Artículo de Investigación Revista Ciencia e Ingeniería. Vol. 32, No 3, pp. 141-152, agosto-noviembre, 2011. ISSN 1316-7081

Sistema semántico para la búsqueda inteligente de información por contexto para la Web

Semantic system for intelligent searching of information by context for the Web

Aguilar, José
CEMISID, Departamento de Computación, EISULA, Universidad de Los Andes
aguilar@ula.ve

Resumen

La Búsqueda por Contexto surge como la herramienta apropiada para explotar el conocimiento de la interacción web–usuario. Ella define modelos que describen los patrones de uso y caracterizan los perfiles de los usuarios de la web. Nosotros proponemos un Sistema Inteligente de Búsqueda por Contexto compuesto por una ontología para modelar el marco conceptual de búsqueda en la Internet, un mecanismo de razonamiento basado en reglas para usar la ontología e inferir patrones de comportamiento sobre Internet, y meta datos para almacenar instancias de información de la Internet. El Sistema de Búsqueda obtiene un conjunto de criterios del usuario cuando accede al sitio, y en base a esos criterios (p .e, preferencias y áreas de interés del usuario) y a la información solicitada por él en un momento dado, establece que será buscado, para tratar de hacer más eficaz ese proceso. Eso que se establece que será buscado es enviado a un motor de búsqueda (en este trabajo se usa a Google).

Palabras clave: Búsqueda por contexto, ontologías, semántica web.

Abstract

Search by Context emerges as an appropriate tool to exploit the knowledge of web-user interaction. It defines models which describe patterns of use and feature profiles of Web users. We propose an Intelligent Search System by context consists of an ontology for modeling the conceptual framework of Internet search, a reasoning mechanism based on rules to use the ontology and to infer patterns of behavior on Internet, and a metadata to store instances of information from Internet. Our search system obtains a set of criteria of the user when accessing the site, and based on them (eg, preferences and areas of interest of the user) and the information requested for him in a given moment, it will defined the information to be searched to try to make this process more efficient. That which is set to be searched is sent to a search engine (in this work is used Google).

Key words: Search by context, ontology, semantic web.

1 Introducción

El crecimiento de Internet ha hecho cada vez más necesario para los usuarios utilizar herramientas automáticas para encontrar, extraer, filtrar y evaluar los recursos de información disponibles. Algunos de los buscadores existentes, que tratan de encontrar información por categoría, o por contenido, son Altavista, Yahoo, Google, etc. (Baeza, 2005), (Jansena y col., 2006) (Li y col., 2004) (Rojas y col., 2010) (Rose y col., 2004), (Strasunskas y col., 2010). A estos buscadores se les introducen palabras claves, y ellos determinan las páginas o sitios web que contienen dichas palabras, tratando de satisfacer así los requerimientos del usuario. Muchas veces estas consultas traen resultados inconsistentes, o documentos que cumplen con el criterio de búsqueda pero no con el interés del usuario.

En los últimos años, algunos sistemas están realizando anotaciones de datos introducidas dentro del código HTML, siguiendo algún esquema de anotación común, normalmente basado en XML (Albrecht , 2001) para tratar de mejorar los procesos de búsqueda. A pesar de eso, los buscadores web al realizar una consulta siguen mostrando información que no es útil. Además, las anotaciones incorporadas en las páginas web no proporcionan igual información, debido a

que no existe un convenio/norma que nos diga qué contenido debemos añadir a las páginas web para su caracterización individual, que permita posteriormente ser usada en procesos de búsqueda. Por otro lado, los sistemas de búsqueda actuales no se diseñan para "comprender" la información que reside en la web.

En los últimos años han surgido una serie de técnicas que permiten el procesamiento avanzado de datos sobre la Internet, que se agrupan en el área llamada web semántica (Shadbolt y col., 2006), (Zhongyu, 2007). Ellas permiten realizar la búsqueda por contexto en la web, que consiste en un agregado semántico a la búsqueda para hacerla más eficaz, que caracteriza las preferencias, desempeños y áreas de conocimiento e interés del usuario que realiza la consulta. Las técnicas proveen formas avanzadas de indexación, patrones de reconocimiento de caminos de búsquedas, entre otras cosas. Algunos proyectos que han estado explotando esas técnicas son: el trabajo propuesto en (Aguilar, 2009) que realiza un proceso de "Minería Web" para extraer conocimiento derivado de la interacción web–usuario. En ese trabajo se emplearon las técnicas patrones secuenciales, análisis de caminos y cubos para construir un Sistema Hibrido de Minería Web, que permite obtener un conjunto de patrones de acceso de los usuarios sobre la web que sirven para descubrir correlaciones entre las páginas web y grupos de usuarios y comportamientos de los usuarios al navegar por el sitio. RODA (Red de Conocimiento Descentralizado a través de Anotaciones) es un proyecto cuyo objetivo es validar herramientas de anotación que faciliten la creación de repositorios con conocimientos claves, permitiendo probar la efectividad de las mismas en diferentes entornos y áreas de conocimiento (Baeza 2005), (Jansena, 2006). Seekport es un proyecto que posee la opción de búsqueda por temas (Baeza 2005), (Jansena, 2006). Los resultados que no se enclaven dentro del tema específico seleccionado aparecerán al final de los listados. El reconocimiento de temas de Seekport no se hace de forma manual, sino que se realiza sobre una base de reconocimiento automático de textos.

