# Modelo de un meta-buscador web semántico basado en una taxonomía general de conocimiento, una ontología de dominio general, ontologías específicas y perfil de usuario

***HUGO ORDOÑEZ ERASO***
*Ingeniero de Sistemas, Magíster en Computación*
*Profesor, Facultad de Ingeniería, Universidad Mariana*
*Miembro del Grupo de I+D en Tecnologías de la Información (GTI), Universidad del Cauca*
*hugoeraso@gmail.com*
*San Juan de Pasto, Nariño, Colombia*

***CARLOS ALBERTO COBOS LOZADA***
*Ingeniero de Sistemas, Magíster en Informática, Ph.D. (c) en Ingeniería de Sistemas y Computación*
*Profesor Titular, Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones*
*Director del Grupo de I+D en Tecnologías de la Información (GTI), Universidad del Cauca*
*ccobos@unicauca.edu.co*
*Popayán, Cauca, Colombia*

***ELIZABETH LEÓN GUZMÁN***
*Ingeniera de Sistemas, Magíster en Ingeniería de Sistemas, Ph.D. en Ciencias e Ingeniería de la Computación*
*Profesora Asistente, Departamento de Ingeniería de Sistemas e Industrial, Facultad de Ingeniería*
*Directora del Grupo de I+D en Minería de Datos (MIDAS), Universidad Nacional de Colombia sede Bogotá*
*eleonguz@unal.edu.co*
*Bogotá, Colombia*

*Fecha de recibido:*12/10/2010
*Fecha de aprobación:* 15/06/2011

**RESUMEN**

La búsqueda web en los últimos años se ha convertido en una de las áreas de investigación más importantes del mundo, debido entre otras cosas: al crecimiento acelerado de las fuentes de información, a la necesidad de contar con información más relevante a los requerimientos específicos de cada usuario, a la exploración de menores tiempos de búsqueda y a la falta de usar la semántica de los términos implicados en las consultas. En este artículo se presenta el modelo de un meta-buscador (usa los recursos indexados por Google, Yahoo! y Bing) web semántico llamado XGhobi, que incorpora una taxonomía general de conocimiento, una ontología de dominio general (WordNet), un conjunto de ontologías de dominio específico y el perfil de los usuarios para mejorar la relevancia de los documentos recuperados tanto en inglés como en español. Se describe en detalle los componentes del meta-buscador, algunas interfaces de usuario y los resultados de su evaluación. La evaluación del sistema muestra la precisión obtenida en pruebas realizadas con usuarios.

**PALABRAS CLAVE:** Meta-buscador web, Taxonomía, Ontología, WordNet, Perfil de usuario.

**ABSTRACT**

Web search has become one of the most important fields of research around the world. They are many reasons including: the fast-growing nature of information sources; the search necessity for information closer to specific user requirements; the need to reduce search time; and the desire to take into account the semantics of terms used when doing search queries. This paper shows a semantic meta-web search model called XGhobi which uses indexed resources by Google, Yahoo! and Bing. The XGhobi engine combines a general taxonomy of knowledge, a general domain ontology –WordNet-, a set of specific domain ontologies, and user profile management to improve the relevance of recovered documents in both English and Spanish. A detailed description of the meta-web search engine's components, some user interfaces and its results and its assessments are shown. The assessment covers the obtained precision on tests done by users.

**KEYWORDS:** Meta-web searcher, Taxonomy, Ontology, WordNet, User profile.

## 1. INTRODUCCIÓN

Desde su aparición, la World Wide Web (WWW o simplemente web), se ha convertido en un instrumento de uso cotidiano en nuestra sociedad. La web es hoy en día uno de los medios de mayor almacenamiento de información, cuenta con fácil acceso y es económicamente accesible para gran cantidad de usuarios. Debido a lo anterior, la web se ha convertido en un instrumento definitivo en el desarrollo social en todos los campos. El modelo de la web ha permitido compartir información entre distintas instituciones o entre unidades y departamentos de la misma organización o grupos de carácter mundial. Sin embargo, la heterogeneidad y descentralización de las fuentes de información que la web presenta ha provocado que cuanta más información hay accesible, más difícil es localizar lo que se busca [1, 2].

Por lo anterior, la búsqueda web se ha convertido en un espacio interdisciplinar de investigación que busca las mejores formas de representar, almacenar, organizar y acceder ítems de información en forma automática [1]. Actualmente, buscadores como Google, Yahoo!, Bing y Ask son muy populares y de gran utilidad cuando se desea recuperar información en la web [1], pero su funcionamiento interno aún presenta falencias en el filtrado, ordenamiento y manejo de la semántica de la información, presentando resultados que muchas veces nada tienen que ver con la las necesidades de información del usuario.

En los actuales sistemas de búsqueda web, el usuario no tiene seguridad de que el sistema le proporcione las respuestas más relacionadas con respecto a sus necesidades [3], o si por el contrario, el sistema le devuelve sólo una parte sesgada de esa respuesta ideal, debido, por ejemplo, al contexto [4, 5] en que se realiza la consulta.

La forma común en que los buscadores tradicionales presentan los resultados a los usuarios es a través de una lista ordenada, lo que en general conlleva a que el usuario lea una serie de páginas con resultados mixtos, gaste mayor tiempo en la lectura de las páginas recuperadas o descarte gran cantidad de resultados, porque sólo lee los primeros documentos recuperados (5 a 10 enlaces de la primera página de resultados) sin tener en cuenta los demás [7]. Esta situación en muchas ocasiones ha generado que el usuario abandone las búsquedas sin obtener los resultado deseados [5].

