capaces de evolucionar de manera eficiente. Los buscadores actuales analizan constantemente la web para mantener al día sus índices de documentos. Adicionalmente, toda esta complejidad debe ser abstraída de forma que los usuarios menos experimentados puedan obtener los resultados deseados prescindiendo de palabras clave u otros criterios avanzados de búsqueda.

A la ya de por sí difícil tarea de obtener los resultados de más alta calidad, se le suma la falta de metainformación en los propios documentos. Los esfuerzos encaminados a resolver este problema se engloban dentro del término «*Web Semántica*». La idea consiste en añadir metadatos, anotaciones semánticas y ontologías a la información existente en la web. De este modo, se facilitaría enormemente la interoperabilidad entre los distintos sistemas informáticos y reduciría la intervención humana. Puesto que la inmensa mayoría de los documentos existentes en la web carecen de dicha información semántica, surge un problema de tratamiento de información no estructurada. La falta de modelos que definan formalmente los contenidos deriva en el empleo de técnicas de procesamiento de lenguaje natural, utilizadas sobre todo en tareas de indexación y clasificación. Aunque para estos propósitos casi todos los buscadores sacan partido del propio formato `<HTML>`, éste no es suficiente cuando

entran en juego características propias del lenguaje natural como la ambigüedad, la sinonimia o la polisemia. Algunos ejemplos:

- Si preguntamos a un buscador sobre el término «`capital`», ¿nos devolverá documentos de economía o sobre capitales de países?
- Si queremos información sobre «`coches`», ¿obtendremos resultados donde se les refiera por «`vehículos`» o «`automóviles`»?
- ¿Encontraremos la misma información si preguntamos por «`firma electrónica`» que por «`certificado digital`»?

Éstas y otras preguntas son las que dan origen al presente trabajo de investigación. Se pretende realizar un estudio del estado arte del conjunto de recursos existentes que mediante anotaciones semánticas permitan potenciar los sistemas de recuperación de información en la web. Este salto cualitativo requiere de técnicas de procesamiento del lenguaje natural e inteligencia artificial, encaminadas a producir dicho conocimiento semántico e incorporarlo a las fases de indexación y búsqueda de los sistemas de recuperación de información.

1.2.1. Definición e historia

La definición del término «Recuperación de Información» puede ser muy amplia; un gesto tan típico como encontrar un número de teléfono en un listín telefónico es una —recuperación de información—. No obstante, dentro del contexto que nos ocupa, podemos definirla como la «**búsqueda de documentos que satisfacen las necesidades de información solicitadas**». Estos documentos contienen generalmente información no estructurada, es decir, texto en lenguaje natural. Las búsquedas pueden lanzarse sobre fuentes de distinto tipo: un fichero, una colección de documentos, un catálogo digital, una base de datos o la propia web.

Historia

A lo largo del tiempo, el hombre siempre ha manifestado interés por dejar constancia de su conocimiento. Desde las pinturas rupestres en la época prehistórica, pasando por la gran biblioteca de Alejandría, hasta la Wikipedia en tiempos modernos. Antiguamente el acceso a las fuentes de conocimiento existentes era común de filósofos y eruditos, no siendo útil para la vida diaria. A partir del siglo V, durante el medievo, será el mundo eclesiástico el que conserve los documentos escritos hasta el momento.

Figura 1.8: Grabado de la antigua biblioteca de Alejandría

Remontándonos de nuevo a la era contemporánea, se muestra evidente la necesidad de almacenar el conocimiento y la información, así como recuperarlos de forma eficiente para llevar a cabo actividades cotidianas y sobre todo económicas. En concreto, a finales de los años 40, podemos fijar un punto de inflexión en el uso de los ordenadores para almacenar información en formato digital. Este hecho denota cada vez más la falta de sistemas de recuperación de información veloces y de alta calidad. Desde los comienzos, cualquier sistema tipo Unix ha incorporado la utilidad «`grep`» que permite realizar búsquedas en ficheros. En los sistemas tipo Windows existe una herramienta de escritorio de propósito similar. El problema de estas sencillas aplicaciones es su ineficiencia, ya que se limitan a buscar la secuencia de caracteres línea a línea en los ficheros de entrada.

Uno de los primeros sistemas de recuperación de información fueron los catálogos digitales para bibliotecas. Su funcionamiento era muy simple: se almacenaban los datos de las fichas existentes para luego facilitar la búsqueda de libros por título, autor o fecha de publicación. El avance era evidente puesto que ya no era necesario consultar ficha a ficha hasta dar con la deseada. Sin embargo, resultaba imposible realizar búsquedas sobre los propios contenidos de los libros. En primer lugar porque la tecnología de recuperación de información no estaba lo suficientemente madura. Y finalmente, para digitalizar los libros hubiera sido necesario una cantidad ingente de trabajo por parte de operadores humanos.

En la década de los 70, con la aparición de los *Sistemas de Gestión de Bases de Datos Relacionales* y la necesidad de consultas mucho más veloces, surgieron conceptos como «*formas normales*», «*clave primaria*», «*índice*», etc... Empezaba a forjarse la idea de que para conseguir búsquedas más eficientes los esfuerzos tendrían que centrarse en optimizar las estructuras de datos y no en los propios algoritmos de búsqueda. Así pues, un ejemplo de estos avances fueron los «*Árboles* B^+», una de las estructuras de datos más utilizada en tareas de indexación. De hecho, podemos encontrar alguna implementación en sistemas de ficheros como NTFS, o en sistemas de bases de datos relacionales como MySQL.

Con la llegada de Internet, los sistemas de recuperación de información se enfrentarían a su más complicado reto: la búsqueda web. Como ya hemos mencionado anteriormente, resulta casi imposible enumerar todos y cada uno de los documentos que almacena la red de redes, la cual y para añadir mayor dificultad, está en un constante proceso de cambio y crecimiento. A mediados de los 90 se produce el «boom» de los buscadores web. Por un lado apareció *Yahoo!*, que presentaba al usuario un directorio con distintas categorías por las que navegar. Al mismo tiempo, otros buscadores como *Excite*, *Lycos* o *AltaVista* lideraban una nueva estrategia demasiado ambiciosa para la época: indexar toda la web. Todas estas *Start-Ups* gozaban de un éxito que sobrepasaba sus expectativas, pero su tecnología flaqueaba en un punto: la ordenación objetiva de los resultados que tan importante es para el usuario final. En este escenario nació *Google*, que gracias a la idea del *PageRank* dieron un salto cualitativo enorme, terminando por desbancar a sus competidores. De hecho, durante un período de tiempo, *Yahoo!* y otras compañías utilizaron los índices de *Google* para proporcionar mejores resultados. Los fundamentos del *PageRank* los veremos más adelante en el apartado 1.2.3.

La tecnología sobre búsqueda e indexación ha avanzado de tal manera que hoy tenemos la posibilidad de incorporar un buscador completo en nuestras aplicaciones web. *Lucene* es una herramienta desarrollada por la fundación Apache que nos proporciona todo lo necesario para indexar documentos de diversos tipos: páginas en HTML, PDFs, documentos en formato Word u OpenOffice, etc... Son muchos los gestores de contenidos que incluyen un buscador basado en *Lucene*. Algunos tan conocidos como *Liferay* o *Alfresco*.

Finalmente, los sistemas de recuperación de información han llegado incluso a los sistemas operativos modernos. Las últimas versiones más recientes incorporan aplicaciones que indexan todo tipo de documentos del usuario: ficheros de texto, marcadores de favoritos, emails, conversaciones de mensajería instantánea, calendarios, etc... Sirva como ejemplo *Spotlight* de Mac OS X, *Instant Search* de Windows Vista o *Tracker* de Linux. Todos estas aplicaciones comparten los mismos principios: un proceso en segundo plano se activa periódicamente e indexa la carpeta principal del usuario.

Al realizar la búsqueda, ésta se lanza directamente sobre los índices devolviendo los resultados de forma inmediata.

1.2.2. Evaluación del rendimiento

Puesto que el objetivo de este trabajo es investigar y profundizar en las técnicas que permitan mejorar el rendimiento de un sistema de recuperación de información, necesitamos definir las métricas que, de forma objetiva, evalúen la eficacia de las distintas estrategias que serán objeto de estudio. Previamente a la definición de dichas métricas, abordaremos una serie de términos básicos relacionados con la recuperación de información:

- **Documento:** En el contexto de la recuperación de información, un documento se define por la unidad mínima de resultado que puede devolver el sistema. Por ejemplo, en un buscador web cada página HTML se considera un documento, definida por su dirección o URL.
- **Colección:** Entendemos por colección al conjunto de documentos sobre el cual se efectuará el proceso de indexación y la posterior recuperación de información.
- **Necesidad de información:** Este es un concepto más bien abstracto ya que intenta definir cuál es el tema o asunto sobre el que necesitamos encontrar información. En relación, definimos el término «**consulta**» como la formalización de nuestra *necesidad de información* para que sea tratada por un sistema informático.
- **Relevancia:** Un documento devuelto por el sistema se considera relevante si contiene información útil para las necesidades de información del usuario.

Para medir el rendimiento de un sistema de recuperación de información y la calidad de sus resultados se utilizan las siguientes métricas. Éstas se enuncian con su nombre original en inglés debido a que su traducción puede dar lugar a confusión:

- ***Precision***[3]***:*** Se define como la proporción de documentos relevantes dentro del conjunto de documentos recuperados por el sistema. También puede verse como la probabilidad condicionada de que un documento recuperado sea relevante. Así pues, un sistema de RI que sólo devuelve documentos relevantes tiene un valor de $precision = 1$.

$$Precision = \frac{|documentos\ relevantes\ recuperados|}{|documentos\ recuperados|} = P(relevante|recuperado)$$

[3]Su traducción al castellano no es «precisión» como cabría esperar. En este contexto se acerca más a «eficacia», es decir, cómo de buenos son los resultados devueltos.

- **Recall**[4]**:** Esta medida define la proporción de documentos relevantes de la colección que fueron recuperados por el sistema. O dicho de otro modo, la probabilidad condicionada de que un documento relevante haya sido recuperado. Por tanto, un *recall* = 1 significa que no quedaron documentos relevantes en la colección sin devolver como parte del resultado.

$$Recall = \frac{|documentos\ relevantes\ recuperados|}{|documentos\ relevantes\ en\ la\ colección|} = P(recuperado|relevante)$$

  De forma trivial, podemos obtener fácilmente un *recall* = 1, simplemente devolviendo todos los documentos de la colección. Claro que de esta manera no obtendríamos un buen valor de *precision*.

Intuitivamente buscamos maximizar tanto *precision* como *recall*, sin embargo, la importancia de cada métrica depende de las necesidades del usuario; un estudiante que busca información de cómo escribir un texto en VERSALITAS con LaTeX, sólo necesita un valor alto de *precision* en los primeros resultados del sistema. Es decir, con que un documento tenga la información necesaria será suficiente, no importando el *recall*. Análogamente, si un usuario está buscando comprar un libro sobre el sistema de control de versiones *Subversion*, necesita sobre todo un alto valor de *recall* para poder comparar todos los ejemplares existentes, siendo deseable un grado alto de *precision* pero no imprescindible.

- **F-score:** La métrica mayormente aceptada para englobar *precision* y *recall* es su media armónica ponderada o *F-score*. De este modo se obtiene un resultado más cercano al menor de los dos. La media armónica simple se formula de la siguiente manera:

$$H_{mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}; a_i \in \mathbb{R} \geq 0$$

  La aplicación de la fórmula a los dos elmentos *precision* y *recall* es la siguiente:

$$F = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

[4]No tiene una traducción directa al castellano. El término que más se aproxima es «cobertura».

En realidad, la fórmula inmediatamente anterior se le conoce como F_1 ya que pondera de la misma forma tanto a *precision* como a *recall*. La fórmula generalizada que permite otorgar distintos pesos y que se emplea en multitud de talleres y *workshops* sobre recuperación de información se denomina F_β. De esta fórmula derivan F_2 donde se le otorga el doble de valor a *recall* (aplicable al ejemplo del libro de Subversion), o $F_{0,5}$ que hace que *precision* valga el doble (ejemplo de las VERSALITAS con LaTeX).

$$F_\beta = \frac{(1+\beta^2) * precision * recall}{(\beta^2 * precision) + recall}; \beta \in \mathbb{R} \geq 0$$

Buscando una perspectiva distinta, un sistema de recuperación de información también puede verse como un clasificador binario: partiendo de la consulta del usuario, hay que clasificar todos los documentos de la colección como relevantes o no relevantes y conformar el resultado con los primeros. En este proceso de clasificación pueden cometerse dos tipos de errores:

- **Falso Positivo:** Clasificar un documento no relevante como relevante. Al devolverlo como parte del resultado penalizamos el valor de *precision*.

- **Falso Negativo:** Clasificar un documento relevante como no relevante. Al no devolverlo, penalizamos el valor de *recall*.

Estos tipos de errores y la relación entre recuperación y relevancia se resumen en el cuadro 1.1. Además permitirán redefinir las fórmulas de *precision* y *recall* en función de dichos conceptos.