Algunos trabajos recientes en el área de búsqueda semántica basada en técnicas de recuperación de información, muy parecidos a nuestra propuesta, son los siguientes. (Zhongyu y col., 2007) mezcla ontologías, agentes inteligentes y mecanismos semánticos, para permitir que la máquina "entienda" la información, lo que facilita la búsqueda inteligente. Para ello usan las anotaciones de las páginas web hechas en XML, crean ontologías para proporcionar un vocabulario RDF y desarrollan agentes inteligentes que sean capaces de hacer uso de ellas. (Pérez-Agüera y col., 2010) analizan los problemas específicos de bibliotecas de Recuperación de Información (IR), como Lucene, para su integración en motores de búsqueda de la web semántica. (Wei y col., 2008) han realizado un estudio exhaustivo acerca de proyectos piloto sobre sistemas de búsqueda semántica. Además, formalizan un marco general de búsqueda semántica y clasifican la investigación de la búsqueda semántica en seis categorías: búsqueda orientada al documento, búsqueda orientada a la entidad y al conocimiento, búsqueda basada en información multimedia, búsqueda basada en Relaciones, búsqueda basada en minería, y análisis semántico. Se ha propuesto un modelo de búsqueda semántica basado en ontologías para mejorar la búsqueda en repositorios de documentos de gran tamaño (Uren y col., 2007). El modelo de recuperación se basa en una adaptación del clásico modelo de vectores de espacio, con un algoritmo de ponderación de anotaciones, y un algoritmo de clasificación. (Vallet y Col. 2007) proponen otro modelo basado en ontología, el cual incluye un esquema de ontologías para la anotación semi-automática de documentos, y un sistema de recuperación. (Iqbal y col., 2009) presentan un enfoque basado en un tipo de consulta para SPARQL en entorno distribuido, que explota los metadatos de los documentos para realizar un indizado estático para una pre-traída en tiempo de ejecución. En (Giunchiglia y col., 2010) se presenta un nuevo enfoque, llamado concepto de búsqueda, que extiende la búsqueda sintáctica con la búsqueda semántica, basada en el cálculo de las relaciones semánticas entre conceptos. La idea clave del concepto de búsqueda es funcionar con conceptos complejos y aprovechar al máximo la información semántica disponible, reduciendo el uso de la búsqueda sintáctica sólo cuando sea necesario (cuando no hay información semántica disponible). (Strasunskas y col., 2010) revisan un conjunto de sistemas de búsqueda semántica y sus métodos de evaluación y conceptualizan y definen una propuesta de evaluación integral de sistemas de búsqueda semántica. (Amudaria y col., 2011) proponen un patrón semántico de los contenidos de la consulta en formato de una matriz de términos de Documento (TDM por sus siglas en ingles). Además, incorporan una técnica de procesamiento de lenguaje natural, junto con synset (WordNet) para el refinamiento y expansión de la consulta. (Cantador y col., 2008) proponen un modelo híbrido de recomendación donde las preferencias del usuario y las características de las páginas web se describen en términos de conceptos semánticos definidos en ontologías de dominio. La explotación de la meta-información que describe los temas recomendados y los perfiles de usuario, junto con la capacidad de inferir conocimiento a partir de las relaciones que se definen en las ontologías, son los aspectos clave de la propuesta. (Martin y col., 2009) proponen usar ontologías contextuales basadas en perfiles de usuarios para personalizar motores de búsqueda. Eso les permite definir perfiles inteligentes de búsqueda, estableciendo formas de correspondencia entre grupos de usuarios y sus preferencias. Finalmente, un experimento interesante es "Wonder Wheel" o "rueda de búsquedas" de Google. Es una aplicación interactiva que se inicia con la palabra clave en el centro y otros términos relacionados a su alrededor. Al hacer clic en un término relacionado crea un nuevo círculo, conectado con más términos relacionados. Cada vez que se hace clic en un término, en el extremo derecho los resultados web cambian para reflejar el tema actual de interés.

Este trabajo propone un sistema de búsqueda por contexto para la web, basado en una ontología de contexto que combina los perfiles de usuario, los documentos y los sitios en la web, el cual puede ser añadido a cualquier buscador clásico en internet (Google, Yahoo, etc.) para personalizarlo. A partir de la ontología se propone un sistema de reglas genéricas, a ser utilizado por el sistema de inferencia en la tarea de pre-procesamiento (expansión) de la consulta. Este trabajo está organizado como sigue, en la sección 2 se presentan algunos aspectos teóricos de base para nuestra propuesta, la sección 3 presenta el diseño del modelo de búsqueda, la sección 4 presenta cómo usarlo, la sección 5 analiza sus rendimientos, para culminar presentando los trabajos futuros y conclusiones.

2 Búsquedas en la Web Actual

La búsqueda en la web actual está constituida por la búsqueda en archivos almacenados en los ordenadores, llamados servidores web, que contienen una extensa base de datos sobre páginas web. El usuario se conecta a un buscador e indica palabras representativas del tema sobre el que está buscando información, que se utilizan como clave de búsqueda. El resultado de la búsqueda se muestra al usuario como una lista de enlaces a páginas web, en cuya descripción o contenidos aparecen las palabras claves suministradas (Aguilar, 2009). Los tipos de buscadores en Internet son (Baeza, 2005), (Jansena y col., 2006) (Rojas y col., 2010), (Strasunskas y col., 2010):

- Buscadores automáticos: Son aquellos que a partir de cierta información entregada en lenguaje natural, o en alguna especificación, puede deducir y recuperar las páginas web que contengan las palabras claves introducidas.
- Buscadores temáticos: Son una guía jerárquica de directorios que va desde temas generales a los más particulares. Listan lugares (URLs) y los clasifican en categorías añadiendo comentarios identificativos sobre ellos. Su objetivo es encontrar los documentos que pertenezcan al área temática seleccionada.
- Buscadores especializados: Son parecidos a los buscadores temáticos pero sólo abordan algún área concreta. Suelen ser grandes recopilaciones del conjunto de recursos sobre un tema específico.

La búsqueda en la web tiene varias etapas:

- Necesidad de información: Inicialmente el usuario busca información genérica respecto a un tema, luego información dentro de las aristas del tema, para finalmente escoger una aproximación y profundizar en ella.
- Transformación de la necesidad: El usuario que tiene la necesidad de información, al utilizar la web puede ingresar directamente la URL de un sitio que puede satisfacer su necesidad, dirigirse a un directorio en que se pueda explorar un listado de sitios web por tema, o dirigirse a un buscador web donde se ingresen palabras clave. La mayoría de los buscadores tienen operadores para mejorar la búsqueda.
- Búsqueda: El proceso de búsqueda es transparente para los usuarios, durante él se realizan varias operaciones que casi siempre implican consultar un índice de páginas, que es una representación compacta del contenido de éstas en una base de datos. Luego los ordena según ciertos criterios, los consolida (ej.: elimina duplicados) y los presenta al usuario.
- Revisión de los Resultados: En esta etapa el usuario se enfrenta a una lista de direcciones (URLS), elige una que le parece interesante, la revisa, escoge otra, navega un rato, vuelve atrás, hace una nueva consulta, etc .