Adicional al problema de la presentación de los resultados, los buscadores tradicionales sólo indexan y buscan en una parte de la Web; hecho que se agrava, si se tiene en cuenta el crecimiento exponencial de los documentos disponibles en Internet. Por lo anterior, la comunidad académica y científica ha desarrollado un abundante marco teórico, de investigación y de aplicación al rededor de los meta-buscadores [1].

Un meta-buscador [6], toma la consulta de un usuario, envía dicha consulta a diversos motores de búsqueda clásicos o tradicionales (como Google, Yahoo! y Bing) y finalmente retorna los resultados de todos los buscadores tradicionales al usuario. En este proceso el meta-buscador realiza entre otras las siguientes tareas:

- Expandir la consulta, por ejemplo, agregándole semántica a los términos en la consulta por medio de conceptos registrados en una ontología o tesauro [7], o basado en un historial de términos registrados como relevantes para el usuario.
- Filtrar los resultados recibidos de los motores tradicionales, por ejemplo, eliminando documentos duplicados y direcciones que no son válidas.
- Ordenar la presentación de los resultados de acuerdo a las necesidades específicas del usuario [8]. Esto se logra si el meta-buscador cuenta con un perfil que permita definir las necesidades a largo y corto plazo del usuario.

Los meta-buscadores se especializan en el uso de diversas fuentes de documentos, para hacer más extensiva la búsqueda, además buscan que el filtrado y el ordenamiento de los resultados sea más apropiado a las necesidades de los usuarios, pero el tiempo de respuesta normalmente es mayor al de un buscador tradicional [6].

En este artículo se propone el modelo de un meta-buscador web que busca mejorar la relevancia de los resultados presentados a los usuarios, desarrollando un mejor proceso de expansión de consulta, filtrado y ordenamiento de los resultados, basado en mejoras en la semántica de los términos de las consultas (en dos idiomas específicos, inglés y español) y aprovechando de una mejor forma la re-alimentación que los usuarios hacen en el sistema sobre los documentos recuperados (relevantes y no relevantes).

A continuación se describen trabajos relacionados con meta-buscadores, búsqueda semántica y perfil de usuario. Luego en la sección 2 se presenta el modelo

del meta-buscador web semántico propuesto. Después en la sección 3, se describe la funcionalidad general de XGhobi, el prototipo software que implementa el modelo propuesto y que sirve de herramienta para la evaluación del mismo. En la sección 4, se muestran los resultados de la evaluación y finalmente se presentan las conclusiones y el trabajo que en el futuro cercano espera desarrollar el grupo de investigación en relación con el modelo propuesto.

## 2. TRABAJOS RELACIONADOS

### 2.1 Meta-Buscadores Web

A continuación se presenta un resumen de los meta-buscadores más destacados [9]:

- **MetaCrawler** [10] es un meta-buscador desarrollado en la Universidad de Washington, que integra un conjunto general de motores de búsqueda web tradicionales. Cuando una consulta se presenta, MetaCrawler despacha la consulta a cada uno de los motores de búsqueda, recupera el código fuente HTML de todos los documentos devueltos, y aplica un análisis para limpiar los vínculos que no están disponibles y documentos irrelevantes. MetaCrawler obtiene alta precisión mediante la combinación de la cima del ranking de resultados de búsqueda de cada uno de los motores de búsqueda, dependiendo del contexto de las palabras clave.
- **Dogpile** [11] es un meta-buscador que busca en Google, Yahoo!, Bing, Ask, entre otros, después decide cuales resultados son más relevantes para la búsqueda dependiendo de la naturaleza e intención de las palabras y elimina los documentos duplicados. Al final, se obtiene una lista de resultados más completos para la consulta planteada.
- **iXmetafind** [6] es un meta-buscador que hace uso de las bases de datos de los motores de búsqueda más populares, asocia un peso con cada documento usando una variación lineal de la combinación del ranking para cada documento recuperado de las fuentes de los motores de búsqueda, lo que refleja su importancia, este peso determina el ranking de importancia del documento, en lo cual las primeras filas de los resultados son los mejores ranqueados y se despliegan al usuario en ese orden de ranking.
- **Ixquick** (www.ixquick.co) es un meta-buscador que hace una búsqueda simultánea y en forma anónima en muchos motores de búsqueda comunes. Excluye los resultados vistos con anticipación, además permite resaltar lo resultados de buena aceptación para tenerlos en cuenta en búsquedas futuras, con esto realiza una re-alimentación y despliega los documentos más relevantes a la consulta.
- **Webferret** (www.webferret.com) permite realizar consultas mediante lenguaje natural especificando que muestre páginas que contengan todas o alguna de las palabras clave, frase exacta o expresión booleana, presentando la lista de resultados con aquellos documentos que cumplan con reglas de agrupamiento booleanas.
- **Monster Crawler** (www.monstercrawler.com) toma los resultados de los motores de búsqueda (como Yahoo!, Google, MSN, Ask), elimina los duplicados y para proporcionar un conjunto más completo de resultados usa la semántica de los términos específicos de la consulta.
- **Mamma** (www.mamma.com) hace uso de los motores de búsqueda tradicionales como fuentes de información, utiliza un sistema de votación llamado (Condorcet) para el ordenamiento de los resultados, clasificación y eliminación del spam resultado de los motores de búsqueda.

Adicionalmente, existe un conjunto de meta-buscadores que se han especializado en presentar los resultados en grupos de temas relacionados o agrupaciones de documentos web, en lugar de hacerlo convencionalmente a través de una lista ordenada de enlaces. Claudio et al en [12], hacen una revisión muy detallada de cada uno de estos meta-buscadores que realizan agrupación de documentos web (Web Clustering Engines), entre los que se pueden mencionar: Clusty (http://clusty.com), Carrot (http://www.carrot2.org), iBoogie (http://www.iboogie.tv), SnakeT (http://snaket.di.unipi.it), Credo (http://credo.fub.it), Grokker (http://grokker.com), KartOO (http://www.kartoo.com), CIIRarchies (http://www.cs.loyola.edu/~lawrie/hierarchies), WebCAT (http://ercolino.isti.cnr.it/webcat), AISearch (http://www.aisearch.de), SRC (http://rwsm.directtaps.net), EigenCluster (http://eigencluster.csail.mit.edu), WhatsOnWeb (http://gdv.diei.unipg.it/view/tool.php?id=wow) y WebClust (http://www.webclust.com).