	Relevante	**No Relevante**
Recuperado	`Positivo (p)`	`Falso Positivo (fp)`
No Recuperado	`Falso Negativo (fn)`	`Negativo (n)`

Cuadro 1.1: Relación recuperación-relevancia

$$Precision = \frac{Positivos}{Positivos + FalsosPositivos}$$

$$Recall = \frac{Positivos}{Positivos + FalsosNegativos}$$

Gracias a éstos últimos conceptos introducidos podemos enunciar una nueva métrica que define la tasa de acierto en la clasificación binaria «relevante-no relevante»:

- **Accuracy**[5]**:** Define la proporción de documentos de la colección clasificados correctamente.

$$Accuracy = \frac{Positivos + Negativos}{Positivos + Negativos + Falsos\ Positivos + Falsos\ Negativos}$$

Hay una razón por la cual *accuracy* no es una métrica muy empleada: en un sistema de recuperación de información, y muy especialmente en aquellos con una inmensa colección de documentos, el 99'9 % de los resultados pertenecerán a la categoría de «no relevante». Este sesgo es un riesgo para la optimización de un algoritmo de clasificación. Maximizar *accuracy* puede dar lugar a resultados no deseados, por ejemplo, obteniendo una alta tasa de «falsos positivos» y consecuentemente disminuyendo el valor de *precision* del sistema.

1.2.3. Búsqueda en la web

Nos centraremos en el estudio de los sistemas de recuperación de información en la web debido a su interés técnico y comercial, gran popularidad y fácil acceso. En primer lugar, enunciaremos las características más importantes de la colección de documentos más grande jamás creada: Internet.

- Los autores de las páginas web han generado una ingente cantidad de contenidos en una multitud de idiomas y en varios miles de dialectos. Por tanto, para una indexación correcta es necesario emplear distintas técnicas de procesamiento de lenguaje natural y otra serie de operaciones lingüísticas adicionales.
- La libertad de expresión[6] y estilo es una característica inherente a la web. Podemos encontrar desde las más correctas formas de redacción, hasta toda una serie de expresiones mal formadas, contradictorias, suposiciones, opiniones objetivas, subjetivas, etc...

[5] En este caso sí podríamos traducir por «exactitud».
[6] Salvo en aquellos países que aún hoy tienen controles de censura a ciertos sitios web.

- El entramado de páginas web puede verse como un grafo: cada página web es un nodo, por otro lado, los hiperenlaces serían las aristas del grafo. Gracias al estudio de la forma de este grafo podrá generarse un ranking de los nodos más importantes.
- Una información muy preciada para tener éxito en la búsqueda web es entender cómo se comportan los usuarios. Los sistemas de recuperación de información tradicionales eran usados típicamente por profesionales con un alto nivel de formación y capacitados para realizar consultas complejas. En los sistemas de búsqueda web, el usuario no suele tener ninguna formación previa sobre la sintaxis que puede emplear. Por tanto, es importante poner énfasis en la flexibilidad y usabilidad de estos sistemas.

Arquitectura de un sistema de recuperación de información en la web

Todos los sistemas de recuperación de información para la web desarrollados hasta el momento tienen una arquitectura bastante común (ver figura 1.9). Por un lado se encuentra la colección de documentos que conforman la web, representados en la figura por documentos `<HTML>` relacionados entre sí. En el lado opuesto tenemos al usuario, que es quien realiza las consultas al sistema. El flujo de información y los agentes involucrados en esta tarea se representan con las flechas de color verde. Finalmente, el sistema de recuperación de información en sí mismo, cuyo proceso de indexación se representa con las flechas rojas. Como puede observarse en la figura, existen dos procesos independientes que involucran al sistema de recuperación de información: la indexación y la búsqueda.

- La **indexación** es una operación que se realiza con cierta periodicidad y consiste en el análisis de los documentos de la colección, es decir las páginas web, para crear los índices de términos que permitan acceso a los mismos de la manera más eficiente posible. Para alimentar al sistema de indexación se necesita de otro proceso que vaya recorriendo el grafo que representa la web en busca de nuevos nodos para analizar. A este último proceso se le conoce como «*crawler*» o «*araña*».
- El proceso de **búsqueda** comienza cuando un usuario realiza una consulta al servidor web del sistema de recuperación de información, éste se encarga de transformar la consulta en una petición a la base de datos de índices donde se buscarán los nodos que conformarán el resultado. Normalmente los buscadores web presentan la lista de resultados ordenándolos según su relevancia estimada, basada en algún algoritmo puntuación como veremos a continuación. Algunos buscadores también presentan sugerencias a la consulta cuando detectan que el conjunto de resultados obtenido es escaso o poco relevante, muchas veces esto se debe a una consulta mal planteada o con faltas de ortografía.

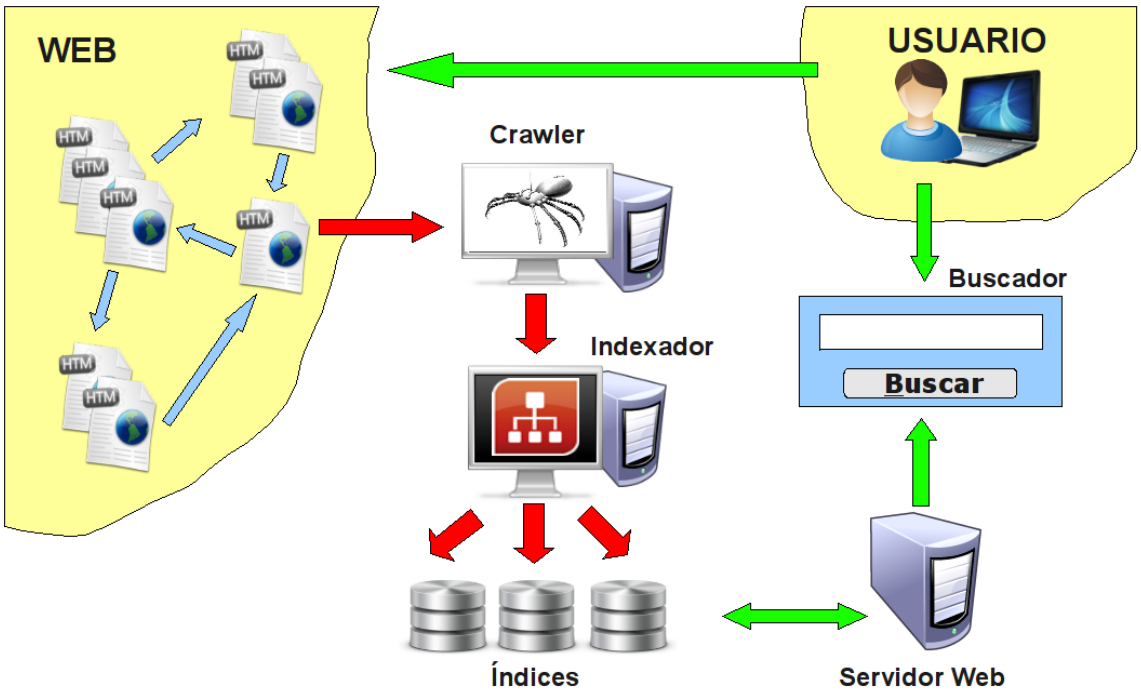

Figura 1.9: Arquitectura de un sistema de RI para la web

Algoritmo de PageRank

Anteriormente se ha mencionado la importancia de ofrecer al usuario el conjunto de resultados ordenados por relevancia. La dificultad de esta tarea radica en encontrar un sistema justo de puntuación, que supere la subjetividad de la web y que se adapte fácilmente a las necesidades de información especificadas por el usuario. Dentro de todos los algoritmos de puntuación existentes, el que mayor éxito ha cosechado hasta el momento es el denominado *PageRank* de Google.

En 1999, los fundadores de Google *Larry Page*[7] y *Sergey Brin*, junto con sus profesores de Stanford *Rajeev Motwani* y *Terry Winograd*, publicaron un artículo titulado «*The PageRank Citation Ranking: Bringing Order to the Web*» [PAG98]. En dicho artículo se enuncian los fundamentos del PageRank: a pesar de que la importancia de una página web no puede medirse de manera objetiva, ya que depende de las distintas necesidades de información de los usuarios, sí podemos confiar ampliamente en la naturaleza democrática de los hiperenlaces que hay en la web. El PageRank interpreta un enlace de la página **A** a la **B** como un voto de **A** para **B**. La clave consiste en no

[7]El principal autor es quien dio origen al nombre de Page-Rank.

contar únicamente el número de votos recibidos, el algoritmo también tiene en cuenta la relevancia de la página que emite dicho voto. De este modo, el voto de una página importante puede valer mucho más que varios votos de páginas menos relevantes. Podemos visualizar este comportamiento con la simpática ilustración de la figura 1.10.

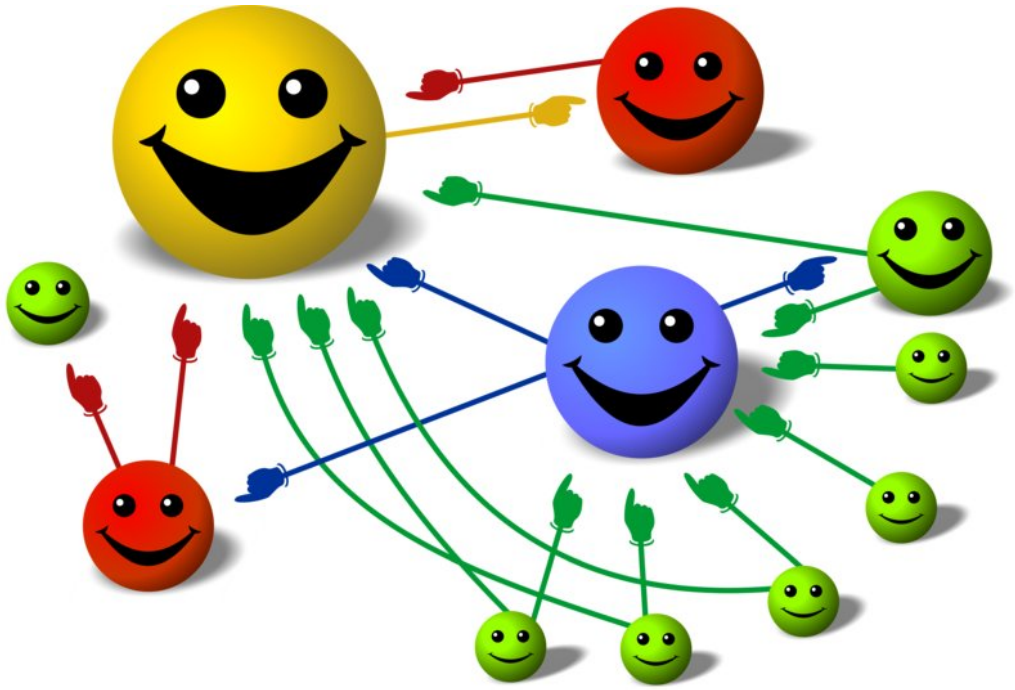

Figura 1.10: PageRank en funcionamiento

Para el cálculo del valor de PageRank de una página «A» se utiliza la siguiente fórmula:

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

Donde:

- **PR(A)** es el PageRank de la página «A».
- El coeficiente **d** representa la probabilidad de que un navegante continúe pulsando en los enlaces y navegando por el grafo web. Según estudios realizados se le estima un valor de 0,85.

- El término **(1-d)** representa por tanto la probabilidad de que un navegante deje de pulsar en los enlaces y escriba una nueva URL aleatoria en la barra de direcciones del navegador. Esta acción sería equivalente a «teletransportarse» de un nodo a otro del grafo que no están unidos por ninguna arista.

- Los $\mathbf{PR}(T_i)$ son los PageRank de las páginas T_i que enlazan hacía la página «A».

- Los $\mathbf{C}(T_i)$ son el número de enlaces totales que salen de T_i.

Debido a la importancia comercial de aparecer entre los primeros resultados, existen varias técnicas para intentar manipular el valor de PageRank asociado a una página. Desde que se crearon las primeras «*páginas spam*» hay una batalla continua entre los desarrolladores de los motores de búsqueda y los creadores de dichas páginas. Por tanto, es muy importante a la hora de diseñar un algoritmo de puntuación tener en cuenta estos detalles si se quieren obtener resultados de calidad. Las fórmulas *anti-spam* son uno de los mejores secretos guardados de los grandes buscadores comerciales.

Indexación web

A grandes rasgos, la indexación web consiste en analizar las páginas seleccionadas por el *crawler* y almacenar su contenido de forma que, al preguntar por un término concreto nos devuelva los identificadores de las páginas que contienen dicho término. Además, la indexación debe realizarse teniendo en cuenta los valores de PageRank asociados a cada página, de modo que los identificadores sean devueltos por orden de relevancia. A esta breve descripción hay que añadir una serie de matices:

- Por almacenar el contenido se entiende únicamente el texto plano; etiquetas de estilo, imágenes, frames y demás componentes visuales no son tenidos en cuenta en este punto.[8]

- Para identificar una página web se suele emplear su URL puesto que es un identificador único válido. En este apartado entran en conflicto las páginas web dinámicas ya que hay que tratar el paso de parámetros dentro de la propia URL tal y como lo hacen tecnologías tipo JSP, ASP.NET o PHP.

- Las estructuras de datos deben estar preparadas para almacenar los términos de forma que puedan recuperarse eficientemente tanto si se pregunta por un término aislado como una combinación de los mismos.

[8] Google lanzó en 2001 su servicio de búsqueda de imágenes por el cual pasaron a almacenar una versión reducida de la misma que es la que se muestra al usuario.