Un buscador web tiene tres subsistemas:

- Un Recolector, tiene la tarea de crear una colección de páginas web, para ello, visita una serie de páginas inícia-les, las incorpora a la colección, les extrae los enlaces de esas páginas y verifica si existen. Después, a esas nuevas páginas de esos enlaces verificadas se les repite el proceso, y así constantemente.
- Un Indexador que convierte la colección de páginas web en una estructura más manejable y pequeña, llamada índice. Lo usual es utilizar un índice invertido. En él, la colección es convertida a una lista de palabras, cada una de las cuales apunta a una lista de documentos.
- Un Buscador que recupera ciertas páginas del índice basado en el requerimiento del usuario. Si se pregunta por dos o más términos, el sistema deberá comparar las listas de cada uno de los términos, realizando una unión o intersección, según corresponda. Así, encontrar las páginas web no es difícil. El desafío es encontrar las mejores páginas web. El proceso de ranking (ordenamiento) es crucial para tener una cantidad razonable de páginas web. Una aproximación es comparar las palabras de la consulta con las palabras que hay en las páginas web encontradas. Si la página encontrada contiene una palabra por la que se preguntó al buscador que no aparece en casi ninguna otra página web, entonces eso es una buena evidencia de que la página que estamos mirando es importante. Otra manera de ver la calidad de las páginas es observar los enlaces a una página dada. na página con buen contenido seguramente es referenciada desde muchos índices. Existen otras técnicas que se utilizan para ordenar los resultados de las búsquedas. Ese es un tema de investigación actual.

3 Desarrollo de la Propuesta

3.1 Análisis y Diseño

En esta sección se plantea el análisis y diseño del sistema utilizando la metodología para el desarrollo de sistemas web planteada en (Pressman, 2002).

3.1.1 Etapa de Formulación

Motivación principal de la aplicación y planteamiento del problema: La búsqueda por contexto en la web, hasta los momentos, no posee un sistema para la administración

de usuarios que optimice y mejore la búsqueda de la cual un usuario determinado necesita en un momento dado.
Definición de metas: surge la idea de implementar un sistema de búsqueda por contexto, para lo cual se requiere:

- Diseñar el modelo ontológico del contexto del sistema: contexto de usuario, del documento y de la plataforma.
- Diseñar reglas genéricas para determinar los criterios en los que se basara la búsqueda de los usuarios.
- Implementar el registro personal y preferencial de los usuarios que buscan en la web (profesión, hobby, etc.).
- Implementar un motor de inferencia basado en reglas, que será el encargado de inferir y determinar el contexto de búsqueda.
- Optimizar la búsqueda, usando la información suministrada por el usuario y el motor de búsqueda Google.

3.1.2 Etapa de Análisis

Análisis de la interacción*:* Las personas que interactúan con el sistema son los usuarios y el administrador. El usuario al entrar al sistema, tiene un perfil de búsqueda según sus preferencias, usado para activar las reglas.

Análisis de configuración*:* El sistema residirá en un servidor Internet o Intranet, con acceso a bases de datos. Estará implementado para ser ejecutado desde cualquier estación remota, pero con conexión al servidor y haciendo uso de un navegador.

SEDERPIC es la definición que se dará al sistema, el cual emplea un motor de búsqueda (en nuestro caso Google) para la búsqueda en Internet.

3.1.3 Descripción de los módulos

Módulo Ingresar: permite al usuario acceder al sistema por medio de un login y un password, o la primera vez hacer una solicitud de acceso al sistema, para lo cual debe completar solo por esa vez los datos que el sistema requiere (preferencias del usuario, etc.).

- Módulo Contactar: permite una comunicación entre los usuarios y el administrador del sistema.
- Módulo Usuarios: Permite modificar y eliminar un usuario que está en el sistema. Solo puede ser accesado por el administrador del sistema.
- Módulo Instancias de Reglas de la Base de Conocimiento (B.C): permite ver las reglas genéricas instan, modificarlas y eliminarlas. Solo accesado por el administrador.
- Módulo Reglas Genéricas de la B.C: permite insertar y eliminar reglas genéricas. Solo accesado por el administrador.
- Módulo Contáctenos: es solo para el administrador y le permite chequear los mensajes que los usuarios del sistema han escrito.
- Módulo Buscar: es el encargado de la búsqueda. Para eso, realiza un conjunto de inferencias con el motor de inferencia del sistema, antes de proceder a llamar a Google. Después hace el llamado a Google usando para la búsqueda la información inferida.

3.1.4 Etapa de Diseño

Diseño arquitectónico: La fig. 1 ilustra la estructura arquitectónica del sistema. El sistema esta compuesto por las ontologías de contexto (taxonomías de los usuarios y sus entornos, que permiten caracterizarlos); por la base de conocimientos (compuesta por las reglas genéricas y preferencias del usuario, determinadas por las relaciones que surgen de las ontologías); por la consulta que suministra el usuario; y por el motor de inferencia (deduce la respuesta más exacta posible de la búsqueda por contexto).