### 2.2 La Semántica en la Búsqueda Web

Varias investigaciones han involucrado la semántica en la búsqueda web, en [13] se hace una propuesta de un marco de trabajo general basado en matrices que contempla la semántica de los términos junto con las relaciones estructurales existentes en los documentos web, destacando el impacto de la semántica en el

ordenamiento (ranking) de los documentos. Además en [14] se muestra el uso de ontologías para generar consultas más inteligentes antes de pasarlas a los buscadores tradicionales.

En [15] se muestra un método para la detección de campos pertenecientes a diccionarios, el cual sirve para la integración de datos de distintas fuentes, garantizando que todos los términos de campos equivalentes contengan valores con la misma representación y sean semánticamente iguales. Muestra cómo es necesario disponer de información del contexto para resolver inconsistencias entre sinónimos y que este proceso automático es posible sólo mediante la utilización de diccionarios de términos. Plantea la solución utilizando un enfoque basado en términos predefinidos que siguen una estructura ontológica.

En [16] se presenta el proyecto OntoBus el cual trabaja con una ontología para la cual incluye la definición de los conceptos y relaciones necesarios para el desarrollo de servicios eficientes de búsqueda de información en el dominio de la bibliotecas universitarias, plantea que debido a la gran cantidad de términos identificados la ontología se organiza en varios espacios de nombres, con vocabularios correspondientes para poder describir más en detalle las instancias más básicas de los términos de la búsqueda.

En [17] se propone un servicio Web basado en ontologías orientado hacia el dominio de la diversidad biológica el cual trabaja con ontologías biológicas que proporcionan descripciones sobre la taxonomía, la evolución y la morfología de las especies, así como las relaciones ecológicas y tróficos (es decir, la posición ocupada por una especie en una cadena alimenticia), con el fin de precisar las consultas que realizan los biólogos cuando desean obtener información adicional acerca de conceptos de plantas, insectos y varios tipos de animales, además combina las ventajas de los servidores, el apoyo a la ontología y la gestión en la Web, en donde aclara la independencia de la tecnología y evolución de las aplicaciones que hacen uso de esta.

En [18] se plantea un método basado en un agente de ontologías llamado SWPAO el cual basa la consulta en los principios de exhaustividad, correlación cruzada y coherencia, en donde los usuarios pueden realizar sus consultas en lenguaje natural, para lo cual el sistema realiza una intensa relación entre el contenido y el título del documento y el agente ontológico compara los términos de la consulta con la información de los documentos que se retornan.

En [19] se muestra un método reflexivo de consulta sobre múltiples fuentes de información inteligente mediante la aplicación de una ontología para mejorar la interoperabilidad, tanto estructural como semántica de la inteligencia de las fuentes de información realizando una transformación a las fuentes para que queden lo más homogéneas posibles. Para reformular una traducción entre la búsqueda del usuario y la consulta de origen, se realiza la descomposición de la consulta en sub consultas que se aplican a diferentes fuentes de datos, los documentos retornados por cada una de las sub consultas son recogidos por un componente de composición que los agrupa y retorna al usuario.

En [20] se propone un método de agrupamiento difuso para construcción de una ontología basada en perfiles de usuario, mostrando que la ontología ofrece algunas oportunidades para mejorar el sistema de Recuperación de Información (RI), manteniendo una representación sofisticada de los perfiles de interés personal. En donde estas representaciones pueden ser utilizadas para una RI eficaz.

El proyecto ALVIS [20] apunta a desarrollar un motor de búsqueda de código abierto, con medios de búsqueda semántica extendida. ALVIS intenta procesar la consulta con más precisión, mientras tiene en cuenta el tema y el contexto de búsqueda para refinar la consulta y el análisis del documento. El desarrollo de ALVIS cuenta con una arquitectura de NLP que enriquece los documentos con información lingüística. Esta plataforma se está diseñando para ser genérica en el procesamiento de documentos.

Las propuestas de Mustafa [7], Song [21], Aufaure [22] ya muestran que el uso de ontologías es una forma viable de mejorar los motores de búsqueda tradicionales basados en palabras claves, y proponen los denominado motores de búsqueda semánticos o modelos de recuperación semántica de información, operando sobre colecciones de documentos no estructurados, que no han sido previamente construidos sobre los conceptos de la Web Semántica.

Las anteriores investigaciones resaltan la fortaleza de la recuperación semántica de información con respecto a la tradicional basada en palabras clave, que radica en la información explícita adicional (tipo, estructura, jerarquía, relaciones, reglas) contemplada por ontologías y almacenada en la base de conocimiento sobre los conceptos referenciados en los documentos, frente a los índices planos clásicos que se desarrollan por palabras clave [7]. Pero ninguna de ellas muestra la

forma como se pueden integrar ontologías de dominio general, de dominio específico y el perfil del usuario, más aún, haciendo que todos estos componentes estén ocultos para el usuario, es decir, que no cambien la interfaz que a la fecha están acostumbrados a usar y que se basa en consultas textuales por palabras claves, similar a la de los buscadores web tradicionales.