En la figura 1.11 se muestra de forma general cómo funciona el módulo de indexación. El *parser* se encarga de la extracción del texto plano a partir de las páginas `<HTML>`. Estas versiones de los documentos «en crudo» se guardan en una caché para un uso posterior, por ejemplo, para mostrar junto a los resultados pequeños resúmenes o «*snippets*» donde se resalta los términos de la consulta en su contexto. Por otro lado, el constructor del índice es el responsable de almacenar qué términos se encuentran en qué documentos, ordenando ya los resultados por relevancia según su PageRank.

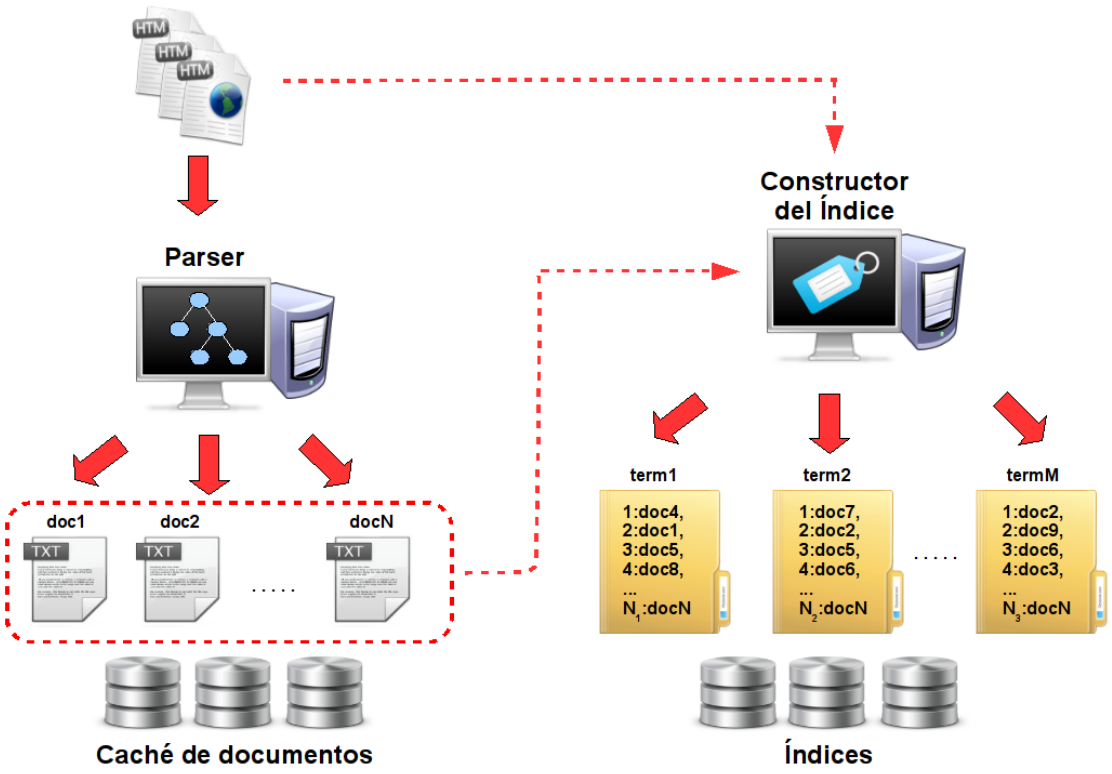

Figura 1.11: Detalle del módulo de indexación

Otro aspecto a tener en cuenta en el proceso de indexación es la localización de los términos dentro de una página; una palabra concreta que aparece en el título de una página **A** debería de otorgarle mayor puntuación respecto a otra página **B** que la contiene en el cuerpo. Esta forma de sacar partido al propio formato `<HTML>` permite ajustar los criterios de búsqueda conforme a la intención original del autor de la página. De hecho, los buscadores más avanzados tienen en cuenta decenas de

criterios como frecuencia y densidad de aparición, colocación dentro de la página, tamaño de la fuente, estilo, etc... El problema es cómo determinar los pesos óptimos que debería tener cada criterio. Esta «*fórmula mágica*» es también uno de los secretos mejor guardados de los buscadores comerciales. Normalmente se establecen unos criterios «a mano», para posteriormente usar técnicas de aprendizaje automático.

Por último, en el proceso de indexación también se deben tratar los aspectos que vienen determinados por las características especiales de la web. Al tener que indexar una cantidad tan grande de documentos es necesaria una arquitectura de *clusters distribuidos*, y por tanto, contar con unas estructuras que permitan dividir el trabajo entre los distintos nodos del cluster. Dichas estructuras tendrán que fusionarse luego para permitir las consultas de los usuarios. Adicionalmente, las estructuras deben facilitar también la actualización parcial de una parte de los índices. Esta necesidad viene del carácter dinámico de la web, ya que indexar por completo todos los nodos una y otra vez no es viable. Para conseguir este comportamiento se utilizan una serie de índices auxiliares que más tarde se fusionarán con el principal.

Capítulo 2

Hipótesis y Objetivos

Una vez introducidas las materias base de «Procesamiento de Lenguajes Naturales» y «Recuperación de Información», en este segundo capítulo trataremos de dar una perspectiva general sobre las hipótesis o nuevas ideas que permitirían mejorar los resultados en la búsqueda web tradicional. Finalmente se plantean los objetivos para realizar un estudio en profundidad sobre la materia.

2.1. Hipótesis

Los sistemas de recuperación de información en la web actuales se basan en técnicas de análisis estadístico superficial de los contenidos para llevar a cabo las tareas de indexación y búsqueda. La tecnología empleada hoy día no permite discernir aún con perfección la relevancia de un documento ante la expresión de una necesidad de información. Tanto es así que a veces no encuentran los documentos solicitados o devuelven documentos no deseados y en un *ranking* que no siempre es el adecuado.

Estos defectos se deben en una gran medida a que los buscadores tradicionales trabajan sin realizar un análisis profundo de los propios contenidos. Las técnicas estadísticas citadas no hacen un uso exhaustivo de las relaciones semánticas que aporta el conocimiento lingüístico. Aunque se han realizado avances como por ejemplo la incorporación del análisis morfológico, que permite relacionar palabras derivadas o distintas conjugaciones de verbos, seguimos sin poder determinar que un usuario que busca documentos sobre «*firma electrónica*» muy probablemente también esté interesado en los documentos que hablen de «*certificados digitales*».

Para conseguir nuevos niveles de refinamiento en las búsquedas creemos que son imprescindibles las aportaciones del procesamiento del lenguaje natural y la lingüística computacional. Estas disciplinas vienen a ayudar donde la simple estadística no puede llegar: generar conocimiento semántico. Como producto principal de esta agrupación tecnológica aparecerán los buscadores semánticos, capaces de entender con-

sultas en lenguaje natural y resolver los problemas de ambigüedad y semántica del lenguaje.

2.2. Objetivos

Los objetivos fundamentales de esta memoria de investigación pasan por desarrollar un estudio del estado del arte sobre la «*búsqueda semántica*» en todas sus variantes. Así como evaluar las distintas técnicas existentes que pueden ayudar en la generación de conocimiento semántico desde información no estructurada o en lenguaje natural.

Se tomará como punto de partida la «*web semántica*», ya que comparte muchos de los objetivos que se persiguen con la aportación del procesamiento del lenguaje natural a la recuperación de información. Realizaremos un estudio de las principales líneas de investigación aplicables así como de otras líneas secundarias relacionadas. Se pretende también elaborar un informe sobre los principales foros, universidades o investigadores relevantes para facilitar el seguimiento del estado del arte a cualquier persona interesada en la materia.

Capítulo 3

Estado del Arte

Para el estudio del estado del arte de una disciplina es importante comenzar sentando las bases de los conceptos a desarrollar. Es por ello que este capítulo empieza describiendo el proceso en el desarrollo web hasta alcanzar la idea de la «*Web Semántica*». Seguidamente se dedica un apartado al principal objeto de estudio de esta memoria: la «*Búsqueda Semántica*». A continuación se presentan las principales líneas de investigación sobre la materia así como algunas de carácter secundario. Por último, se realiza un estudio del contexto investigador detallando los congresos, revistas, entidades y personalidades más relevantes en el estudio de la «*Búsqueda Semántica*».

3.1. Hacia la web semántica

Actualmente, la web está poblada con numerosos documentos en formato `<HTML>`. Este lenguaje, tal y como hemos mencionado anteriormente, sólo sirve para definir la apariencia gráfica de los documentos, limitando las posibilidades de clasificación de los mismos. Con estas carencias resulta bastante difícil que los creadores de páginas web puedan aportar algo positivo a las tareas de indexación de los buscadores. Se hace patente la necesidad de un nivel adicional de especificación donde se determine la semántica de los contenidos.

La evolución tecnológica que viene produciéndose en la web en los últimos años va claramente encaminada a conseguir dos objetivos: involucrar a los usuarios en el desarrollo de los contenidos y la interoperabilidad entre todos los sistemas informáticos sin intervención humana. El primero de ellos es una realidad patente hoy día; sitios web como la *Wikipedia*, *YouTube*, *Tuenti* o los propios *blogs* tienen un denominador común: la participación de los usuarios en la creación de los contenidos. A este fenómeno se le conoce como *Web 2.0* y está resultando todo un éxito.

En cuanto a la interoperabilidad de sistemas, el objetivo es conseguir que la metainformación semántica añadida describa los contenidos web y las relaciones en-

tre los mismos. Para dotar de una utilidad práctica a dicha información ésta debe definirse en un formato fácilmente procesable por un computador. En los últimos años se han producido grandes avances en el desarrollo de ontologías y lenguajes dedicados a su tratamiento como *RDF*, *OWL* o *SPARQL*. A esta nueva tendencia donde los contenidos web se autodescriben se le conoce como *Web Semántica* o *Web 3.0*. Afortunadamente, la Web 2.0 está influyendo positivamente en la creación de su sucesora ya que fomenta en los usuarios la costumbre de «etiquetar» sus contenidos; por ejemplo, etiquetando un vídeo subido en YouTube o añadiendo etiquetas de los amigos que aparecen en una foto de Tuenti. De esta manera se consigue poblar el nivel semántico y el uso de dicha información por parte de software de terceros sin intervención humana.

La ilustración «Evolución de la web» (ver fig. 3.1) sirve para hacernos una idea gráfica de la transformación tecnológica y social que ha tenido lugar en los últimos años, además de mostrar el escenario ideal que se quiere alcanzar con la implantación de la web semántica:

- **Web 1.0:** La idea original de la web era separar a los creadores de contenidos por un lado, dejando a los lectores por otro e inhabilitando a estos últimos para cualquier colaboración creativa. Este ha sido el modelo de funcionamiento de la web durante todos estos años y de hecho es muy probable que lo siga siendo en determinadas áreas.

- **Web 2.0:** La acuñación del término se le atribuye a *Tim O'Reilly*, y engloba a todas aquellas tecnologías relacionadas con el ámbito web que han permitido involucrar a los usuarios en la creación de contenidos web. Este nuevo modelo se sirve del comportamiento humano, que tiende a agruparse según sus gustos y/o necesidades. Así pues, bajo este paradigma han surgido nuevas aplicaciones como las redes sociales, los wikis o los blogs.

- **Web Semántica o Web 3.0:** En contra de lo que pudiera parecer, la idea de la web semántica es previa a la Web 2.0. La incorporación de metadatos semánticos a los contenidos web era parte del plan original de *Tim Berners-Lee* en el desarrollo de la *World Wide Web*[1], sin embargo, diversas razones le impidieron llevar a cabo su cometido. El objetivo principal de la web semántica es conseguir que los distintos sistemas informáticos sean capaces, de manera autónoma, de analizar los contenidos publicados y a su vez generar nuevos contenidos. Los usuarios saldrían ganando con aplicaciones que permitieran la obtención de información de distintas fuentes: los vuelos más baratos con ofertas turísticas adicionales, ofertas inmobiliarias con información sobre créditos hipotecarios, etc... Estas aplicaciones podrían mezclar tanto la información en formato texto como contenido multimedia, obteniendo también mejores estudios de mercado

[1] Principios de los años 90 en los laboratorios del *CERN*.

o la publicidad personalizada para el cliente que tanto valor tiene para las agencias marketing y publicidad.