Diseño de las ontologías de contexto: se basa principalmente en las características más importantes de un usuario, de los sitios desde donde un usuario accesa al sistema, y de las páginas web instanciadas por el usuario después de realizar la búsqueda. Esas tres ontologías permiten contextualizar la búsqueda. La fig. 2 ilustra el diseño detallado de los tres modelos ontológicos:

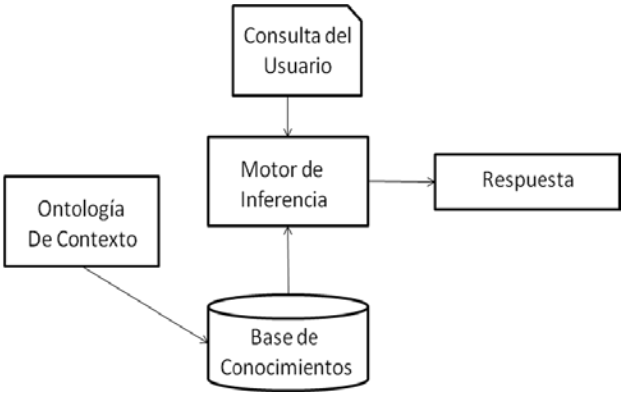

Fig. 1. Estructura arquitectónica del sistema

1. Contexto de Usuario: Es una clase de la superclase de ontología por contexto que categoriza a una persona desde el punto de vista personal, laboral e intelectual. Ella tiene definida los siguientes atributos:

- Nombre y Apellido: atributos que serán el identificador textual de un usuario determinado.
- Sexo (F o M): Es un campo que influye en la búsqueda por contexto ya que determina puntos de vistas diferentes.
- Edad: Es un valor característico que categoriza al usuario, ya que la información de interés de un adulto no es la misma que la de un adolescente.
- País: se refiere al país de origen de un usuario.
- Idioma: Es la lengua del usuario, y permite ver en que idioma un usuario se maneja o desenvuelve.
- Preferencias: son los atributos que caracterizan a un usuario como su: profesión, hobby, deportes, etc. Estos campos son de gran importancia para la búsqueda por contexto.

2. Contexto de documento: Es una clase que describe las características de un documento (página web). Sus atributos son:

- Palabras Claves: contiene las palabras que caracterizan al documento, ese es el campo usado para realizar búsquedas de documentos.
- Tipo de documento: es el formato en que un usuario desea conseguir un determinado documento, puede ser de tipo pdf, ps u otros.

3. Contexto de Plataforma: Es una clase que describe la ubicación actual de un usuario. Sus atributos son:

- Ubicación Actual (# IP): Es una clase a través de la cual el sistema captura por medio del número de IP la ubicación geográfica específica del usuario en ese momento (estado, país, institución, etc.).

Diseño de la Base de Conocimientos: El diseño de la base de conocimientos se basa principalmente en las Reglas Genéricas surgidas de la Ontología del Contexto. Esta base de conocimientos está basada en estructuras condicionales: Si (antecedente) entonces (consecuente). Antecedente y consecuente corresponden a atributos de las taxonomías ontológicas previamente indicadas. De la información de las ontologías, nosotros usamos cuatro atributos del contexto de usuario, las cuales fueron: Profesión, Hobby, Deporte, Noticia; mientras que de los otros contextos ontológicos nosotros usamos los atributos tipo de documento y ubicación actual. Con esas variables se construyen las reglas genéricas de la base de conocimientos. Ese es un aspecto innovador de nuestra propuesta, que fácilmente puede ser escalable a más reglas genéricas, si incorporamos otras variables de las ontologías que podrían ser interesantes de considerar. A continuación un ejemplo de las reglas genéricas existentes en la Base de Conocimiento.

Si (profesión) entonces búsqueda (hobby).

Si (profesión) entonces búsqueda (deporte)

Si (profesión) entonces búsqueda (noticia).

Si (profesión) entonces búsqueda (noticia y tipo de documento).

Si (profesión) entonces búsqueda (deporte y ubicación actual).

Si (tipo de documento) entonces búsqueda (ubicación actual).

Si (tipo de documento) entonces búsqueda (hobby y deporte)

Si (ubicación actual) entonces búsqueda (hobby).

Si (ubicación actual) entonces búsqueda (deporte).

Si (profesión y hobby) entonces búsqueda (noticia y deporte).

Si (profesión y hobby) entonces búsqueda (tipo de documento y noticia).

Y (deporte) entonces búsqueda (noticia y ubicación actual).

Si (noticia y ubicación actual) entonces búsqueda (profesión).

Si (profesión y deporte) entonces búsqueda (noticias).

Si (profesión y deporte) entonces búsqueda (hobby).

Si (noticia y tipo de documento) entonces búsqueda (profesión).

Si (tipo de documento y ubicación actual) entonces búsqueda (el hobby).

Si (tipo de documento y ubicación actual) entonces búsqueda (profesión)

Si (profesión, hobby y noticia) entonces búsqueda (deporte).

Si (profesión, hobby y deporte) entonces búsqueda (ubicación actual y noticia).

Si (profesión, noticia y deporte) entonces búsqueda (hobby y tipo de documento).

Si (profesión, hobby, noticia y tipo de documento) entonces búsqueda (deporte y ubicación actual).

Si (profesión, noticia, deporte y ubicación actual) entonces búsqueda (hobby y tipo de documento).

Las reglas genéricas antes presentadas representan la combinación interesante entre los atributos seleccionados (existe una relación de causalidad entre ellos), pero como dijimos antes, nuestro enfoque permite añadir otras.

3.2 Implementación

3.2.1 Arquitectura del sistema

La arquitectura de nuestro sistema está compuesta por tres capas: La capa de presentación que está constituida por las interfaces a los usuarios y al administrador del sistema. La capa intermedia corresponde al servidor de la aplicación y al motor de inferencia, el cual trabaja directamente con la capa de base de datos donde se almacenan los datos de los usuarios y la base de conocimientos. La entrada, el proceso y la salida del sistema de búsqueda por contexto del sistema son:

Entrada: Datos suministrados por el usuario y datos capturados automáticamente por el sistema (del sitio). Los datos del usuario pueden ser el perfil del usuario (sus preferencias), y palabras claves de información a buscar.

Proceso 1: Deducciones del Motor de Inferencia a partir de las reglas que se activen. Esto permite darle valor semántico (contextualizar) a la búsqueda.

Proceso 2: Una vez definido la información a buscar (ya contextualizada), se pasa a Google para que haga la búsqueda por Internet.