La búsqueda semántica introduce un paso adicional con respecto a los modelos clásicos de RI: En lugar del barrido de un índice por palabras, la búsqueda semántica procesa una consulta semántica contra la base de conocimiento, y devuelve un conjunto de instancias [7]. Esto puede verse como una forma de expansión de la consulta, donde las instancias devueltas representan un nuevo conjunto de términos de búsqueda, que conduce a un mayor nivel de recuperación. Esta expansión de la consulta se amplía también mediante la inferencia sobre reglas, jerarquías y relaciones.

Por lo anterior, se pueden obtener las siguientes mejoras con respecto a la búsqueda por palabra clave:

- Mayor recuperación en las consultas por clase.
- Mejor precisión por la utilización de consultas semánticas estructuradas.
- Mayor recuperación mediante la utilización de jerarquías de clases y reglas.
- Es posible combinar condiciones sobre conceptos y contenidos.
- Las mejoras con respecto a la búsqueda por palabra clave crecen con el número de condiciones en la especificidad de la consulta formal.

Uno de las herramientas ontológicas más usadas en las investigaciones recientes es WordNet. La cual es una base de datos léxica estructurada a partir de las principales relaciones conceptuales que vinculan entre sí a los lexemas intra- y trans- categorialmente, también puede verse como una ontología para el procesamiento de lenguaje natural [23], que contiene alrededor de 100,000 términos organizados en jerarquías taxonómicas (ver Fig 1), está dividida en cinco categorías gramaticales: sustantivos, verbos, adjetivos, adverbios y palabras funcionales, estas jerarquías se organizan en estructuras léxicas: los nombres en jerarquías léxicas sobre la base de relaciones de hiponimia y meronimia; los verbos con base en relaciones de implicación (entailment), y finalmente, los adjetivos y adverbios se organizan como hiperespacios N-dimensionales.

WordNet se basa en el supuesto teórico de matriz léxica (ver Fig 2). En la matriz léxica el encabezamiento de las columnas ($F_1$) corresponde a las unidades léxicas de una lengua y el encabezamiento de las filas ($M_1$) a los conceptos. Una entrada en una celda de la matriz ($E_{1,1}$) implica que esa forma ($F_1$) puede ser utilizada para expresar el concepto ($M_1$). Esta presentación en columnas y filas permite observar gráficamente dos de los principales temas de la semántica léxica: la polisemia (en caso de que la misma columna cuenta con dos entradas, $E_{1,2}$ - $E_{2,2}$) y la sinonimia (en caso de que la misma fila cuenta con al menos dos entradas, $E_{1,1}$ - $E_{1,2}$) [24].

En esta matriz léxica los conceptos se representan por la lista de unidades léxicas que pueden ser usadas para expresarlo (todas las entradas que pertenezcan a una misma fila), es decir, el conjunto de sinónimos (synset) no explica al concepto sino que simplemente indica que el concepto existe. No obstante, la representación propuesta por esta matriz léxica no puede trabajar directamente con los conceptos, sino que lo hace con las unidades léxicas. La relación léxica principal en WordNet es la sinonimia, pero también están presentes la antonimia, la hiperonimia, la hiponimia, la meronimia y las relaciones morfológicas.

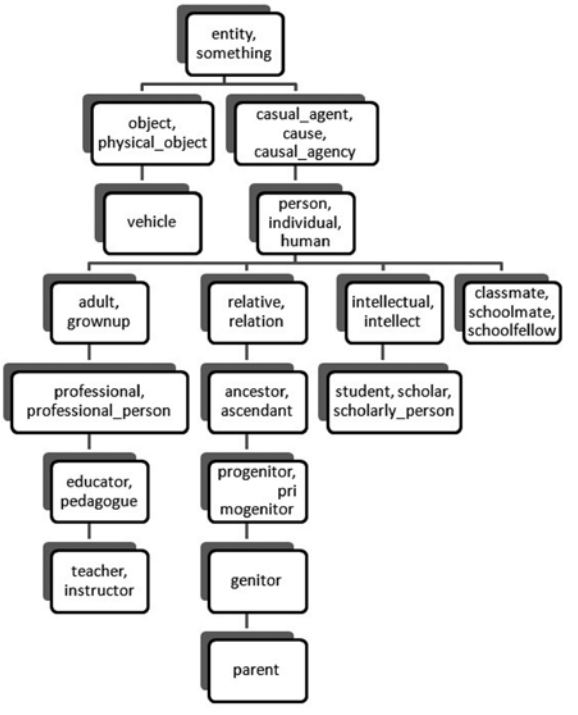


**Figura 1.** *Una parte de la base de conocimiento semántica en WordNet (adaptada de [25])*

| Significado de las palabras | Formas de las palabras | | | | |
|---|---|---|---|---|---|
| | $F_1$ | $F_2$ | $F_3$ | ... | $F_n$ |
| $M_1$ | $E_{1,1}$ | $E_{1,2}$ | | | |
| $M_2$ | | $E_{2,2}$ | | | |
| $M_3$ | | | $E_{3,3}$ | | |
| ... | | | | ... | |
| $M_m$ | | | | | $E_{m,n}$ |

Matriz léxica: F1 y F2 son sinónimos; F2 es polisémico

**Figura 2.** *Matriz léxica WordNet (adaptado de [24])*

### 2.3 Perfil de Usuario

La información del usuario, en especial el perfil, es usado junto con los consultas realizadas por los usuarios para realizar un proceso de RI personalizado. Esta personalización busca estimar de una mejor manera las necesidades de los usuarios y seleccionar el conjunto de documentos más relevantes a dichas necesidades [26]. En este proceso, la consulta describe la búsqueda actual del usuario, conocido como su interés local [13], mientras que el perfil del usuario describe las preferencias del usuario sobre un largo periodo de tiempo, conocido como su interés global. Dependiendo de la forma en que los interés globales afectan los locales, las operaciones de consulta se clasifican dentro de dos operaciones: expansión de consulta y reponderación de términos [1]. Un sistema puede tener una combinación de las dos técnicas, cambiando los pesos de los términos (teniendo en cuenta, inclusive la re-alimentación que hacen los usuarios en los resultados de consultas previas) y adicionando nuevos términos a la consulta (expansión).