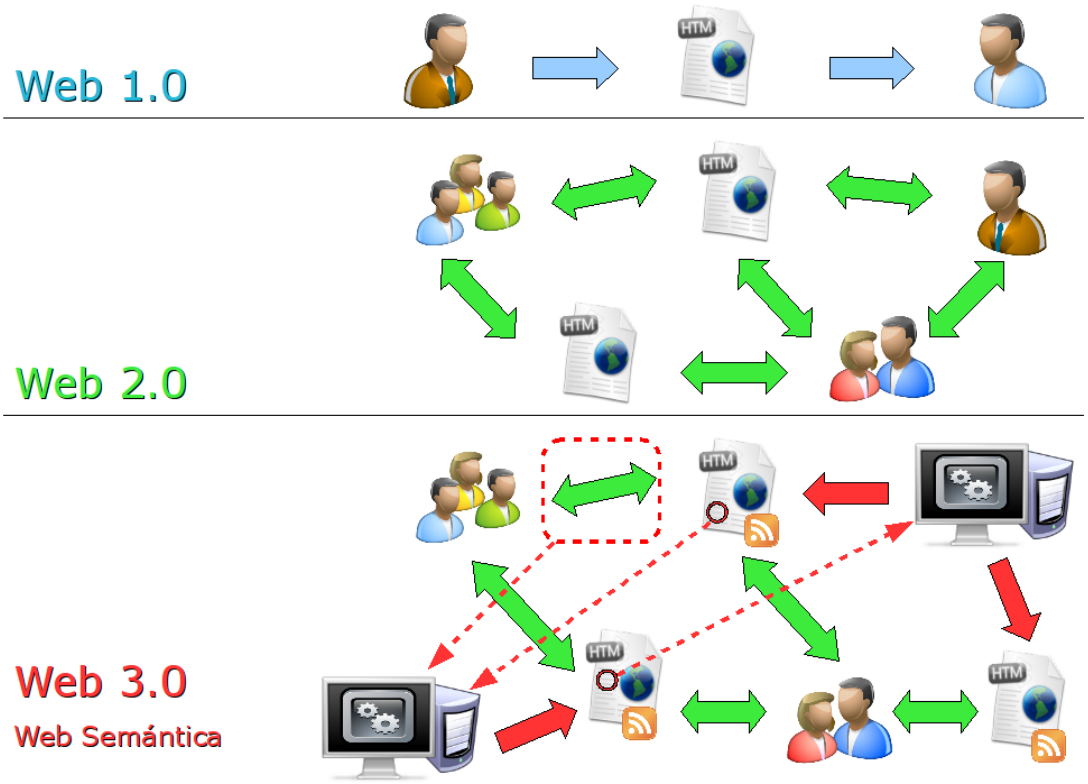

Figura 3.1: Evolución de la web

Centrándonos ya exclusivamente en la web semántica, son multitud de proyectos los que giran alrededor de este concepto. En primer lugar hay que destacar a los lenguajes que se están imponiendo como estándares y que permiten tanto «etiquetar la web» como recuperar la información. Estos son los tres pilares fundamentales en los que se describe la metainformación ontológica hoy día:

- **RDF:** Sus siglas responden a «Resource Description Framework». Se trata de un lenguaje ideado para proporcionar una información descriptiva de los recursos que se encuentran en la web. Se fundamenta en tripletas «*sujeto-predicado-objeto*». El *sujeto* es el recurso que se describe, el *predicado* es la propiedad o la relación que se establece sobre el sujeto, y finalmente el *objeto* es el valor de la propiedad o el otro sujeto con el que se relaciona.

- **SPARQL:** Estas siglas definen al lenguaje estándar de consulta sobre RDF, y significan «SPARQL Protocol and RDF Query Language». Permite búsquedas sobre recursos de la web semántica utilizando distintas fuentes de información. Ha sido concebido con la idea de que una sola consulta a través de diferentes almacenes de datos es mejor que múltiples consultas por separado.
- **OWL:** Es un lenguaje tipo `<XML>` que permite definir ontologías serializables en formato RDF. Su acrónimo significa «Ontology Web Language». Se le considera pieza fundamental en el desarrollo de la web semántica, ya que la definición de ontologías en OWL son el punto de partida para lograr que otras aplicaciones puedan realizar razonamientos sobre los datos de forma automática. Estas ontologías en formato OWL incluyen definiciones de conceptos dentro de un dominio y la relaciones existentes entre dichos conceptos.

Una de las principales necesidades para conseguir los objetivos de la web semántica consiste en publicar, en formato RDF, grandes cantidades de información útil para las aplicaciones de uso cotidiano: enciclopedias, catálogos de bibliotecas, discografías musicales, información geográfica, etc... Si tal cantidad de información fuese pública, se podrían emplear para mejorar los resultados de búsqueda y la experiencia de usuario al navegar por internet, dando la sensación de que está todo interconectado. Esta es la meta del proyecto *LOD,* «*Linking Open Data*» del W3C, que está construyendo una enorme red de fuentes de información abierta y de uso cotidiano en formato RDF. La gran ventaja de este proyecto radica en la facilidad con la que podemos navegar desde un elemento dentro de una fuente determina, a otro elemento relacionado de otra fuente distinta. Gracias a estos enlaces RDF, también es posible que un *crawler semántico* devuelva información estructurada de distintas fuentes y no simples enlaces a páginas `<HTML>`. En la figura 3.2 puede verse el estado actual de la nube del proyecto «*Linking Open Data*» a fecha de Marzo de 2009, con aproximadamente 4.5 billones de tripletas RDF y unos 180 millones de enlaces RDF. Algunas de las fuentes de información más importantes que ya están vinculadas al proyecto son:

- DBpedia: Es un proyecto destinado a extraer información estructurada directamente desde la Wikipedia.
- FOAF: El proyecto «*The Friend of a Friend*» almacena descripciones sobre personas y los vínculos entre los mismos. Muy similar a las conocidas redes sociales como *Tuenti* o *Facebook*.
- Geonames: Es una base de datos geográfica mundial.
- Freebase: Base de datos con información de temas de muy diversa índole.
- ACM: La prestigiosa sociedad científica permite el acceso a toda la base de datos de sus publicaciones en formato RDF.

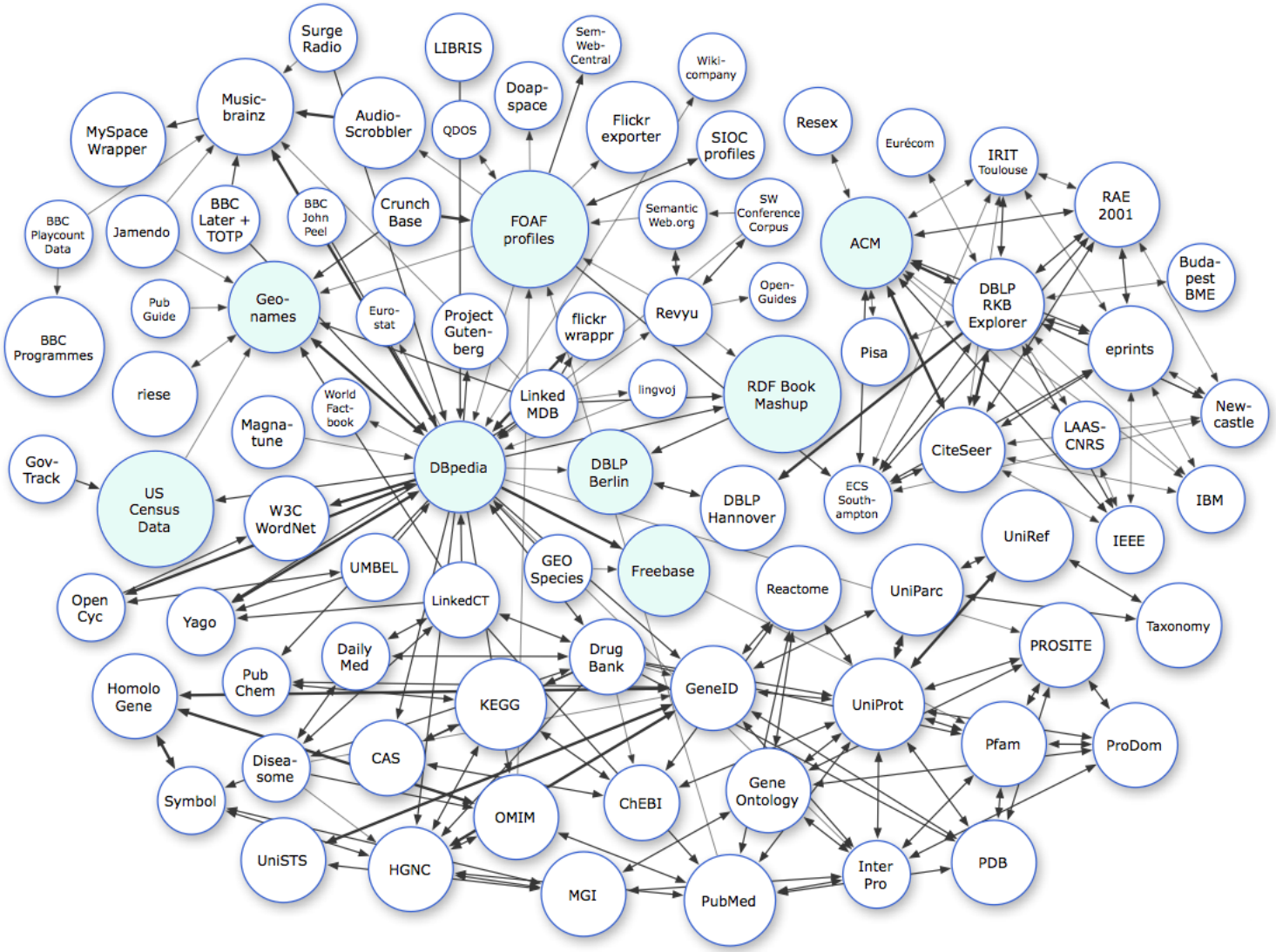

Figura 3.2: Nube del proyecto «*Linking Open Data*» a Marzo de 2009

3.2. Búsqueda semántica

Durante la elaboración de esta memoria se ha realizado una recopilación de artículos relacionados con la «*Búsqueda Semántica*». Una de las primeras conclusiones es que dicho término no tiene una definición aceptada por toda la comunidad científica. Tras una lectura detallada de los mismos se induce que por «*Búsqueda Semántica*» se pueden entender en realidad tres variantes:

- «**Búsqueda en la web semántica**» (apartado 3.2.1).
  Variante de la recuperación de información que saca provecho de las características de la web semántica para enriquecer los resultados.
- «**Indexación y búsqueda con información semántica**» (apartado 3.2.2).
  Uso de información semántica en la fase de indexación para mejorar los resultados de búsqueda.
- «**Búsqueda en lenguaje natural**» (apartado 3.2.3).
  Recuperación de información mediante consultas realizadas en lenguaje natural en lugar el uso tradicional de palabras clave.

Teniendo presente esta triple visión sobre qué es la «*Búsqueda Semántica*», se ha realizado un análisis de la bibliografía (pág. 59) clasificando los 28 artículos recogidos con el objetivo de visualizar en qué líneas se mueven los investigadores. En la figura 3.3 se muestra un diagrama de *Venn* con las intersecciones de los tres conjuntos. Cuando un artículo aparece en una zona de intersección es porque es aplicable a más de una variante. A priori, llama la atención que la mayoría de los artículos se concentren en las intersecciones entre la «*Búsqueda en la web semántica*» y la «*Indexación y búsqueda con información semántica*». Otra de las conclusiones a extraer es que la «*Búsqueda en lenguaje natural*» parece no tratarse como un problema aislado sino que guarda una estrecha relación con la «*Indexación semántica*». En los siguientes subapartados estudiaremos en profundidad las distintas variantes de la «*Búsqueda Semántica*».

3.2.1. Búsqueda en la web semántica

La acepción que tiene más adeptos es considerar la «búsqueda semántica» como la nueva generación de algoritmos de «búsqueda en la web semántica». Esta corriente tiene como premisa que, en la nueva web semántica, una búsqueda no puede limitarse a un simple problema de recuperación de documentos. Hay que sacar provecho de las nuevas interconexiones para dotar de un valor añadido a los resultados. De hecho, poco a poco se va vislumbrando esta nueva tendencia en los buscadores comerciales, los cuales incorporan publicidad orientada, enlaces a elementos multimedia de otras

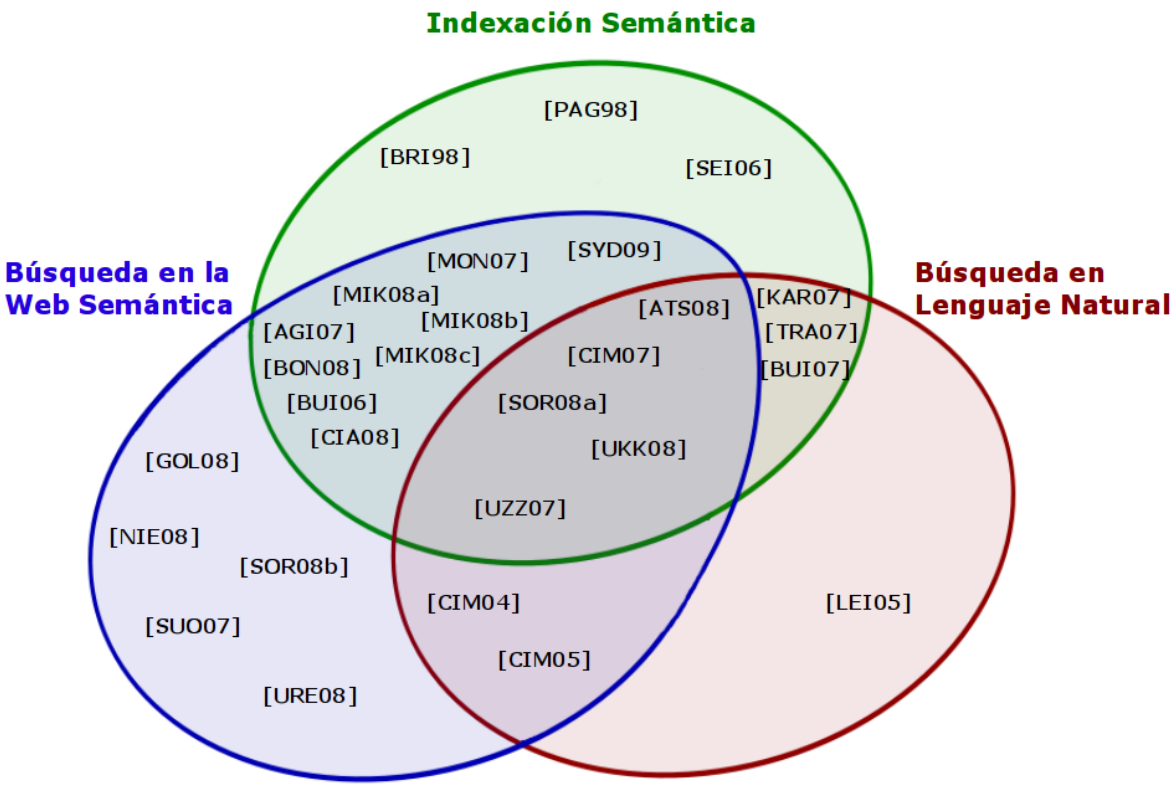

Figura 3.3: Clasificación de publicaciones

webs (e.g.: vídeos de YouTube, fotos de Flickr, etc...) o incluso mapas con información geográfica de localización. Sirva como ejemplo la búsqueda del término «pizza» en Google (ver fig. 3.4).