Salida: Resultado de la búsqueda realizada por Google.

4 Uso de la Plataforma

4.1 Menú del SEDERPIC: Administrador

Permite eliminar usuarios del sistema, insertar y modificar preferencias de los mismos, e insertar y eliminar reglas de la base de conocimientos.

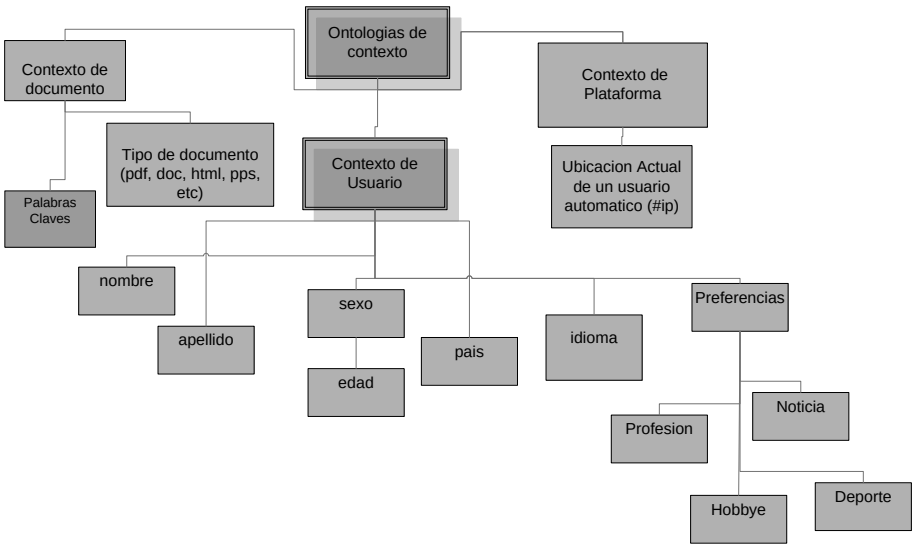

Fig. 2. Diseño del Modelo Ontológico del Contexto

4.1.1 Módulo de Reglas Genéricas de la B.C

Este módulo permite insertar o eliminar una regla genérica (haciendo combinaciones de los atributos de la ontología de contexto). A partir de dichas reglas genéricas se generan las instancias de reglas, usando para ello las preferencias de los usuarios (ver fig. 3). En el momento de insertar una regla genérica, el sistema permite elegir de un listado de atributos el antecedente y consecuente de la regla (fig. 4), validando que sean atributos diferentes en el antecedente y en el consecuente. También permite ver las instanciaciones de las reglas genéricas (ver fig. 5).

Fig. 3. Reglas genéricas de la base de conocimientos

4.1.2. Módulo Preferencias

Este módulo permite al administrador insertar, modificar o eliminar una preferencia de la base de datos y actualizar y depurar la misma, agregando o eliminando valores a los atributos o atributos a la ontología de contexto. (ver fig. 6 para el atributo profesión los valores que puede tomar.

Fig. 4. Insertar una nueva regla genérica

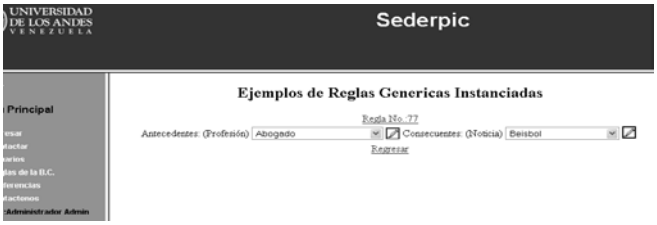

Fig. 5. Ejemplo de Reglas Genéricas Instanciadas

4.2. Menú del SEDERPIC. Usuario

En este menú se ingresa al sistema por medio de un nombre y una clave (login y password), o se solicita un ingreso nuevo al sistema. En la fig. 7 se muestra el ingreso de un usuario que ya se encuentra registrado. Si un usuario no se encuentra registrado debe solicitar el ingreso por medio de un enlace que se encuentra en la página principal del sistema llamada solicitud de acceso al sistema, para crear su perfil de usuario.

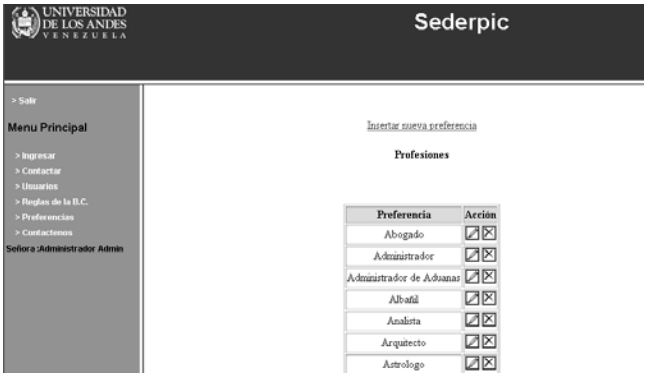

Fig. 6. Instancias de cada preferencia (ejemplo: Profesiones)

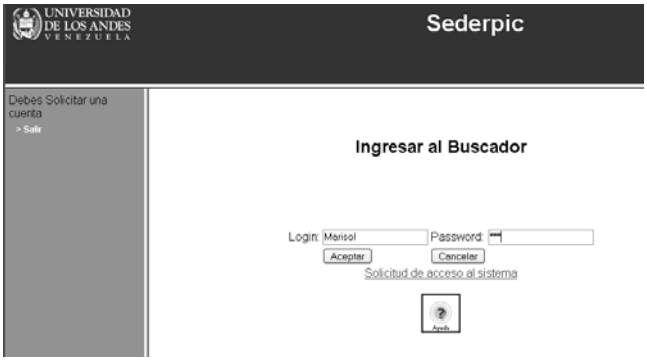

Fig. 7. Ingreso de un usuario

El usuario suministra sus datos y preferencias: Idiomas, Profesiones, Hobby, etc. Si una preferencia no existe, el usuario puede agregarla.