La expansión de la consulta es usada frecuentemente en meta-buscadores personalizados. El meta-buscador adiciona a las consultas de los usuarios, los términos o componentes del perfil del usuario y envía la consulta extendida a cada motor de búsqueda [27]. La re-alimentación (una página es relevante o no) o *feedback* del usuario [28, 29] también puede ser usada para expandir la consulta y re-ponderar los términos, en fórmulas muy conocidas como la de Rochio [1].

## 3. MODELO PROPUESTO

El modelo de meta-buscador semántico de documentos web propuesto, incorpora una taxonomía general de conocimiento (TGC), una ontología de dominio general (WordNet), ontologías de dominio específico y perfil de usuario para mejorar la relevancia (basada en precisión) de los resultados entregados a un usuario en un sistema de búsqueda web. El modelo es desarrollado en la Universidad Mariana de la Ciudad de San Juan de Pasto, Nariño, Colombia y en la Universidad del Cauca, Popayán, Cauca, Colombia, y se basa en un modelo compuesto por cinco (5) módulos, a saber: Módulo de expansión de consulta, módulo de consulta remota bilingüe, módulo de indexación y filtrado bilingüe, módulo de ordenamiento y filtrado, y módulo de perfil de usuario.

### 3.1 Módulo de Expansión de Consulta

Este módulo integra una taxonomía general de conocimiento (TGC), como una estructura organizada de forma jerárquica que representa el conocimiento humano en múltiples lenguajes (en este trabajo en español e inglés). Cada nodo o rama de la TGC enlaza ontologías de dominio específico, que describen en detalle los conceptos y las relaciones de dichos conceptos en un dominio particular del conocimiento humano. Además se usa una ontología de dominio general como por ejemplo WordNet para encontrar los sinónimos e hiperónimos más frecuentes de los términos que el usuario ha digitado en la consulta y con ellos realizar un proceso de expansión de consulta que tenga en cuenta la semántica de los términos digitados por el usuario (ver Fig 3). A continuación se describen los componentes y procesos principales de este módulo:

- **Interfaz web.** El modelo provee una interfaz Web para acceso a los usuarios. Esta interfaz permite conexiones mediante cualquier tipo de navegador Web y tiene como objetivo principal, soportar el ingreso al sistema (login) y un proceso asistido e interactivo de formulación de la consulta. Quiere esto decir que el usuario digita una consulta y el sistema le permite (a través de una lista de auto completar) ir complementando los términos de búsqueda.

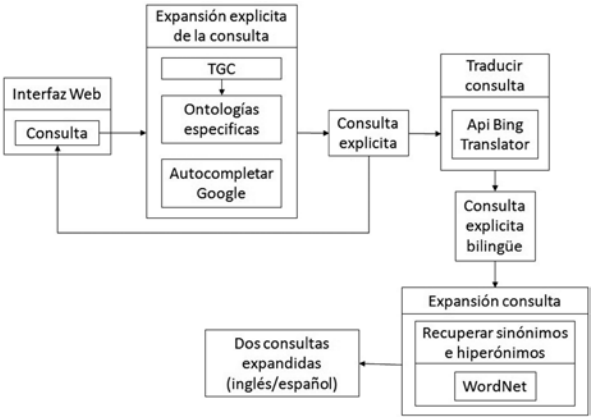


**Figura 3.** *Módulo de Expansión de Consulta*

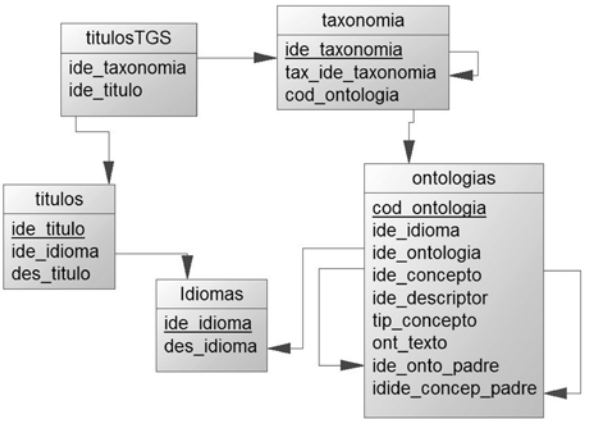


**Figura 4.** *Modelo relacional de taxonomías y ontologías*

- **Expansión explicita de la consulta.** Como se menciono anteriormente, el usuario puede usar una opción de autocompletar cuando esta digitando la consulta. En esta acción se recuperan los conceptos de la taxonomía general de conocimiento, que vincula conceptos desde lo general a lo concreto y las ontologías de dominio específico que definen conceptos, relaciones, funciones, instancias y axiomas de un dominio en particular. Para este modelo en las ontologías se tienen en cuenta sólo: Conceptos o ideas básicas que se intentan formalizar; Relaciones que representan la interacción y enlace entre los conceptos del dominio (Por ejemplo: subclase-de, parte-de, parte-exhaustiva-de, conectado-a, entre otros) e instancias que representan objetos determinados de un concepto el cual puede estar en los idiomas trabajados (inglés o español). En este proceso, si el sistema no encuentra información relacionada con los intereses del usuario (paso anterior), acude a un servicio de autocompletar externo, como por ejemplo el de Google (basado en análisis de registros de consultas de sus usuarios, un enfoque centrado en filtrado colaborativo).

