ROLANDO CLAUDIA
RATTI DANIELA
MARCONI LUCIA
CUTUGNO PAOLA
Istituto per i Circuiti Elettronici
Consiglio Nazionale delle Ricerche
Génova, Italia
E-Mail: {rolando, ratti, marconi, cutugno}@ice.ge.cnr.it

Estudio de la problemática relativa a la navegación en INTERNET y perspectivas para la creación de instrumentos que se puedan facilitar al usuario

Introducción

Imagínate encontrarte en el mercado, un mercado enorme, donde se vende de todo, la primera impresión que experimentas es de turbación y, si el tiempo a tu disposición es poco, te pones a pensar: ¿Por dónde empiezo? ¿Qué estoy buscando? ¿Qué necesito y qué quiero ver realmente?

Con una situación problémica, análoga al contenido de nuestro trabajo, queremos dar inicio a este artículo que trata sobre la posibilidad de crear un instrumento o soporte de distintos motores de búsqueda o más ambiciosamente, un motor de tipo distinto que pueda facilitar la búsqueda al usuario en WEB.

En los últimos años se ha asistido a un desarrollo siempre creciente de INTERNET sobre todo en lo que concierne a dos servicios que han llegado a ser esenciales en todo el mundo y de los que se habla cotidianamente: *E–mail y WEB*; mientras que otros servicios, a pesar de ser importantes, han quedado relegados a usos muy específicos (FTP, TELNET, etc.), ya que no han constituido un estímulo desde el punto de la investigación.

Las páginas HTML y de sitios WEB en estos años han tenido un crecimiento exponencial, ya sea de tipo cuantitativo, o ya sea cualitativo y, más en general, se ha verificado un aumento y un cambio de la tipología de los usuarios, se ha pasado de hecho, desde una de tipo especializado, ligado a la realidad de la investigación y de la industria, a una de masa que ha provocado el cambio, no solo de la cantidad, sino también de la modalidad misma de presentación de la información.

En el mercado el usuario tenía frente a sí una multitud de objetos y podía a través de la vista, del olfato, del tacto, del gusto elegir la mercancía que se le proponía, aquí, en la era de la información, el usuario se encuentra frente a un mercado donde sólo tiene una pantalla y un punto de partida por donde empezar su viaje en un mercado multilenguas, multidisciplinario, donde cada uno vende su producto, de cualquier género que se trate.

Los datos estadísticos revelan que hay tres millones de nombres de dominio, cuarenta y cinco millones de HOST, doscientos países conectados, cien millones de usuarios y que el 75 % del tráfico es generado por WEB; en Italia los datos observados en una investigación hecha por el Observatorio INTERNET Italia (Bocconi, junio 99) revelan que hay catorce millones de italianos que tienen en su casa una computadora, cuatros millones y medio un módem, un millón una suscripción a INTERNET.

Se calcula que actualmente el entero WEB está compuesto por 800 millones de páginas.

Para acceder a esta enorme cantidad de informaciones el recurso principal que el usuario tiene a su disposición es, sin dudas, el motor de búsqueda; de hecho el 80 % de los usuarios llega a los sitios a través de un motor de búsqueda.

Por otro lado, es opinión común que los actuales motores tendrían que mejorar el tipo de contestación con relación a los resultados, que, frecuentemente, suelen ser incongruentes y por demás genéricos.

La dificultad mayor para el usuario es hallar la dirección URL; se trata de resolver, o por lo menos facilitar al usuario, la recuperación cualitativa de la información, es decir crear instrumentos innovadores que filtren las informaciones y presenten solo las pertinentes a las preguntas hechas.

El usuario, en la actualidad, se encuentra frente al dilema de elegir el motor de búsqueda o un metamotor, buscar por medio de palabras claves y aún más, por categorías, con la sensación de haber perdido, a pesar de sus elecciones, algunas informaciones fundamentales para su investigación.

En realidad la calidad, la pertinencia de los resultados de cada investigación, depende de la organización de las informaciones a disposición en Internet.

Problemáticas en el análisis de los datos

Un análisis de las páginas HTML y de las distintas tipologías de los motores de búsqueda es esencial para procurar crear un instrumento que pueda ser de ayuda y facilitar la recuperación de las informaciones.

Páginas HTML

Las páginas en INTERNET son estructuradas de modos muy distintos, algunas son muy descriptivas, otras tienen muchas imágenes y poco texto, otras tienen LINKS a otras páginas, etc.

Es mucha la diferencia de las páginas científicas a las páginas comerciales, hay una estructura muy distinta ya que, creciendo la red en el tiempo, las páginas han sido creadas con HTML y generadores distintos y una total anarquía en su composición por la ausencia o la presencia de determinados TAGS.

Otro fenómeno que merece tenerse en consideración consiste en el hecho de que en muchas páginas la información es casi totalmente estática, es decir el contenido varía muy poco, mientras en otras la información varía cotidianamente y también más veces al día, por ejemplo las informaciones relativas a los diarios, o a bolsa, etc.