4.3 Programa principal de búsqueda

Este Módulo funciona como el motor de inferencia del sistema de búsqueda por contexto. El mismo se activa en el momento en que el usuario escoge en la página principal de búsqueda (ver fig. 8) lo que desea buscar (palabras claves), a lo cual se le agrega la contextualización hecha basada en los atributos del marco ontológico seleccionado por el usuario (deporte, mi hobby, noticia y mi profesión), el sitio actual, el formato y titulo de documento, etc., tal que puedan ser usados por el motor de inferencia para dicha tarea. Es decir, esos serán los atributos que indican los antecedentes de las reglas que se deben activar en la B.C para determinar que otros aspectos agregar a la búsqueda. Con esa información (palabra clave, más la generada por el motor de inferencia), y antes de ir a Google, se le presenta al usuario todas las posibles combinaciones de búsqueda como resultado del proceso de contextualización (ver fig. 9), para que el seleccione una que será la que definitivamente se le pase Google.

La fig. 10 ilustra el resultado de la búsqueda en una primera llamada a Google con el primer resultado del proceso de contextualización mostrado en la fig. 9. En ese caso, la consulta Proyectos de Puentes ha sido contextualizada (expandida) con los atributos de la ontología Profesión= Arquitectura, Hobby=Comer Dulces, Hobby= Creyones (se refiere al hobby de pintar con creyones). Podemos constatar que para la regla genérica 75 de la fig. 3 a la base de esa extensión hay dos instancias posibles, por eso la contextualización con dos hobbies distintos.

Fig. 8. Búsqueda de un usuario y menú de preferencias

Fig. 9. Resultado del proceso de contextualización a pasar a Google

Fig. 10. Búsqueda de Google usando primera contextualización

5 Evaluación de Rendimiento del Sistema

No hay un marco de referencia estándar para analizar los sistemas semánticos de recuperación de información, hay algunos trabajos que hacen un análisis del estado de artes (Strasunskas y col., 2010), algunas métricas clásicas re-

ferenciados por varios autores (ver Baeza, 2005), sin embargo sigue siendo un tema abierto de investigación. Ahora bien, cuando se evalúa un sistema de búsqueda semántico se persiguen dos objetivos: probar sus ventajas con respecto a motores de búsqueda existentes, y mostrar sus potenciales usos. La primera parte de esta sección cubre el primer objetivo, y en la segunda parte el otro objetivo.

5.1 Comparación con otros enfoques

Hay dos enfoques clásicos para poder comparar sistemas como el propuesto en este trabajo con otros trabajos. Uno donde usuarios reales son consultados sobre los resultados arrojados por el sistema de búsqueda; y el segundo donde de manera automática, ciertas métricas son calculadas usando información proveniente de la consulta y de la información recuperada. Nosotros en esta fase consideramos ambos, comenzando por el segundo. Para el segundo caso definiremos métricas que nos permitan medir la calidad de la recomendación de la búsqueda por contenido de nuestro sistema. Para comprender esto, nuestro enfoque de recuperación debe ser visto como una evolución de las técnicas de búsqueda basadas en claves de los motores de búsqueda clásicos (Yahoo, Google, etc.), cuyos índices claves se enriquecen semánticamente. Nuestro sistema toma la consulta (palabras claves, una frase en lenguaje natural, etc.) y le agrega contenido semántico según el perfil del usuario. El motor de consulta que instancia nuestro sistema (en este caso Google) regresa un listado de páginas web que supuestamente corresponden a la preferencias del usuario Dichas páginas web tienen anotaciones, entre los cuales están las palabras claves que lo describen. Nosotros definimos una métrica de similaridad semántica para determinar la calidad de respuesta de nuestro sistema de búsqueda. D es el conjunto de páginas web en el espacio de búsqueda y U el conjunto de preferencias de los usuarios. Cada página web en el espacio de búsqueda será el vector $d \in D$, y dx es cada uno de las anotaciones de interés para nuestro estudio (palabras claves descriptoras de la página web). dx indica cuales son las palabras claves que indexan las páginas recuperadas de una consulta por Internet. $Um \in U$ es un vector de preferencia que representa los valores de los atributos del perfil de cada usuario (sus preferencias), tal que Um= (um1, um2,...,umk) donde k es el numero de preferencias (atributos) que tiene un usuario dado (cada um guarda el tipo del atributo y su valor para ese usuario específicos). Basado en esos vector podemos calcular una métrica, que llamaremos similaridad, para comparar los vectores de anotación y preferencia para las primeras páginas recuperadas en la consulta a Google (motor de búsqueda usado por nuestro sistema). Esa medida de similaridad entre una consulta de un usuario con perfil Um y una página web recuperada con descriptores dx's es computada como:

$$\text{Similaridad}(U_m, d_x) = \text{match}(U_m, d_x)/|U_m|$$

donde $|U_m|$ es el numero de atributos del perfil del usuario U_m a cubrir, y *match* determina cuantos de los atributos de U_m son descriptores de la página web con descriptores d_x. Esta muestra la eventual satisfacción del usuario ante la respuesta que se le da.

Hemos hecho dos pruebas, una usando los usuarios U1, U2, U3 y U5 de la tabla 2 (ver fig. 11), y otra en la cual el usuario U1 hace varias consultas (caso 1: "Proyectos de Puentes", caso 2: "Vuelos entre Caracas y Mérida", caso 3: "Cursos de Linux en la ULA"), ver fig. 12. Los puntos que se muestran, en el caso de la fig. 11 son los promedios de la métrica de similitud para los diferentes usuarios, y en la fig. 14 el promedio de las 20 primeras páginas para cada caso. Estos resultados se comparan con Google y (Vallet y col., 2007).

Nuestro sistema es capaz de mejorar los resultados de Google, por incorporarle información semántica a la consulta. Con respecto a (Vallet y col., 2007), nuestros resultados son similares, si bien es cierto que en ese trabajo se proveen otros aspectos de interés como la capacidad de generar procesos de inferencia desde las ontologías existentes (nuestro sistema se limita a las reglas genéricas preestablecidas en el diseño). Ambos agregan explicita información al proceso de búsqueda, de manera de personalizarla (expanden la consulta).