  Con el fin de aumentar el rendimiento en las consultas, la estructura de representación de conocimiento se migró a un modelo relacional, debido a que el acceso a los documentos o textos OWL (Ontology Web Language) que se usan para almacenar las ontologías es muy costoso en tiempo y procesamiento. La Fig 4 muestra dicho modelo.

- **Traducir consulta.** Se encarga de tomar la cadena de consulta explícita y detectar el idioma en que está escrita para realizar la traducción al segundo idioma (inglés o español) apoyado en la Api *Bing Translator*, con el fin de realizar la consulta de forma bilingüe.

- **Expansión consulta** (tácita u oculta para el usuario). La primera tarea que realiza consiste en **Recuperar sinónimos e hiperónimos de WordNet.** Para ello, elimina palabras vacías (stop words) de la consulta, caracteres especiales, acentos, convertir la cadena a minúsculas, con el resultado de estas tareas se toman cada uno de los términos de la consulta y se examina la ontología para retornar los sinónimos e hiperónimos con los que se realiza la expansión de la consulta. Para la elección de los sinónimos se toma cada término $\{T_i.....T_n\}$ de la consulta, con el cual se examina la relación de sinonimia que este pudiera tener, si tiene estas relaciones, se recorre el árbol de relación semántica de WordNet y se retornan las dos primeras unidades que aparecen en los niveles superiores (quienes constituyen la relación de sinonimia más común) que representan a $T_i$, además de cada una de estas unidades léxicas retornadas (sinónimos) se retorna el elemento de primer nivel que aparece en las relaciones de hiperonimia (hiperónimos más comunes) los cuales están vinculados directamente con cada sinónimo, lo anterior con el fin de aportar sentido semántico a los términos que forman la consulta del usuario (ver Fig 5).

Es decir, del vector de términos que conforman la consulta $Q=\{T_1,T_2,.....T_n\}$ posteriormente se toma cada uno de los términos del vector de consulta y se forman los conceptos de tal forma que $C_i=(T_i+(S_{1,i}+H_{1,i})+(S_{2,i}+H_{2,i}))$. Cada concepto es igual al término digitado por el usuario y los términos semánticamente relacionados que fueron recuperados de la ontología. Es así como la consulta expandida queda como $Q^E=\{T_1, T_2,\ldots T_n, S_{1,1}, H_{1,1}, S_{2,1}, H_{2,1}, S_{1,2}, H_{1,2}, S_{2,2}, H_{2,2},\ldots S_{1,n}, H_{1,n}, S_{2,n}, H_{2,n}\}$.

Con los conceptos de la consulta expandida se examina la base de datos de conceptos del usuario, donde se recupera los conceptos de mayor ponderación, teniendo en cuenta la importancia relativa del concepto en la colección (historial de conceptos de usuario).

Finalmente, se toman los resultados de la expansión de consulta formado dos cadenas, una para el idioma inglés y otra para español así: $Q^E_{español}=\{T_{1\text{-}ES}, T_{2\text{-}ES},\ldots T_{n\text{-}ES}, S_{1,1\text{-}ES}, H_{1,1\text{-}ES}, S_{2,1\text{-}ES}, H_{2,1\text{-}ES}, S_{1,2\text{-}ES}, H_{1,2\text{-}ES}, S_{2,2\text{-}ES}, H_{2,2\text{-}ES},\ldots S_{1,n\text{-}ES}, H_{1,n\text{-}ES}, S_{2,n\text{-}ES}, H_{2,n\text{-}ES}\}$ y $Q^E_{ingles}=\{T_{1\text{-}EN}, T_{2\text{-}EN},\ldots T_{n\text{-}EN}, S_{1,1\text{-}EN}, H_{1,1\text{-}EN}, S_{2,1\text{-}EN}, H_{2,1\text{-}EN}, S_{1,2\text{-}EN}, H_{1,2\text{-}EN}, S_{2,2\text{-}EN}, H_{2,2\text{-}EN},\ldots S_{1,n\text{-}EN}, H_{1,n\text{-}EN}, S_{2,n\text{-}EN}, H_{2,n\text{-}EN}\}$

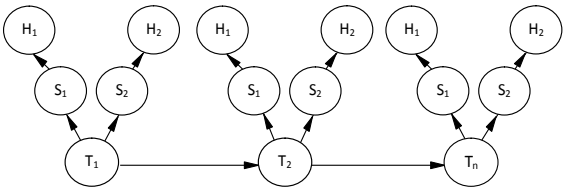


**Figura 5.** *Árbol de relación semántica de sinonimia e hiperonimia para cada término*

### 3.2 Módulo de Consulta Remota Bilingüe

Este módulo tiene dos tareas fundamentales (ver Figura 6), a saber:

- **Procesar cadenas de consulta.** Encargado de procesar la cadena de consulta antes de enviarla a las fuentes primarias de búsqueda en este caso los buscadores (Google, Yahoo! y Bing), dando el formato de cadena de búsqueda específica de cada buscador (API, Interfaz de Programación de Aplicaciones).

- **Búsqueda en motores tradicionales.** Con las cadenas de consulta formateadas apropiadamente, este paso se encarga de hacer los llamados a los buscadores web tradicionales, en este caso Google, Yahoo! y Bing. Este llamado se realiza en forma asincrónica, basado en hilos y con un punto en común de sincronización. Los resultados de las tres fuentes en los dos idiomas se almacenan en una estructura única denominada, documentos recuperados.