Otra de las metas que persiguen los algoritmos de búsqueda orientados a la web semántica consiste en modelar la *intención* del usuario y facilitarle las tareas. Siguiendo con el ejemplo anterior, si en lugar de «pizza» hubiéramos buscado por «pedir una pizza», Google nos habría devuelto una serie de documentos donde aparecerían las palabras «pedir» y «pizza», pero no nos ayuda en la tarea de saciar nuestro apetito. Lo ideal sería que un buscador fuese capaz de extraer la semántica de nuestra petición: estamos hambrientos y nos apetece una pizza pero no estamos muy seguros así que buscamos algunas sugerencias. Lo primero que debería hacer el buscador semántico es obtener mi situación geográfica para únicamente tener en cuenta las pizzerías de alrededor. Posteriormente accedería a mi perfil de Tuenti para ver qué amigos ya han probado las pizzas de los restaurantes cercanos, para volcar sus comentarios al respecto. Finalmente, debería mostrar algunas fotos de pizzas de dichos restaurantes que seguro terminan de ayudarnos en la decisión.

Figura 3.4: Búsqueda del término «pizza» en Google

Del anterior ejemplo se extraen una serie de conclusiones que definen los retos que debe afrontar la investigación en este tipo de búsquedas:

- ¿Cómo modelamos la *intención* que expresa el usuario? Indudablemente hay un peso muy importante en tareas de procesamiento del lenguaje natural y aprendizaje automático que tendrán que evolucionar en este sentido.
- ¿Cuántos resultados se deberían mostrar? El usuario quiere realizar una tarea, ya no está buscando documentos que le ayuden a realizarla. Quiere saltarse el último paso de filtrar los resultados devueltos por los buscadores convencionales.
- ¿Qué modelo debería emplearse en la web para conseguir dar respuesta a las intenciones de los usuarios? ¿Es suficiente el modelo de ontologías?
- ¿Cómo medimos ahora la relevancia de los documentos? En los modelos tradicionales, la relevancia se mide en función de criterios totalmente objetivos, tal y como vimos al estudiar el PageRank (pág. 23) Al expresar intención, el usuario está implícitamente solicitando que se atiendan sus intereses particulares. Las

métricas de *precision* y *recall* ya no son suficientes para evaluar la satisfacción del usuario.

En definitiva, este enfoque de la búsqueda semántica plantea que un buscador ya no es una interfaz para recuperar documentos, sino un medio para realizar tareas basado en la web.

3.2.2. Indexación y búsqueda con información semántica

Los buscadores tradicionales intentan localizar los términos de la consulta dentro de la colección de documentos que tienen indexada; priorizan que las palabras aparezcan en el mismo orden y luego buscan en otros documentos pero flexibilizando la colocación de las mismas. En los últimos años se han introducido técnicas simples de procesamiento del lenguaje natural, por ejemplo las de corrección ortográfica, que detectan errores sencillos y realizan sugerencias. Otro caso es el de las reglas morfológicas que utilizan los buscadores para devolver los mismos resultados con términos en singular y en plural, diminutivos, etc...

Con estas nuevas técnicas de procesamiento de lenguaje natural se abren nuevas posibilidades para mejorar los resultados. Como ya vimos en la introducción a la recuperación de la información (pág. 14), existen situaciones en las que un análisis más profundo de la consulta y un conocimiento semántico de los contenidos indexados, permitiría la resolución de distintos problemas como los siguientes:

- **Términos polisémicos:** Tomando el castellano como ejemplo de partida, existen multitud de palabras polisémicas que ante una determinada búsqueda generan muy diversos y relevantes resultados. Sirva de ejemplo «*capital*», «*masa*»[2] o «*herramienta*»; que respectivamente dan lugar a resultados en Google sobre economía, física y utensilios, dejando de lado a las capitales de países, la repostería y las utilidades software. Una buen buscador semántico debería, por un lado, advertirnos de que hay otros significados del término por si queremos refinar la búsqueda, y por otro, indexar las palabras polisémicas en su contexto, para de este modo, no mezclar las distintas acepciones del término en los resultados devueltos al usuario.
- **Sinonimia:** Este es otro de los grandes retos de la indexación semántica. La riqueza de un lenguaje suele a veces medirse en la cantidad de palabras que contiene su vocabulario para referirse a un mismo elemento. Por ejemplo, en zonas de lluvia seguro que agradecen poder referirse a la misma como: llovizna, aguacero, chaparrón, diluvio, tromba, chirimiri o calabobos. Sería interesante que ante la búsqueda de un término con tantos sinónimos, la indexación actuase

[2] La Wikipedia da hasta 38 acepciones distintas del término «masa».

de forma semántica y los interpretara con el mismo significado. De esta forma obtendríamos resultados similares ante consultas sobre «*coches*», «*vehículos*» o «*automóviles*».

- **Expresiones equivalentes:** Esta es una forma especial de sinonimia. En este caso son varios los términos los que pueden combinarse generando siempre el mismo significado. Un indexador semántico tiene que tenerlos en cuenta en su contexto para enlazarlos posteriormente en una búsqueda. Un claro ejemplo es el uso indistinto de «*firma electrónica*» o «*certificado digital*» para referirse al mismo concepto.

En conclusión, estas técnicas de procesamiento del lenguaje natural presentan otra acepción del término «búsqueda semántica», donde en lugar de estudiar lo relacionado con la web y sus modelos ontológicos, se centran en sacar partido a la propia semántica del lenguaje para producir mejores resultados. La gran ventaja respecto a la solución basada en ontologías y metainformación, es que las técnicas de procesamiento del lenguaje natural pueden aplicarse de manera inmediata a los contenidos actuales sin necesidad de cambio. Por tanto, este es el movimiento estratégico que están desarrollando las soluciones comerciales como paso previo a la implantación definitiva de la web semántica.

3.2.3. Búsqueda en lenguaje natural

Desde que aparecieron los primeros buscadores en los años 90, se ha «educado» al usuario de forma que consulte sobre las palabras clave más importantes, ignorando preposiciones y otras partículas sin relevancia y refinando los términos de la consulta a cada paso hasta dar con los resultados esperados. Este es el modelo que tenemos vigente en los buscadores comerciales actuales y al que los usuarios están bastante bien adaptados.

Recientemente han surgido nuevos proyectos con el lenguaje natural como herramienta de búsqueda. ¿Por qué expresar una necesidad de información «hablando como los indios» cuando podríamos hacerlo en nuestra propia lengua? Este es el objetivo de compañías como *Powerset*, *Lexxe*, *TrueKnowledge* o *Ask.com*. Lejos de ser proyectos lo suficientemente maduros para su uso cotidiano, sí consiguen reflejar el valor añadido respecto a las búsquedas tradicionales. El uso intensivo de técnicas de procesamiento del lenguaje natural, sobre todo en lo relativo a la semántica de las consultas y contenidos, hacen que a este tipo de búsquedas también se les refiera como «*búsquedas semánticas*» o «*búsquedas en lenguaje natural*».

De las tres acepciones ya presentadas para la «*búsqueda semántica*», es quizás esta última la que más dificultades está planteando a la comunidad científica. La mayoría de estos sistemas persiguen una indexación con información semántica de toda la

web. La infraestructura y los recursos necesarios para conseguir estos objetivos tan ambiciosos plantean unos costes que no terminan de ser del todo asumibles. Sólo hay que realizar un cálculo mental para imaginar la cantidad de recursos lingüísticos que harían falta en cada idioma, y que además permitan obtener una estructura semántica de cada frase de cada documento existente en la web. Por otro lado, los usuarios actuales se encuentran lo suficientemente *entrenados* en el manejo de los buscadores tradicionales que es difícil plantearles la necesidad de dar el salto a este tipo de tecnologías. No obstante, las grandes compañías siguen viendo en la búsqueda en lenguaje natural un filón a explotar. Prueba de ello es la reciente adquisición de Powerset por Microsoft.[3]

Para concluir este apartado sobre búsqueda en lenguaje natural veamos un ejemplo de la tecnología desarrollada por la empresa Powerset. Por el momento, Powerset ha creado toda una capa semántica encima de la Wikipedia —de momento sólo en su versión inglesa—. Este buscador permite realizar preguntas en lenguaje natural usando las típicas cláusulas *"Wh"*: What, Which, When, Where, Who, Why, etc... En la figura 3.5 se muestra el resultado de la consulta «*Who was the first man on the moon?*».[4] Puede observarse como Powerset también hace uso de la web semántica, en concreto de *Freebase*, para mejorar los contenidos devueltos por la Wikipedia sobre el célebre astronauta *Neil Amstrong*.

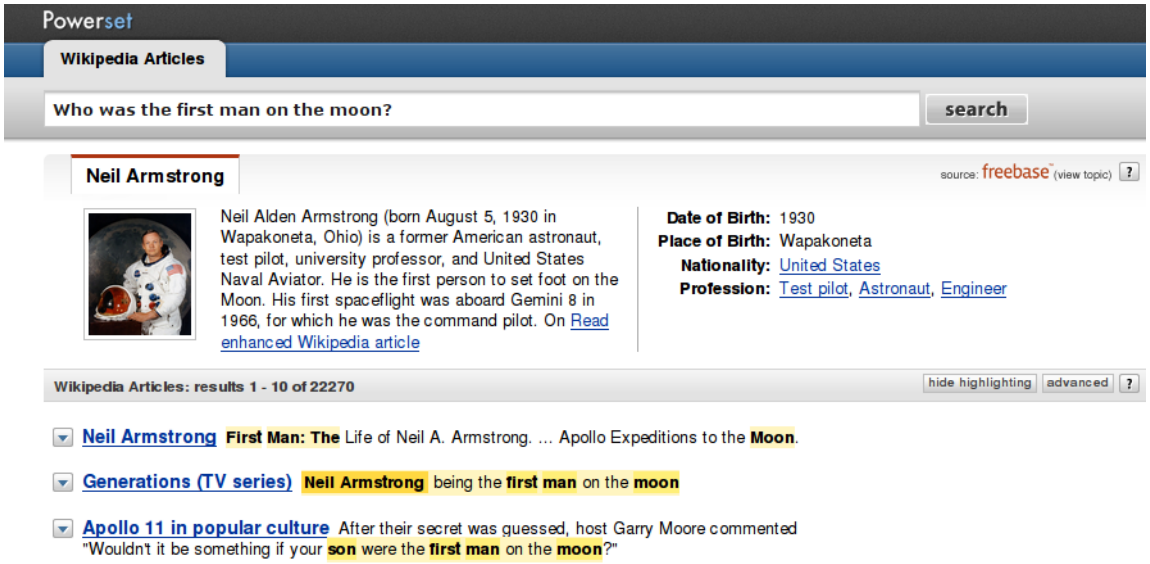

Figura 3.5: Búsqueda en lenguaje natural mediante Powerset

[3] Aunque la cifra oficial no ha trascendido se estima la operación en casi 100 millones de dólares.
[4] En castellano: «¿Quién fue el primer hombre en la luna?».

Tras el estudio de los distintos significados que se le pueden atribuir al término «*búsqueda semántica*», es interesante realizar una serie de conclusiones en cuanto a la relación entre los mismos:

- Los tres problemas tienen un nexo común: necesitan de conocimiento semántico para mejorar la calidad de los buscadores actuales.
- La comunidad científica sigue desarrollando tanto las técnicas como las propias herramientas para el tratamiento de ontologías y otras estructuras semánticas.
- Todas las grandes corporaciones que directa o indirectamente se dedican a la recuperación de información, han puesto sus miras en la semántica como el siguiente escalón tecnológico a alcanzar.

3.3. Principales líneas de investigación

Una vez definido y acotado el campo de estudio dentro de la Recuperación de Información y la Búsqueda Semántica en las cuales se centra esta memoria, dedicaremos el presente apartado a enunciar las principales líneas de investigación que se están desarrollando en la comunidad científica: estudio de ontologías, aprendizaje automático, obtención automática de corpus, cross-lingual, publicidad personalizada y redes sociales.

3.3.1. Ontologías

Las ontologías son una herramienta extraordinaria en el mundo informático para la representación de conocimiento sobre un dominio determinado. Modelar mediante ontologías significa definir una serie de entidades con atributos y el conjunto de relaciones entre dichas entidades. Dada una ontología podemos realizar operaciones de clasificación o de razonamiento inductivo. La idea subyacente en el uso de ontologías en la web es obtener un mejor modelo de la semántica de los contenidos publicados para así mejorar los resultados de búsqueda. A continuación se exponen algunos de los enfoques más relevantes.