La inserción de páginas WEB en los motores de búsqueda

Las contestaciones al usuario dependen de la organización de los motores de búsqueda.

La metodología para la grabación de los datos en los archivos de los motores de búsqueda se hace en la actualidad según dos modalidades:

- Grabación manual de los varios sitios en «directory», estructuradas en categorías o temas por parte de un trabajador del equipo de redactores de la directory (en general se pone una sola referencia para cada sitio, de norma la «home page»).
- Grabación automática de páginas por medio de «search engine» que analizan todas las páginas de un sitio utilizando a la vez dos sistemas distintos: el sistema add–url o el sistema spider. El primero permite señalar cada página del sitio al motor de búsqueda dando la especifica de la dirección WEB, el segundo analiza el WEB buscando todas las páginas que encuentra y las graba en el archivo general. Este consigue visitar diariamente millones de sitios WEB poniendo al día las informaciones existentes e insertando las nuevas. Los dos tipos de grabación: manual y automático, a veces coexisten en el mismo motor pero no consiguen, a pesar de estar juntos, superar algunos límites de inserción y búsqueda:

➢ <u>en la estructura a directory:</u>

- El error de filtraje por parte del WEB–master que inserta el sitio en una determinada categoría
- La subdivisión en categorías a describir los sitios WEB existentes es insuficiente y no es homogénea entre los varios motores de búsqueda, es decir no existe un criterio unívoco de catalogación;
- Es difícil para un equipo, a pesar de su número de componentes, categorizar los datos en tiempos breves, considerando el crecimiento exponencial de los datos insertos en la red, y sobre todo, ponerlos al día;
- La caracterización por parte de los autores con meta–tag, keywords, etc. de las páginas WEB de forma que estas acaban por pertenecer a una determinada categoría en detrimento de otras.

➢ <u>spider</u>

- Excesivo ruido en los datos visualizados;
- Demasiado elevada la cantidad de páginas que tienen que ver.
- No podemos decir que las primeras páginas sean las más relevantes, su importancia puede estar determinada por varios factores, como por ejemplo LINKS a otros sitios, etc.
- Redundancia en la información con relación a la visualización de páginas del mismo sitio que se refieren al mismo argumento.

Es evidente que resulta difícil moverse con agilidad en este mar de informaciones tan diversamente estructuradas y que resulta arduo buscar estrategias válidas para la creación de tools innovadores.

Nueva metodología para búsqueda e indización

Probemos a analizar qué problemáticas se encuentran si un usuario escribe la palabra "pietra" utilizando un motor de búsqueda, por ejemplo Altavista, uno de los más conocidos.

El resultado de la búsqueda es de 50 895 páginas encontradas; resulta imposible consultarlas todas: se refieren a producción, elaboración, piedras preciosas, piedras para construcciones, hoteles, sitios turísticos, catálogos de productos de arte, complemento de mobiliario, películas, exposiciones, etc. y, en todas, obviamente aparece la palabra "pietra".

Conseguir extraer las informaciones útiles sólo se podría lograr por casualidad, pues está claro que se necesita mejorar la metodología de búsqueda.

Estos problemas se presentan obviamente en todas las palabras que tienen más de un significado y también en otras lenguas diferentes al italiano.

Si se analiza la definición de la palabra piedra que ofrece un diccionario general de la lengua italiana hallamos que esta tiene más significaciones :

Pietra: s.f.

pietra da affilare

pietra di paragone

pietra serena

pietra miliare

pietra preziosa

…..

Pietra Ligure

…..

Según nuestro parecer si conseguimos organizar los datos de un diccionario de forma que puedan representar una solución a las preguntas más pertinentes, esto podría ser el punto de partida para llegar a obtener respuestas más razonables.

Se podría pasar, por ejemplo, de una búsqueda inicial por palabras a una búsqueda sucesiva por lexías complejas: (pietra di paragone, pietra dello scandalo) o a palabras con relaciones semánticas (pietra preziosa: diamante, smeraldo, agata, etc.:) lo que significa una reorganización por campos del diccionario, que tenga en cuenta las relaciones paradigmáticas y sintagmáticas de las palabras, las relaciones de tipo sinonímica y las relaciones de tipo ideológico.

Es evidente que la nueva base de datos de un diccionario de referencia no es todavía suficiente para tratar todas las posibles elecciones que se encuentran en la red, seria necesario por lo tanto, enriquecerla y ponerla al día, constantemente y en particular:

- Averiguar si lemas/palabras son utilizadas en el ámbito del corpus de INTERNET o si son desusadas;
- Incrementar el diccionario con las nuevas palabras extraídas de la red y con los nuevos significados, e individualizar los usos específicos , sectoriales, etc.
- Insertar informaciones cuantitativas con relación al uso de los significados de cada palabra de forma que se pudieran presentar al usuario las usadas, en orden de prioridad.

Queda el problema de cómo presentar las posibles elecciones al usuario; de hecho, mientras algunos lemas tienen un número reducido de significados, otros tienen muchos u originan muchos caminos, como en el caso de piedra. Es necesario, por lo tanto, estudiar el problema tomando en consideración por ejemplo: la visualización de los datos en la pantalla, el tiempo de acceso y el tiempo de enlace.