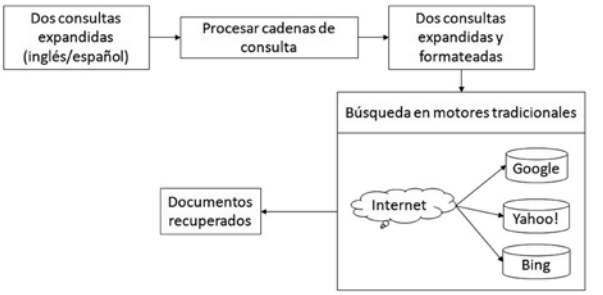


**Figura 6.** *Módulo de Consulta Remota Bilingüe*

### 3.3 Módulo de Indexación y Filtrado Bilingüe

Este módulo toma los resultados de las fuentes principales de documentos (documentos recuperados), lee los snippets (textos resumen que elaboran los buscadores web de cada uno de los recursos registrados en sus bases de datos) de cada URL retornada y la indexa en memoria (ver Figura 7).

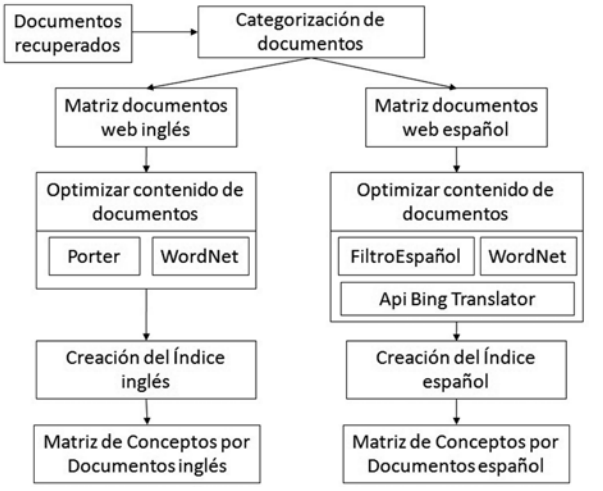


**Figura 7.** *Módulo de Indexación y Filtrado Bilingüe*

Las tareas que realiza este módulo son las siguientes:

- **Categorización de documentos.** Toma los documentos recuperados por las fuentes de búsqueda primarias (Google, Yahoo!, Bing) y los clasifica en dos conjuntos dependiendo del idioma (uno para inglés y otro para español).

- **Matriz de documentos web** (inglés, español). Encargado de tomar los resultados en los formatos retornados por los buscadores y transformarlos al estándar de la aplicación. En el proceso de creación de la matriz se eliminan los documentos (URLs) repetidos o duplicados (debido a que algunos de estos documentos son recuperados por más de un buscador), las palabras vacías (stop words), caracteres especiales, etiquetas HTML, acentos y saltos de línea. La matriz se forma con los campos IdDocumento, Título, Contenido (o snnipet), ContenidoStemming (snippet pre-procesado), Url, BuscadorOrigen, en este caso se toman aproximadamente sesenta (60) documentos por cada buscador (este parámetro se puede ajustar en el modelo), obteniendo un resultado de documentos únicos que suma en promedio 120, lo que muestra que existe un promedio de 60 resultados (33%) compartidos por los motores de los 180 posibles resultados. La Figura 8 muestra la forma de la matriz de documentos Web (inglés, español).

| idr | iddcument | title | content | contenStemming | url | source browser |
|---|---|---|---|---|---|---|
| 1 | 1 | title..1 | content..1 | contenStemm..1 | url..1 | google |
| 1 | 2 | title..2 | content..2 | contenStemm..2 | url..2 | google, yahoo |
| 1 | 3 | title..3 | content..3 | contenStemm..3 | url..3 | bing |
| 1 | 4 | title..4 | content..4 | contenStemm..4 | url..4 | google,bing |
| 1 | 5 | title..5 | content..5 | contenStemm..5 | url..5 | yahoo,bing |
| 1 | 6 | title..6 | content..6 | contenStemm..6 | url..6 | google |
| 1 | 7 | title..7 | content..7 | contenStemm..7 | url..7 | yahoo,bing |
| 1 | 8 | title..8 | content..8 | contenStemm..8 | url..8 | yahoo |

**Figura 8.** *Matriz de documentos web*

- **Optimizar contenido de documentos.** Analiza cada uno de los documentos, buscando la aparición de los conceptos representados por una listas de términos (Sinónimos e Hiperónimos) que se encuentran en el contenido del documento y los remplaza por el concepto original que aparece en la consulta con el propósito de aumentar la ponderación del concepto en el documento cuando se cree la matriz de conceptos por documentos que se utiliza en el modelo vectorial. Inicialmente se toma un documento d como una lista de términos así: $d=\{T_1,T_2,...T_D\}$ y se obtiene un documento como una lista de conceptos que relacionan varios términos, apoyado nuevamente en sinónimos e hiperónimos de WordNet, igual que en el proceso de la expansión de la consulta. Es así como al final se obtiene un documento $d= \{C_1,C_2,....C_M\}$ donde M es el número de conceptos y $M<D$. Este proceso se realiza tanto para los documentos en inglés como en español. Para los documentos en español se usa además la API de traducción de Bing, debido a que WordNet sólo contempla términos en inglés. Para no usar ningún servicio de traducción externa de términos, se puede incluir EuroWordNet o MultiWordNet que son ontologías de dominio general multilenguaje, dentro de los que se encuentra el inglés y el español.