Interpretación de palabras clave basada en ontologías

Ya hemos mencionado anteriormente la existencia de una gran cantidad de usuarios perfectamente adaptados a la búsqueda tradicional, es decir, la redacción de consultas mediante palabras clave o *keywords*, ignorando partículas importantes desde el punto de vista sintáctico como preposiciones o determinantes pero carentes de

significado desde el punto de vista semántico. Por tanto, la idea de realizar una interpretación ontológica de las palabras clave es un primer paso lógico de investigación como así lo expresan en [TRA07].

Una de las primeras conclusiones es que, aunque parezca lo contrario, es más fácil generar una consulta sobre una ontología dando por supuesto cierto entendimiento del lenguaje natural, que desde palabras clave. Para este propósito es necesario realizar un estudio de cómo suelen los usuarios redactar sus necesidades de información mediante palabras clave, para finalmente, realizar un *mapping* a un lenguaje de consulta tipo SPARQL.

Otras de las conclusiones a la que llegan estos estudios es que la naturaleza estadística de las soluciones basadas en palabras clave y la madurez de las técnicas desarrolladas en este ámbito se adaptan mejor a las dimensiones de la web. Parece ser que el uso de ontologías en tales órdenes de magnitud adolece de problemas de rendimiento que no son factibles a nivel de explotación comercial. Éste es quizás unos de los problemas por los que la web semántica no termina de despegar, y por tanto, una interesantísima línea de investigación.

Generación automática de ontologías

Uno de los grandes problemas a los que se enfrentan los investigadores de la web semántica es la falta de recursos ontológicos sobre dominios concretos. Proyectos como «*Linking Open Data*» (pág. 35) vienen a paliar en parte estas necesidades, pero aún hay un largo camino por recorrer para conseguir obtener recursos con la misma facilidad y completitud con que se obtienen en otras disciplinas. Por ejemplo, para obtener un corpus de entrenamiento sobre artículos deportivos es suficiente con descargar unas cuantas páginas de la prensa especializada. En este sentido, existen estudios como [BUI07] o [CIM07] que intentan generar ontologías partiendo de los propios corpus de entrenamiento de las herramientas de procesamiento del lenguaje natural. El objetivo que se persigue es realizar un intenso análisis y procesamiento de los textos hasta generar las entidades y relaciones que den forma a una ontología.

Desambiguación mediante ontologías

La ambigüedad es una característica inherente a lenguaje natural. Tal y como vimos en la introducción sobre los lenguajes naturales (apartado 1.1.2), la ambigüedad es un problema complejo y requiere de distintas técnicas para resolverla. En el ámbito del procesamiento del lenguaje natural ya existen diversos algoritmos que resuelven la tarea con éxito, especialmente en los niveles léxico y sintáctico. Sin embargo, en el nivel semántico aún existen problemas de desambiguación. Para este propósito se están investigando algoritmos que mezclan la estadística con el conocimiento ontológi-

co y que están dando muy buenos resultados (ver [KAR07]).

3.3.2. Aprendizaje automático

Podemos encontrar numerosas líneas de investigación abiertas en lo concerniente a «*Machine Learning*» o aprendizaje automático. La aportación de esta disciplina dentro de la web sirve, sobre todo, para mejorar la calidad de las distintas herramientas implicadas en la web semántica: indexadores, clasificadores, *crawlers*, *taggers*, filtros anti-spam, etc... Algunos de los trabajos pioneros en la utilización de aprendizaje automático para la web semántica van en la línea de [CIM05] y [CIM04]. Seguidamente veremos algunos de los estudios que relacionan el aprendizaje automático con la búsqueda semántica.

Enriquecimiento automático de la metainformación

Como ya vimos en el apartado 3.1, una de las carencias de la *web tradicional* o Web 1.0 es la falta de metainformación en los contenidos. A pesar de los esfuerzos realizados para paliar este problema, y muy especialmente por parte de las aplicaciones de la Web 2.0, la cantidad de información semántica asociada a los contenidos publicados sigue siendo insuficiente. Es por ello que la búsqueda de fórmulas para el enriquecimiento de dicha metainformación es una línea de investigación muy activa.

El principal problema radica en lo costoso y lento que está resultando alcanzar los objetivos propuestos por la web semántica, sobre todo porque no todos los usuarios tienen suficiente interés en etiquetar sus contenidos o simplemente porque no están concienciados de la importancia de dicha labor y las repercusiones positivas a largo plazo. Por consiguiente, se están desarrollando técnicas de aprendizaje para el enriquecimiento de la metainformación de los contenidos de forma automática. Existen proyectos que, mezclando técnicas de procesamiento de lenguaje natural y *machine learning*, están obteniendo nuevos enlaces y datos que posteriormente se añaden a los contenidos. Uno de los ejemplos más ilustrativos es el enriquecimiento de la DBPedia utilizando un etiquetador entrenado sobre la Wikipedia. Los detalles de este trabajo están disponibles en [MIK08b].

Búsqueda contextualizada

Cuando abrimos la página de algún buscador y escribimos nuestra consulta, el sistema no tiene ninguna noción de contexto sobre la cual apoyarse para refinar los resultados. Veamos un ejemplo; si estamos visitando una página de turismo y escribimos «`jerez`» en el buscador, muy probablemente estamos solicitando información sobre la localidad gaditana. Sin embargo, si nos encontramos navegando en páginas

sobre vinos, al escribir de nuevo «`jerez`» seguramente es porque queremos información sobre el famoso *Vino Sherry* o *Vino de Jerez*. Por último, si tecleamos «`jerez`» mientras vemos páginas sobre motocicletas de gran cilindrada es porque estamos interesados en conocer las últimas novedades del G.P. de Motociclismo.

Partiendo del ejemplo anterior es evidente que conocer la información contextual de navegación puede convertirse en una pieza muy importante para mejorar las métricas de *precision* y *recall*. Las técnicas que suelen emplearse para este propósito se centran en realizar una expansión de la consulta con dicha información de contexto. Para procesar esta tupla $search = \{query, context\}$ es necesario también indexar con información ontológica. Finalmente, se realizan procesos de aprendizaje para conseguir asociar los contextos de entrada con los contextos indexados (ver [UKK08]).

3.3.3. Obtención automática de corpus

En los últimos años se ha incrementado notablemente el uso de los contenidos web en lenguaje natural como corpus de entrenamiento, muy especialmente para las investigaciones en el campo de la recuperación de información. En concreto, la Wikipedia ha sido una de las fuentes más utilizadas gracias al gran valor que aportan sus contenidos «democráticos» en la realización de las pruebas. Sin embargo, los recursos necesarios para indexar corpus de tal envergadura no están al alcance de todos. Distintos grupos de investigación se afanan en la tarea de generar herramientas que permitan generar corpus de entrenamiento de gran tamaño como es el caso de la *"Semantically Annotated Wikipedia"* de *Yahoo! Research*; en [ATS08] los autores presentan su trabajo sobre cómo generar un corpus anotado para entrenamiento utilizando software libre. En concreto, el corpus obtenido sobre aproximadamente millón y medio de entradas de la Wikipedia inglesa contiene la siguiente información:

- Separación por frases y *tokens*.
- Etiquetado morfosintáctico (part-of-speech), de entidades y semántico.
- Dependencias en análisis sintáctico.
- Lista de entidades y relaciones.

3.3.4. Multilingüismo

El interés por aplicaciones multilingües viene de varios años atrás con el estudio de la traducción automática (ver pág. 6). En continentes como el europeo conviven lenguas tan importantes como el inglés, el castellano, el alemán, el francés o el italiano. Las actividades económicas entre los países que manejan dichas lenguas repercute

positivamente en la inversión para la investigación de tecnología multilingüe.

Existe una línea de investigación llamada «*Cross-Language Information Retrieval*» cuyo objetivo consiste en realizar consultas a un sistema de recuperación de información en un idioma determinado, el cual ha indexado los contenidos web en distintos idiomas. Sería como preguntar a un buscador en castellano por «*estadística*» y obtener no sólo los documentos que contengan dicho término sino también aquellos en los que aparezcan «*statistics*», «*statistik*», «*statistique*» o «*statistica*». Adicionalmente, los documentos recuperados podrían devolverse ya traducidos a la lengua de la consulta.

Aunque los congresos y *workshops* para la publicación de artículos relacionados con la recuperación de información los enunciaremos en detalle en el apartado 3.5.1, en el apartado del multilingüismo toma especial interés el *CLEF Workshop, «Cross-Language Evaluation Forum»*. Es un foro subvencionado por la Comunidad Europea para la investigación del acceso a la información multilingüe. Algunas de las tareas consisten en mejoras de la recuperación de información multilingüe mediante análisis semántico como en [SOR08a]. O en la búsqueda de nuevos enlaces entre contenidos en distintos idiomas como en [SOR08b].

3.3.5. Publicidad personalizada

La publicidad es una estrategia encaminada a incrementar las ventas de un producto. Normalmente consiste en un anuncio informativo en el cual se expresan las supuestas bondades de dicho producto. Una buena campaña publicitaria suele basarse en dos factores: la forma de transmitir el mensaje es buena y hace mella en el público al que va dirigido y el canal de difusión ha sido el adecuado. En este último aspecto es donde internet ha dado todo un giro en la forma de trabajar de las empresas publicitarias.

Con la llegada de la Web 2.0 y sus aplicaciones, especialmente aquellas que recogen como información el perfil de los usuarios, la agencias se rifan el acceso a tal información ya que les permitiría generar una publicidad personalizada de gran impacto y penetración. Hay que tener en cuenta que los beneficios de estas empresas se basan en saber acertar en a quién dirigen sus campañas publicitarias. No es de extrañar que se estén invirtiendo importantes sumas de dinero en desarrollar una tecnología para la publicidad personalizada en la web.

Hace ya mucho tiempo que los principales buscadores incorporan un apartado de «*Enlaces Patrocinados*» donde muestran a los anunciantes relacionados con los términos de la consulta. Otro de los medios de publicidad en la web consiste en la selección automática de los anuncios en función de los contenidos. Algunos trabajos como en [CIA08] investigan esta línea en base a la relevancia de la página y el análisis semántico de los contenidos.

3.3.6. Redes sociales

Las redes sociales como Tuenti o Facebook entre otras, se está erigiendo como abanderadas de cómo deben desarrollarse aplicaciones en la web 2.0 y con claras miras hacía conseguir una web semántica. En el caso concreto de las redes sociales, son los propios usuarios quienes generan una enorme cantidad de información de gran valor para terceros. Acabamos de ver en el apartado anterior como podrían utilizarse con fines publicitarios. En el apartado 3.2.1 vimos incluso como se enriquecería la búsqueda de «pedir una pizza» con información de nuestros contactos.

Por tanto, no es de extrañar que en actualmente se estén incrementando los estudios en el campo de la redes sociales. Principalmente en hay dos variantes de investigación: el uso de la información semántica contenida en las redes sociales y la búsqueda de contenidos dentro de la propia red. En la primera ya hemos visto la cantidad de escenarios en los cuales se le podría sacar provecho a dicha información. Respecto a la segunda podemos decir que hay una pequeña divergencia entre cómo los buscadores comerciales buscan información dentro de la red social y cómo lo hacen los propios buscadores de la red. Trabajos como [MIK08a] o [AGI07] estudian este último problema.

3.4. Otras líneas de investigación

En este apartado nos centraremos en algunas líneas de investigación relacionadas con la búsqueda semántica pero que no tienen la trascendencia y el impacto de las anteriores. Son el caso de los portales de dominio específico o restringido, los manuales de ayuda, la recuperación de información musical y la optimización.

3.4.1. Portales de dominio específico

Entendemos por portal de dominio específico aquel cuyos contenidos se basan en una temática concreta; un portal de venta de vehículos, un portal para los fans de «*El Señor de los Anillos*» o un portal dedicado a la salud. En este último caso se han elaborado proyectos en colaboración con las administraciones públicas para mejorar el acceso de los ciudadanos a la información. En trabajos como [SUO07] se propone un portal con características Web 2.0 que ayuden a la inserción de metainformación ontológica por parte de las autoridades sanitarias, de forma que se favorezca el posterior uso de la búsqueda semántica por parte de los ciudadanos.

En [BUI06] se realiza un curioso trabajo de aplicación directa de las técnicas ontológicas sobre un portal futbolístico. La idea original es extraer el conocimiento semántico del portal para crear una nueva web que respete los contenidos originales pero mejore la navegación y la experiencia de usuario. Todo ello partiendo de la información ontológica original sin alterarla. El objetivo es demostrar, por un lado, el desaprovechamiento actual de la metainformación, y por otro, motivar a los desarrolladores a utilizar tales recursos semánticos para crear nuevas interfaces de usuario.

3.4.2. Manuales de ayuda

Todos estamos acostumbrados a trabajar con manuales de ayuda cuando utilizamos un nuevo software. La forma de consultar dichos manuales suele ser mediante búsquedas por palabras clave. Resulta extraño ver como la búsqueda en Internet ha mejorado sustancialmente en los últimos años gracias a las técnicas de recuperación de información y sin embargo no termina de implantarse para la búsqueda de contenidos dentro de un manual de ayuda. Al final, los usuarios no contentos con los resultados terminan por no utilizar el manual y recurriendo a Internet. En este sentido, existen trabajos como [UZZ07] que investigan la posibilidad de mezclar técnicas de recuperación de información y búsqueda contextualizada.

3.4.3. Recuperación de información musical

La recuperación de información musical o MIR (del inglés "Music Information Retrieval") es análoga a la recuperación de información vista hasta ahora, con la salvedad de que en lugar de tener un corpus de documentos de texto trabajamos con archivos de audio musicales. Algunos de los usos prácticos de esta disciplina son:

- Clasificación automática por características musicales.
- Reconocimiento e identificación musical.
- Selección musical contextual y por estados de ánimo.