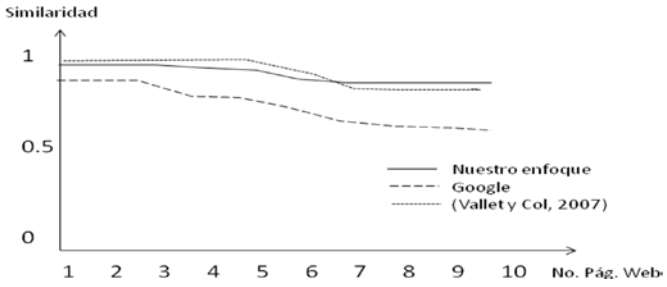

Fig. 11. Comparación de la métrica de similaridad en diferentes usuarios

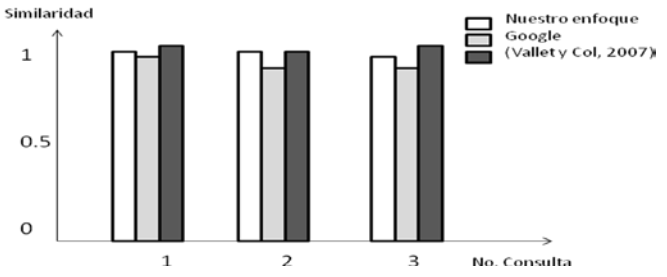

Fig. 12. Comparación de la métrica de similaridad en diferentes Consultas

La segunda experiencia está basada en el primer enfoque de comparación, donde se involucran a los usuarios en el proceso. En este caso se hacen las tres consultas definidas anteriormente, definidas como caso 1, 2 y 3. Las 10 primeras páginas web recuperadas se dieron a usuarios reales (en nuestro caso 2), quienes las evaluaron de acuerdo al interés que para ellos representaban, para calcular una métrica de

relevancia propuesta, (Strasunskas y col., 2010). Esta métrica sustituye la clásica métrica llamada "precision and recall" propuesta en (Baeza, 2005)). Ella se define como:

$$\text{Relevancia} = \sum_{i=1}^{2}\sum_{j=1}^{10}\frac{Pij}{20}$$

donde cada Pij es la evaluación del individuo i de la página web j (sí la página web j es muy relevante para el usuario i él colocará 1, y 0 en el otro extremo). Otras métricas existen en la literatura que son modificaciones de ella: Geometric Mean Average Precision, Precision after X documents, R-Precision. Como todas son variantes de la primera, usamos esa en nuestro análisis. En este caso comparamos nuestro trabajo con (Pérez-Agüera y col., 2010) y (Strasunskas y col., 2010).

Tabla 1: Resultados según la relevancia para diferentes consultas

Técnica vs Consulta	1	2	3
Nuestro	0,95	0.85	0.93
(Pérez-Agüera y col., 2010)	0.97	0.83	0.92
(Strasunskas y col., 2010).	0.93	0.86	0.90
Google	0.59	0.51	0.60

Tabla 1 nos muestra los resultados de la búsqueda. Vemos que los tres trabajos comparados obtienen resultados similares debido a que proveen a los usuarios páginas web parecidas. Es decir, las páginas web recuperadas por cada uno de los trabajos comparados son similares. Con respecto a Google, vemos también como los tres lo superan en todos los casos. Ahora bies, se nota que la relevancia en la segunda consulta no es tan alta, indicando una incertitud de los usuarios sobre la relevancia de las páginas web recuperadas.

5.2 Comparación de la calidad de Agrupamiento de la información

En este experimento hemos partido del grupo de 10 perfiles de usuarios definidos en la tabla 2, haciendo todos la consulta del caso 1. Para poder comparar con (Cantador y col., 2010), en el sistema de ellos hemos considerados el grado de interese de los usuarios en esos tópicos alto, y cuando no lo tiene se colocan en bajo.

Tabla 2. Atributos para comparar capacidad de agrupamiento

Usuario	Profesión	**Hobby 1**	**Hobby 2**	Deporte
U1	Arquitecto	Pintar (con creyones)	Comer	Fútbol
U2	Ingeniero			Beisbol
U3	Medico	Animales		
U4	Arquitecto	Pintar (con creyones)	Comer	Fútbol
U5	Arquitecto	Pintar		Futbol

Según esa tabla, los usuarios U1 y U4 tienen perfiles similares, y ciertamente, en (Cantador y col., 2010) los colocan en un mismo grupo de usuario. Según fig. 9, nuestro sistema les provee las mismas opciones de contextualización a ambos (una forma indirecta de establecer esa idea de grupos). Para el usuarios U5 que comparte ciertos atributos con U1 y U4, usando el sistema de (Cantador y col., 2010) son agrupados junto al mismo grupo de usuarios anterior. En el caso de nuestro sistema eso se refleja en las opciones de contextualización que nuestro sistema provee, donde algunas de ellas son comunes y otras no (ver fig. 13). Para el resto, por ser diferentes, en nuestro sistema se proveerían opciones de contextualización diferente para cada usuario, y en el caso de (Cantador y col., 2010), los coloca en grupos distintos. Nuestro sistema no hace una agrupación, sino que provee opciones de contextualización más o menos similares. Analizando el trabajo de (Cantador y col., 2010), ellos usan un criterio de similaridad basada en la distancia Euclidiana entre los conceptos; nuestro sistema, sin usar esa medida, permite regenerar esa idea en las opciones de contextualización que provee, las cuales pueden ser usadas (según sus semejanzas) para realizar un proceso similar al propuesto en (Cantador y col., 2010). Nuestro sistema automatizaría la expansión de la consulta según los perfiles de usuarios (usando su ontología). A partir de allí, se podría hacer el proceso agrupamiento usando métricas semejantes a las usadas en (Cantador y col., 2010).