En la Fig.1 se puede ver el esquema del camino que el usuario puede recorrer:

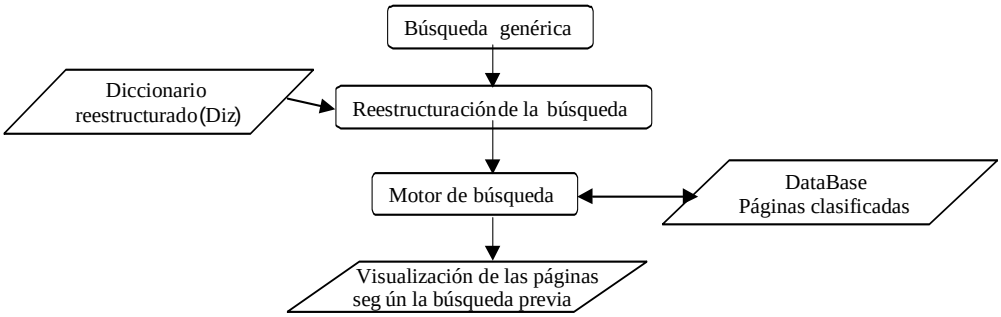

Fig. 1

Después de la organización semántico–ideológica del diccionario se puede pensar en la realización de un software que, analizada la página HTML, sea capaz de separar, junto al conocimiento de la base de datos, las palabras relevantes en el texto.

Se trataría de pasar de un formato HTML a un formato texto y hacer un análisis del texto de la página. La idea es desarrollar un paquete que pueda funcionar para la mayoría de las páginas y que, sobre todo, asigne pesos distintos a la palabras que forman el texto. El nuevo texto generado no tendría por lo tanto, una estructura donde todas las palabras tienen valor equivalente sino, al contrario, tendría palabras o conjuntos de palabras con más valor que otras y palabras con valor igual a cero. De tal forma se conseguiría evidenciar las palabras que caracterizan el texto, que hacen declarar que un texto habla de un tema y no de otro y este nuevo subconjunto representaría las nuevas palabras claves del documento.

En fig. 2 se puede ver esquematizada la metodología para clasificar la página y para insertar en el diccionario los nuevos datos.

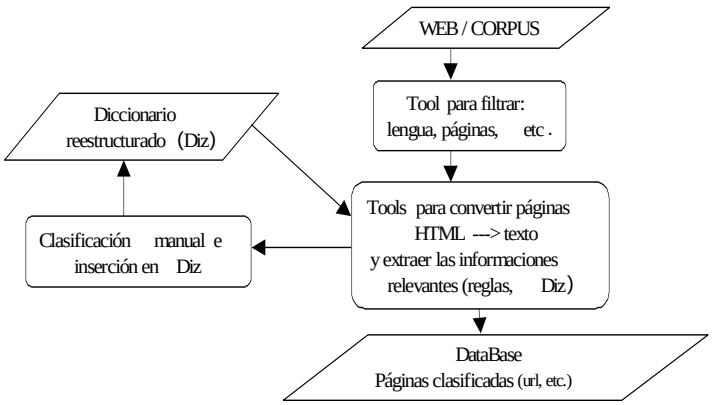

Fig. 2

Conclusiones

Nuestro proyecto de investigación no se propone desarrollar un motor novedoso de búsqueda, sino ofrecer al usuario, en algunos ámbitos, la posibilidad de hacer preguntas más pertinentes y enriquecer un motor de búsqueda con un nuevo tipo de indexación.

El desarrollo de las metodologías descritas, la reestructuración del diccionario, la creación del software para la extracción de las informaciones relevantes, el nuevo tipo de indización de los datos, tendrían que permitir una búsqueda inteligente que posibilite llegar a los resultados que el usuario desea, aunque queda abierta una serie de problemas tales como:

- Automatizar cuanto sea posible, procurando evitar siempre el trabajo manual;
- El software no funciona seguramente por todas las páginas HTML y, por lo tanto, se tendría que estudiar estrategias distintas;
- No es posible pensar que el software funcione en sitios que varíen cada día, o varias veces al día;
- Para los sitios o las páginas donde la información textual es escasa se tendrán que estudiar soluciones distintas;
- Hallar y tratar una gran cantidad de datos podría generar algunos problemas y por lo tanto llegar a ser necesaria toda una fase de verificación de los resultados;
- En algunos sectores específicos podría ser más oportuno organizar de otro modo la información;
- Cómo y cuándo prever la puesta al día de los datos.

El trabajo que nos espera es muy grande, aunque se trata de una sola lengua y a pesar de la gran cantidad de datos que ya tenemos y de nuestra experiencia en tratar datos lingüísticos, pensamos que merezca la pena concentrar nuestros esfuerzos en este sector, ya que la información cuando llega a ser excesiva y no manejable tiene un valor nulo y no es imaginable pensar en llegar al día de mañana con instrumentos que traten solo lenguas más usadas y conocidas.