Lo interesante de esta variante musical de la recuperación de información es que al guardar numerosas similitudes con la recuperación de documentos, muchas de las técnicas desarrolladas en para una pueden aplicarse en la otra y viceversa. Además de lo atractivo de contar con un corpus del gusto de los investigadores se le suma cierta facilidad a la hora de evaluar el rendimiento de los algoritmos. Un caso práctico de indexación semántica musical puede estudiarse en [SEI06].

3.4.4. Optimización

Todas las líneas de investigación analizadas hasta el momento se centran en la mejora de los resultados obtenidos por los sistemas de recuperación de información. No obstante, la ejecución de todas estas nuevas técnicas tiene un coste temporal asociado a una consulta y que por tanto hay que evaluar. Actualmente, los buscadores comerciales trabajan prácticamente en tiempo real, así pues los nuevos algoritmos deben desarrollarse con la premisa de que los usuarios muy probablemente no estén dispuestos a sacrificar los tiempos de respuesta a favor de un mejor refinamiento de los resultados. En esta línea podemos encontrar trabajos como [MON07], donde analizan las desventajas temporales de los sistemas de búsqueda semántica y proponen soluciones para el trabajo óptimo con ontologías.

3.5. Contexto investigador

El objetivo de este apartado es agrupar toda la información relativa a congresos, *workshops*, revistas, universidades, centros de investigación, empresas privadas e investigadores relevantes que de un modo u otro están relacionados con la búsqueda semántica o la recuperación de información. Así pues, esta información puede utilizarse para profundizar en los contenidos de esta memoria o para iniciarse en la investigación sobre dichos temas.

3.5.1. Congresos y talleres

Se enuncian a continuación los congresos y *workshops* de mayor interés para la investigación en el procesamiento del lenguaje natural orientado a la indexación semántica y la recuperación de información en la web.

Congresos

En el caso de los congresos podemos resaltar los siguientes: TREC y CLEF son los más importantes en cuanto a recuperación de la información. El primero es organizado por el gobierno norteamericano y por tanto se centra en el inglés como lengua. El CLEF está patrocinado por la Unión Europea por lo que hace hincapié en la investigación de la recuperación de información multilingüe. Adicionalmente podemos destacar: el CICLing, que se especializa en el procesamiento del lenguaje natural, ESWC y ISWC que son los congresos más importantes sobre web semántica o el WSDM que es el congreso del ACM sobre búsqueda web.

- **CICLing:** «Conference on Intelligent Text Processing and Computational Linguistics»
  http://www.cicling.org
- **CLEF:** «Cross Language Evaluation Forum»
  http://www.clef-campaign.org
- **EACL:** «The European Chapter of the ACL»
  http://www.eacl2009.gr
- **ESWC:** «European Semantic Web Conference»
  http://www.eswc2009.org
- **ICAIL:** «International Conference on Artificial Intelligence and Law»
  http://idt.uab.cat/icail2009
- **ICDAR:** «International Conference on Document Analysis and Recognition»
  http://www.cvc.uab.es/icdar2009
- **ICSC:** «IEEE International Conference on Semantic Computing»
  http://icsc.icsi.berkeley.edu/icsc
- **ICWSM:** «International AAAI Conference on Weblogs and Social Media»
  http://www.icwsm.org
- **ISMIR:** «The International Society for Music Information Retrieval»
  http://www.ismir.net
- **ISWC:** «International Semantic Web Conference»
  http://iswc2009.semanticweb.org
- **LREC:** «International Conference on Language Resources and Evaluation»
  http://www.lrec-conf.org
- **NAACL-HLT:** «North American Chapter of the Association for Computational Linguistics - Human Language Technologies»
  http://www.naaclhlt2009.org
- **NODALIDA:** «Nordic Conference on Computational Linguistics»
  http://beta.visl.sdu.dk/nodalida2009
- **PALC:** «Practical Applications in Language and Computers»
  http://palc.ia.uni.lodz.pl
- **RANLP:** «International Conference on Recent Advances in Natural Language Processing»
  http://lml.bas.bg/ranlp2009

- **SEPLN:** «Sociendad Española para el Procesamiento del Lenguage Natural»
  http://www.sepln.org
- **TAC:** «Text Analysis Conference»
  URL: http://www.nist.gov/tac
- **TALN:** «Traitement Automatique des Langues Naturelles»
  http://www-lipn.univ-paris13.fr/taln09
- **TREC:** «Text REtrieval Conference»
  http://trec.nist.gov
- **WSDM:** «ACM International Conference on Web Search and Data Mining»
  http://www.wsdm2009.org

Workshops

Uno de los *workshops* a destacar es el ESAIR, organizado por Yahoo! Research. Otros de gran importancia son el IWCS o el SemSearch.

- **AND:** «Workshop on Analytics for Noisy Unstructured Text Data»
  http://and2009workshop.googlepages.com
- **ESAIR:** «Exploiting Semantic Annotations in Information Retrieval»
  http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=esair_2009
- **IWCS:** «International Workshop on Computational Semantics»
  http://iwcs.uvt.nl
- **NTCIR:** «NII Test Collection for IR Systems»
  http://research.nii.ac.jp/ntcir
- **OntoLex:** «The OntoLex Workshop»
  URL: http://www.loa-cnr.it/ontolex08
- **SDoW:** «Social Data on the Web»
  http://sdow2008.semanticweb.org
- **SemSearch:** «Semantic Search Workshop»
  http://km.aifb.uni-karlsruhe.de/ws/semsearch09
- **WAC:** «Web As Corpus»
  http://www.sigwac.org.uk/wiki/WAC5
- **WeST:** «Web & Semantic Technology»
  http://airccse.org/west/west2009.html

3.5.2. Revistas

En este apartado se enuncian algunas de las publicaciones más relevantes en los campos del procesamiento del lenguaje natural y la recuperación de información. Destacar especialmente las de «*ACM Transactions on Information Systems*», «*Computational Linguistics*» e «*International Journal on Document Analysis and Recognition*».

- «*ACM Transactions on Information Systems*»
  ISSN: 1046-8188
  http://tois.acm.org
- «*Artificial Intelligence Research*»
  ISSN: 11076-9757
  http://www.jair.org
- «*Computational Linguistics*»
  ISSN: 0891-2017
  E-ISSN: 1530-9312
  http://www.mitpressjournals.org/loi/coli
- «*Information Processing & Management*»
  ISSN: 0306-4573
  http://www.elsevier.com/wps/find/journaldescription.cws_home/244/description
- «*Information Retrieval*»
  ISSN: 1386-4564
  E-ISSN: 1573-7659
  http://www.springer.com/computer/database+management+%26+information+retrieval/journal/10791
- «*Intelligent Information Systems*»
  ISSN: 0925-9902
  E-ISSN: 1573-7675
  http://www.springer.com/computer/security+and+cryptology/journal/10844
- «*International Journal on Document Analysis and Recognition*»
  ISSN: 1433-2833
  E-ISSN: 1433-2825
  http://www.springerlink.com/content/1433-2825
- «*Natural Language Engineering*»
  ISSN: 1351-3249
  E-ISSN: 1469-8110
  http://journals.cambridge.org/action/displayJournal?jid=NLE

- «*Research on Language and Computation*»
  ISSN: 1570-7075
  E-ISSN: 1572-8706
  http://www.springer.com/linguistics/computational+linguistics/journal/11168

- «*Web Semantics*»
  ISSN: 1570-8268
  http://www.elsevier.com/wps/find/journaldescription.cws_home/671322/description

3.5.3. Universidades

En este apartado se pretende agrupar a las universidades y grupos de investigación asociados que más destacan en publicaciones sobre recuperación de información, procesmiento del lenguaje natural y búsqueda semántica. Especial mención merecen los grupos de las Universidades de Edimburgo, Karlsruhe, Cambridge y Stanford.

- *Computational Linguistics & Phonetics*
  Universität Des Saarlandes
  http://www.coli.uni-saarland.de

- *Computational Linguistics Laboratory*
  Nara Institute of Science and Technology
  http://cl.naist.jp

- *Human Communication Research Centre*
  The University of Edimburgh
  http://www.hcrc.ed.ac.uk

- *Institute of Applied Informatics and Formal Description Methods (AIFB)*
  Univeristy of Karlsruhe
  http://www.aifb.uni-karlsruhe.de

- *Institute for Natural Language Processing*
  Univeristy of Stuttgart
  http://www.ims.uni-stuttgart.de

- *Knowledge Engineering Group*
  Tsinghua University
  http://keg.cs.tsinghua.edu.cn

- *LINC Lab - Language and Information in Computation*
  Penn - University of Pennsylvania
  http://www.cis.upenn.edu/~linc

- *Linguistic Computing Laboratory*
  Univeristy of Roma «La Sapienza»
  http://lcl2.uniroma1.it
- *Natural Language and Information Processing Group*
  Univeristy of Cambridge
  http://www.cl.cam.ac.uk/research/nl
- *Semantic Computing Research Group*
  Helsinki University of Technology
  http://www.seco.tkk.fi
- *Semantic Grid Research Group*
  University of Southampton
  http://www.semanticgrid.org/OGF
- *The Semantic Web Research Group*
  Univeristy of Maryland
  http://www.mindswap.org
- *The Stanford Natural Language Processing Group*
  Stanford Univeristy
  http://nlp.stanford.edu
- *Tsujii Laboratory*
  Univeristy of Tokyo
  http://www-tsujii.is.s.u-tokyo.ac.jp

3.5.4. Empresas y centros de investigación

Existen numerosas empresas y centros de investigación que se dedican al desarrollo de nuevas técnicas para el procesamiento del lenguaje natural y la recuperación de información. Casi resulta obvio que las empresas más importantes en este ámbito son las que disponen de un buscador comercial on-line; son los caso de Microsoft, Google y Yahoo!. Especialmente esta última compañía es la que más publicaciones sobre búsqueda semantica presenta en congresos internacionales.

- DFKI - German Research Center for Artificial Intelligence
  http://www.dfki.de
- EML Research - European Media Laboratory
  http://www.eml-r.org
- Google
  http://research.google.com

- Hewlett-Packard
  http://www.hpl.hp.com
- IBM
  http://www.research.ibm.com
- KMi - Knowledge Media Institute
  http://kmi.open.ac.uk
- Microsoft
  http://research.microsoft.com
- PARC - Palo Alto Research Center
  http://www.parc.com
- Yahoo!
  http://research.yahoo.com

3.5.5. Investigadores relevantes

Por último queremos destacar a una serie de investigadores que han publicado recientemente trabajos relevantes sobre búsqueda semántica. Siendo coherentes con el apartado anterior, algunos de los más importantes pertenecen a Yahoo! Research como Ricardo Baeza-Yates, Peter Mika o Hugo Zaragoza. Otros investigadores relevantes son Paul Buitelaar y Philip Cimiano.

- Ricardo Baeza-Yates
  Yahoo! Research
  http://research.yahoo.com/bouncer_user/70
- Paul Buitelaar
  DFKI, Saarbrücken
  http://www.dfki.de/~paulb
- Philip Cimiano
  Universität Karlsruhe
  http://www.aifb.uni-karlsruhe.de/Personen/viewPersonenglish?id_db=98
- Sara Cohen
  The Hebrew University of Jerusalem
  http://www.cs.huji.ac.il/~sara
- Jennifer Golbeck
  Universidad de Maryland
  http://www.cs.umd.edu/~golbeck

- Daniel Jurafsky
  Stanford University
  http://www.stanford.edu/~jurafsky
- Anastasia Karanastasi
  Technical University of Crete
  http://www.ced.tuc.gr/Staff/KaranastasiAnastasia.htm
- Juanzi Li
  Tsinghua University http://keg.cs.tsinghua.edu.cn/english.htm
- Christopher Manning
  Stanford University
  http://nlp.stanford.edu/~manning
- Peter Mika
  Yahoo! Research
  http://research.yahoo.com/bouncer_user/66
- Magnus Niemann
  Free University of Berlin
  http://www.ag-nbi.de/members
- Philip Sorg
  Universität Karlsruhe
  http://www.aifb.uni-karlsruhe.de/Personen/viewPersonenglish?id_db=70
- Osma Suominen
  Helsinki University of Technology
  http://www.seco.tkk.fi/u/oisuomin
- Antti Ukkonen
  Helsinki University of Technology
  http://www.cis.hut.fi/aukkonen
- Victoria Uren
  Knowledge Media Institue
  http://people.kmi.open.ac.uk/victoria
- Naushad UzZaman
  University of Rochester
  http://www.cs.rochester.edu/~naushad
- Hugo Zaragoza
  Yahoo! Research
  http://research.yahoo.com/bouncer_user/37

Capítulo 4

Conclusiones

A lo largo de esta memoria hemos profundizado en la «Recuperación de Información» como una de las tareas dentro del «Procesamiento del Lenguaje Natural» y su aplicación directa en la recuperación de documentos en la web. Se han definido las métricas que permiten la evaluación objetiva del rendimiento de los distitos motores de búsqueda además de exponer brevemente el funcionamiento de dichos sistemas, prestando especial atención a la fase de indexación y a los algoritmos de puntuación.