Fig 13. Resultado del proceso de contextualización para U5

6 Conclusiones

Este trabajo ha propuesto un sistema de Búsqueda por Contexto por Internet, empleando como motor de búsqueda a Google. La Búsqueda por Contexto es una herramienta apropiada para explotar el conocimiento generado de la interacción web – usuario, usando modelos que formalizan patrones de uso y caracterizan los perfiles de los distintos grupos de usuarios que hacen uso de este medio.

La búsqueda es basada en la información aportada por las reglas genéricas instanciadas, que se activan según la información del usuario, a la cual se le asocian las palabras claves que el usuario introduce en ese momento.

Este trabajo representa un posible inicio para la creación de sistemas de búsqueda por contexto más enriqueci-

dos, que soporten más datos sobre el perfil de un usuario, creando ontologías de contexto con muchas más características, conceptos y relaciones, y de esta manera poder crear y generar reglas que le den mucho más valor semántico a la búsqueda por contexto de un usuario. Ella permite: I) Mejor recuerdo en las consultas al añadirles palabras de interés. II) Mejor precisión mediante el uso de consultas estructuradas semántica, que expresan las necesidades de información más precisa. Nuestro modelo de recuperación semántica depende de la calidad de la información guardad sobre los usuarios (sus perfiles) y de las reglas genéricas.

Futuros trabajos vincularán nuestro esquema de búsqueda por contexto a enfoques de Manejo de la Web Semántica basada en Ontologías Dinamices, como la presentada en (Rodríguez y col., 2010), de manera de enriquecer el proceso de contextualización con la información contenida en esas ontologías. También, otros trabajos tendrán como objetivo desarrollar una versión dinámica de las reglas genéricas (que se puedan aprender y desaprender según el contexto donde se usen).

Agradecimiento

Al proyecto del CDCHT I-1237-10-02-AA de la Universidad de los Andes por su apoyo financiero.

Referencias

Aguilar J A, 2009, Web Mining System, WSEAS Transactions on Information Science and Applications, Vol. 6, No. 9, pp. 1523-1532.

Amudaria S, Sasirekha S, 2011, Design of Content Oriented Information Retrieval Based on Semantic Analysis, International Journal of Computer Science and Information Security, Vol. 9, No. 1, 92-97.

Albrecht S, Florian W, Martin K, Florescu D, Manolescu L, Carey M, Busse R, 2001, The XML Benchmark Project, Information Systems, No. 3, pp.1-17.

Baeza R, 2005, Excavando la web, El Profesional de la Información, Vol. 13, No. 1, pp. 4 – 10.

Cantador A, Bellog P y Castells P, 2008, A multilayer ontology-based hybrid recommendation model. AI Commun. Vol. 21, No. 2-3, pp. 203-210.

Giunchiglia F, Kharkevich U y Zaihrayeu, I, 2010, Concept Search: Semantics Enabled Information Retrieval. Technical Report DISI-10-004, Ingegneria e Scienza dell'Informazione, University of Trento.

Iqbal A, Ott M y Seneviratne, A, 2009, Semantic Information Retrieval in a Distributed Environment, Proceeding 6th IEEE Consumer Communications and Networking Conference, pp. 1 – 5.

Jansena B y Spink, A; 2006, How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, Vol. 42, No. 1, pp. 248-263.

Li D, Finin T, Anupam J, Pan R, Scott R, Peng, Reddivari P, Vishal D y Sachs J, 2004, Swoogle: a search and metadata engine for the semantic web. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, New York, USA, pp. 652-659.

Martin A, Celestino S, Valdenebro A, Mensaque J, 2009, Perfil de ontología para la recuperación de la información, Proceeding IX Congreso ISKO, pp 155-169

Pérez-Agüera J, Arroyo J, Greenberg J, Perez-Iglesias J, Fresno V, 2010, Using BM25F for Semantic Search. Proceeding Semantic Search Workshop at the 19th Int. World Wide Web Conference WWW2010 April 26, 2010 (Workshop Day), Raleigh, NC, USA

Pressman R, 2002, "Ingeniería de Software: Un enfoque Practico", 5ta.Edicion, McGraw-Hill.

Rojas M, Silva E, Cristo M, Philippe T, Soares A, 2010, Exploring features for the automatic identification of user goals in web search, International Journal in Information Processing and Management, Vol.46 No.2, p.131-142,.

Rodríguez T, Aguilar J, Puerto E, 2010, Dynamic Semantics Ontological Framework for web Semantics, Proceeding of the 9th WSEAS Intl. Conference on Computational Intelligence, Man-Machine Systems and Cybernetics (CIMMACS '10), pp. 91-98.

Rose D; Levinson D, 2004, Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web, New York, USA, pp. 13-19.

Shadbolt, Nº Hall, W; Berners-Lee, T., 2006, The Semantic Web Revisited, IEEE Intelligent Systems, Vol. 21, No. 3, pp. 96 – 101.

Strasunskas D, Tomassen, S, 2010, 'On Variety of Semantic Search Systems and Their Evaluation Methods', Proceedings of the International Conference on Information Management and Evaluation, pp. 380-387.

Uren V, Lei Y, Lopez V, Liu H; Motta E, Giordanino M, 2007, The usability of semantic search tools: a review. The Knowledge Engineering Review, Vol. 22, No 4, pp. 361–377.

Vallet D, Castells P, Fernandez M, Mylonas M, Avrithis Y, 2007, Personalized Content Retrieval in Context Using Ontological Knowledge. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on "The Convergence of Knowledge Engineering, Semantics and Signal Processing in Audiovisual Information Retrieval", Vol. 17, No. 3, pp. 336-346.

Wei W, Barnaghi P, Bargiela A, 2008, Search with Meanings: An Overview of Semantic Search Systems, Int. J. Communications of SIWN, Vol. 3, pp. 76-82.

Zhongyu L, Umair R, 2007, Semantic search technology for information retrieval on the web , Int. J. of Agent-Oriented Software Engineering, Vol. 1, No. 2, pp. 225 - 243.

Recibido: *15 de enero de 2011*

Revisado: *17 de junio de 2011*