A continuación, se han enunciado una serie de hipótesis sobre cómo podrían mejorarse los sistemas de búsqueda si se aplicaran técnicas de procesamiento del lenguaje natural para extraer la semántica de los contenidos y no ceñirse únicamente a un análisis estádistico basado en la frecuencia y localización de los términos. Seguidamente se han expuesto una serie de objetivos sobre los cuales puede plantearse toda una labor de investigación sobre el estado del arte de la denominada «*búsqueda semántica*».

Finalmente se ha realizado un estudio sobre la evolución tecnológica que ha sufrido la web en los últimos años, centrándonos especialmente en la denominada web semántica y su influencia directa en los algoritmos de recuperación de información. En dicho estudio se ha llegado a la conclusión de que el término «*búsqueda semántica*» puede tener hasta tres acepciones en las cuales se utilizan técnicas de procesamiento del lenguaje natural para mejorar los resultados de los motores de búsqueda tradicionales:

- «Búsqueda en la web semántica»
- «Indexación y búsqueda con información semántica»
- «Búsqueda en lenguaje natural»

Adicionalmente se ha llevado a cabo una labor de recopilación de las líneas de investigación principales en las que trabaja la comunidad científica para mejorar los sistemas de recuperación de información en la web. Las estrategias de mayor

seguimiento son aquellas que emplean ontologías tanto para modelar el conocimiento semántico como para mejorar la eficacia de los motores de búsqueda. Cabe mencionar los estudios basados en aprendizaje automático para la creación y utilización de la metainformación de los contenidos. Además, se han enunciado otras vías de interés como la publicidad personalizada o la recuperación de información en las redes sociales. En este sentido hay que destacar la evolución de los sistemas de búsqueda para relacionar semánticamente elementos de distinta naturaleza tales como texto, fotos, video y audio e integrarlos en un mismo conjunto de resultados.

El estudio sobre «*búsqueda semántica*» concluye con un compendio de todo tipo de información útil relacionada con el contexto investigador: congresos, revistas, universidades, empresas privadas, investigadores relevantes, etc... con el doble objetivo de completar los contenidos de esta memoria y facilitar a cualquier persona interesada la iniciación en la materia.

Como posible trabajo futuro y de cara a desarrollar una tesis doctoral, cabe destacar que el abanico de posibilidades para nuevas líneas de investigación sobre la «*búsqueda semántica*» es aun muy amplio. En primer lugar hay mucho por descubrir en lo relativo a cómo relacionar semánticamente toda la información no estructurada existente en la web. Creemos que el uso de ontologías tiene mucho que decir en este sentido. En segundo lugar, hay bastante trabajo por desarrollar para conseguir una indexación semántica de los contenidos; en este aspecto es donde parece que el procesamiento del lenguaje puede jugar un papel decisivo. Por último, para la búsqueda en lenguaje natural hay que investigar no sólo en los algoritmos de PLN sino también en la manera de afrontar los recursos necesarios para la comprensión semántica de todos los documentos existentes en la red.

En definitiva, existe una interesante oportunidad para trabajar en la materia como así lo demuestran la cantidad de congresos de carácter internacional, además del claro interés de empresas multinacionales que dedican una gran parte de su presupuesto a la investigación y desarrollo de la «*búsqueda semántica*» como son los casos de Yahoo!, Google o Microsoft.

Bibliografía

[AGI07] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis and Gilad Mishne
Finding High-Quality Content in Social Media with an Application to Community-Based Question Answering
Yahoo! Research - Technical Report

[ALE05] H. Peter Alesso and Craig F. Smith
Developing Semantic Web Services
A K Peters, Ltd.
2005

[ALL95] James Allen
Natural Language Understanding
Benjamin Cummings
1995

[ATS08] Jordi Atserias, Hugo Zaragoza, Massimiliano Ciaramita and Giuseppe Attardi
Semantically Annotated Snapshot of the English Wikipedia
In proceedings of the 2008 LREC Conference, Marrakech, Marruecos.

[BAE99] Ricardo Baeza-Yates and Berthier Ribeiro-Neto
Modern Information Retrieval
ACM Press
1999

[BON08] Francesco Bonchi, Carlos Castillo, Debora Donato and Aristides Gionis
Topical Query Decomposition
In proceedings of the 2008 ACM KDD Conference, Las Vegas, Nevada, Estados Unidos.

[BRI98] Sergey Brin and Lawrence Page
The Anatomy of a Large-Scale Hypertextual Web Search Engine
Journal of Computer Networks and ISDN Systems, Vol. 30, Pages 107-117.

[BUI06] Paul Buitelaar, Thomas Eigner, Greg Gulrajani, Alexander Schutz, Melanie Siegel, Nicolas Weber, Philipp Cimiano, Günter Ladwig, Matthias Mantel and

Honggang Zhu
Generating and Visualizing a Soccer Knowledge Base
In proceedings of the EACL06 Demo Session, Trento, Italia.

[BUI07] Paul Buitelaar, Philipp Cimiano and Bernardo Magnini
Ontology Learning from Text: An Overview
In proceedings of the 2007 Conference on Traitement Automatique des Langues Naturelles, Tolouse, Francia.

[CIA08] Massimiliano Ciaramita, Vanessa Murdock and Vassilis Plachouras
Semantic associations for contextual advertising
Journal of Electronic Commerce Research, Vol. 9, Nº1.

[CIM04] Philipp Cimiano, Aleksander Pivk, Lars Schmidt-Thieme and Steffen Staab
Learning Taxonomic Relations from Heterogeneous Sources of Evidence
In proceedings of the ECAI-2004 Workshop on Ontology Learning and Population, Valencia, España.

[CIM05] Philipp Cimiano, Andreas Hotho and Steffen Staab
Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis
Journal of Artificial Intelligence Research, Vol. 24, Pages 305-339.

[CIM07] Philipp Cimiano, Peter Haase, Matthias Herold, Matthias Mantel and Paul Buitelaar
LexOnto: A Model for Ontology Lexicons for Ontology-based NLP
In proceedings of the 2007 OntoLex Workshop held in conjunction with ISWC'07, Busan, Corea del Sur.

[DAV06] John Davies, Rudi Studer and Paul Warren
Semantic Web Technologies: Trends and Research in Ontology-based Systems
John Wiley & Sons Ltd.
2006

[GOL08] Jennifer Golbeck, Peter Mika and Michael Uschold
Introduction to the Special Issue on the Semantic Web Challenge 2006 and 2007
Journal of Web Semantics, Vol. 6, Issue 4, Pages 241-242.

[GRO04] David A. Grossman and Ophir Frieder
Information Retrieval: Algorithms and Heuristics
Springer
2004

[JUR08] Daniel Jurafsky and James H. Martin.
Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics
Prentice-Hall
2008

[KAR07] Anastasia Karanastasi and Stavros Christodoulakis
Ontology-Driven Semantic Ranking for Natural Language Disambiguation in the OntoNL Framework
In proceedings of the ESWC'07 Conference, Innsbruck, Austria.

[LEI05] Jochen L. Leidner
A Wireless Natural Language Search Engine
In proceedings of the 2005 ACM SIGIR Conference, Salvador, Brasil.

[MAN99] Christopher D. Manning and Hinrigh Schütze.
Foundations of Statistical Natural Language Processing
MIT Press, Cambridge, Massachusets,
1999

[MAN08] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze
Introduction to Information Retrieval
Cambridge University Press
2008

[MIK08a] Peter Mika
Semantic Search and the Social Web
In proceedings of the SDoW 2008 Workshop at the 7th International Semantic Web Conference, Karlsruhe, Alemania.

[MIK08b] Peter Mika, Massimiliano Ciaramita, Hugo Zaragoza and Jordi Atserias
Learning to tag and tagging to learn: A case study on Wikipedia
IEEE Intelligent Systems, Vol. 23, Issue 5, Pages 26-33.

[MIK08c] Peter Mika
Microsearch: An Interface for Semantic Search
In proceedings of the SemSearch 2008 Workshop at ESWC Conference, Tenerife, Spain.

[MON07] Marius Monton, Jordi Carrabina, Carlos Montero, Javier Serrano, Xavier Binefa Ciro Gracia, Mercedes Blázquez, Jesús Contreras, Emma Teodoro, Núria Casellas, Joan-Josep Vallbé and Pompeu Casanovas
Accelerating Semantic Search with Application of Specific Platforms
In proceedings of the 2007 International Conference on Artificial Intelligence and Law, Stanford, California, Estados Unidos.

[NIE08] Magnus Niemann, Malgorzata Mochol and Robert Tolksdorf
Enhancing Hotel Search with Semantic Web Technologies
Journal of Theoretical and Applied Electronic Commerce Research, Vol. 3, Issue 2, Pages 82-96.

[PAG98] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd
The PageRank Citation Ranking: Bringing Order to the Web
Stanford Digital Library Technologies Project.

[SEI06] Frank Seifert
Comparison and Partial Ordering of Music by Applying a Generic Semantic Index
Journal of Systemics, Cybernetics and Informatics. Vol. 4, Issue 2, Pages 1-5.

[SYD09] Marcin Sydow, Francesco Bonchi, Carlos Castillo and Debora Donato
Optimising Topical Query Decomposition
In proceedings of the 2009 workshop on Web Search Click Data, New York, NY, Estados Unidos.

[SOR08a] Philipp Sorg and Philipp Cimiano
Cross-lingual Information Retrieval with Explicit Semantic Analysis
In proceedings of the CLEF 2008 Workshop, Aarhus, Dinamarca.

[SOR08b] Philipp Sorg and Philipp Cimiano
Enriching the crosslingual link structure of Wikipedia - A classification-based approach
In proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence, Chicago, Estados Unidos.

[SUO07] Osma Suominen, Kim Viljanen and Eero Hyvönen
Semantic Faceted Search in a Citizens' Health Portal
In proceedings of the ESWC'07 Conference, Innsbruck, Austria

[TRA07] Thanh Tran, Philipp Cimiano, Sebastian Rudolph and Rudi Studer
Ontology-based Interpretation of Keywords for Semantic Search
In proceedings of the ISWC'07 Conference, Busan, Corea del Sur.

[UKK08] Antti Ukkonen, Carlos Castillo, Débora Donato and Aristides Gionis
Searching the Wikipedia with Contextual Information
In proceedings of the 17th ACM conference on Information and knowledge mining, Napa Valley, California, Estados Unidos.

[URE08] Victoria Uren, Yuangui Lei and Enrico Motta
SemSearch: Refining Semantic Search
In proceedings of the SemSearch 2008 Workshop at ESWC Conference, Tenerife, Spain.

[UZZ07] Naushad UzZaman
Semantic Search on Help Files
Project for Problem Seminar (CSC 400), University of Rochester, 2007.

Índice de figuras

Índice de cuadros

Currículum Investigador

Se añade este anexo para recoger el currículum investigador del autor de esta memoria.

Situación Profesional

- Intelligent Dialogue Systems S.L. - INDISYS.
  Coordinador de Ingeniería.
- Grupo de Investigación Julietta
  Universidad de Sevilla.

Líneas de Investigación

- Procesamiento del lenguaje natural.
- Sistemas de diálogo multimodal y multilingüe.
- Síntesis y reconocimiento de voz.
- Recuperación de información.
- Indexación semántica.

Formación Académica

- Ingeniero en Informática.
  Escuela Técnica Superior de Ingeniería Informática de Sevilla.
  Febrero 2006.

Estancias en Centros de Investigación

- IBM - Centro Tecnológico de la Lengua. Sevilla.
  Etiquetadores morfosintácticos para sistemas de síntesis de voz.
  Aplicaciones `<VXML>`
  Octubre 2004 - Junio 2005.

▪ Participación en Proyectos de Investigación

- **Título: TALK**
  «*Talk and Look, Tools for Ambient Linguistic Knowledge*»
  - Grupo de Investigación Julietta, Universidad de Sevilla.
  - VI Programa Marco, UE.
  - Enero 2004 - Diciembre 2006.

- **Título: ATLANTIDA**
  «*Aplicación de tecnologías líder a aeronaves no tripuladas para la investigación y desarrollo en ATM*»
  - INDISYS
  - III Programa Cénit, Ministerio de Ciencia e Innovación.
  - Septiembre 2007 - Septiembre 2011.

▪ Publicaciones

- «*Una Comparación de ESBs desde la Perspectiva de la Integración de Aplicaciones*»
  Rafael Corchuelo, Rafael Z. Frantz y Jesús González.
  Proceedings of the 13th Conference on Software Engineering and Databases, JISBD 2008, pp. 403-408.
  ISBN: 978-84-612-5820-8

- «*Advances in a DSL for Application Integration*»
  Rafael Z. Frantz, Rafael Corchuelo y Jesús González.
  Proceedings of the 6th Workshop on Integrating Web Applications ZOCO, JISBD 2008. Vol. 2, No. 2, pp. 54-66.
  ISSN: 1988-3455

- «*Semi-Automated Testing of Real World Applications in Non-Menu-Based Dialogue Systems*»
  Pilar Manchón, Guillermo Pérez, Gabriel Amores y Jesús González.
  Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pp. 181-182. DECALOG 2007

- «*Adaptación del Método de Etiquetado No Supervisado TBL*»
  Jesús González Martí, David González Maline, y José Antonio Troyano Jiménez.
  Procesamiento del Lenguaje Natural Nº37, pp. 5-10. SEPLN'06
  ISSN: 1135-5948

- «*Integrating OWL Ontologies with a Dialogue Manager*»
  Gabriel Amores, Guillermo Pérez, Pilar Manchón, Fernando Gómez y Jesús González.
  Procesamiento del Lenguaje Natural Nº37, pp. 153-160 . SEPLN'06
  ISSN: 1135-5948