- **Creación del índice.** Encargado de crear los índices de los documentos en memoria. En estas estructuras se indexan los documentos uno a uno haciendo uso del FiltroEspañol, que se basa en el algoritmo de porter (para más detalles ver [30]) y para hacer *stemming* (reducción de términos a su raíz léxica) a los documentos en español y el algoritmo de porter para los documentos en inglés. Con lo anterior, se convierte cada uno de los términos del documento a su raíz léxica (por ejemplo "Corredor" en "Correr" o fishing", "fished", en "fish"), estos analizan cada uno de los documentos de la matriz, creando un documento filtrado, considerando que un documento está bien filtrado cuando su longitud mínima de conceptos o palabras en el contenido del documento es mayor que 2, si esto se cumple se agrega el documento al índice con los campos IdDocumento, Título, Contenido, ContenidoStemming, Url, BuscadorOrigen, con esto el índice queda listo para realizar el filtrado y ordenamiento. La Figura 9 muestra el resultado de esta fase, una matriz de conceptos por documentos en cada uno de los idiomas (español, inglés), similar a la matriz de términos por documentos del modelo vectorial [4] de representación de documentos en RI. Donde cada celda refleja la importancia del concepto en su raíz léxica contra los documentos basado en la fórmula (1) propuesta por Salton [1, 31], común en el modelo vectorial de representación de documentos en recuperación de información, donde $F_{i,j}$ es la frecuencia observada del concepto j en el documento i. $Max(F_i)$ es la mayor frecuencia observada en el documento i. N es el número de documentos en la colección y $n_j$ es el número de documentos en los que aparece el concepto j.

$$w_{i,j} = \frac{F_{i,j}}{\max(F_i)} \times \log\left(\frac{N}{n_j + 1}\right) \quad \textbf{(1)}$$


**Figura 9.** *Matriz de conceptos por documentos*

### 3.4 Módulo de Ordenamiento y Filtrado

Este módulo ordena y filtra los resultados basado en la distancia de cosenos (comúnmente usada en el modelo vectorial de recuperación de información), muestra en forma ordenada en dos pestañas (una para inglés y otra para español) los resultados que superan un umbral de similitud entre el documento y la consulta. La Figura 10 muestra un resumen de las tareas que realiza este módulo; tareas que son explicadas a continuación.

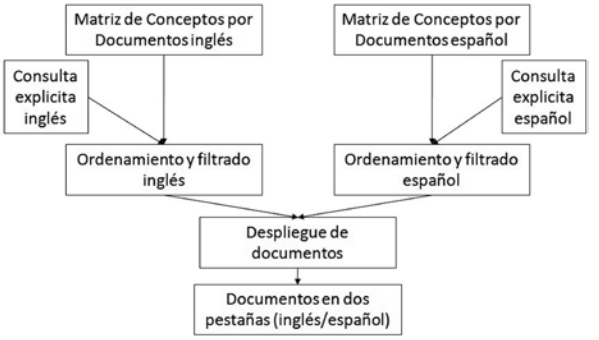


**Figura 10.** *Módulo de Ordenamiento y Filtrado*

- **Ordenamiento y filtrado inglés/español.** Haciendo uso del FiltroEspañol y el algoritmo de Porter sobre la cadena de consulta del usuario (definida como cadena explícita en el primer módulo), se genera una cadena de consulta con los términos reducidos a su raíz léxica, tanto en inglés como en español. Con esta cadena reducida se ejecuta la consulta sobre la matriz de conceptos por documentos obteniendo una lista ordenada dependiendo de la similitud del contenido de estos con la consulta digitada por el usuario. La similitud se establece basado en la distancia de cosenos, según la fórmula (2) tanto para la matriz en ingles como para la matriz en español. En el sistema se puede definir un parámetro de similitud mínima que varía de 0 a 1, donde 1 corresponde a totalmente igual y 0 a totalmente diferente, para que el sistema filtre los documentos que estén por debajo de este parámetro o umbral.

$$Cos(d,q) = \frac{\sum_{i=1}^{n}(d_i * q_i)}{\sqrt{\sum_{i=1}^{n} d_i^{\,2} * \sum_{i=1}^{n} q_i^{\,2}}} \quad \textbf{(2)}$$

- **Despliegue de documentos.** Se listan los documentos en dos pestañas (resultados en inglés y resultados en español) para que el usuario los visualice. Los resultados se listan en orden de acuerdo a la similitud (más similares a menos similares) que presentan con respecto a la consulta realizada por el usuario.

### 3.5 Módulo de Perfil de Usuario

Este módulo hace una representación a largo plazo de las necesidades del usuario, para ello se basa en el contenido de los documentos y se establece como una estructura de grano fino determinado por las palabras clave (términos) que se encuentran en los documentos recuperados. Con base en la calificación que el usuario hace sobre los documentos que le han sido mostrados y que son relevantes para cada consulta, el sistema toma las palabras encontradas en los documentos, las pondera (con un peso dependiendo de la frecuencia de la palabra en cada documento) y almacena en una base de datos (formando el perfil) propia de cada usuario para emplearlas en futuras búsquedas y reordenar los resultados con base en las características propias del perfil (ver Figura 11).

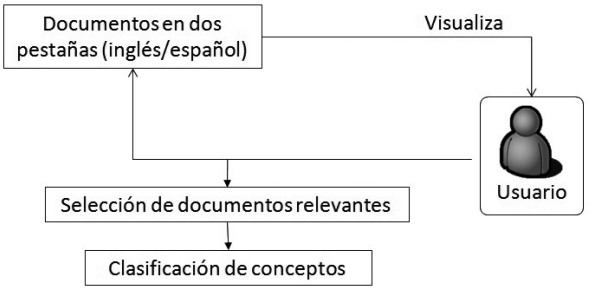


**Figura 11.** *Módulo de Perfil de Usuario*

Las tareas desarrolladas por este módulo son:

- **Selección de documentos relevantes.** Cuando el usuario visualiza los documentos, tiene la opción de marcar las páginas (documentos) que considere

más relevante a la búsqueda que previamente definió, de las cuales se sacan los conceptos que se emplearon en la búsqueda y se ponderan con la media de frecuencia de aparición del término en el documento (denominado frecuencia de término en el documento o TF por sus siglas en inglés). La Figura 12 muestra la relación de los conceptos con los intereses del usuario que realiza la consulta.

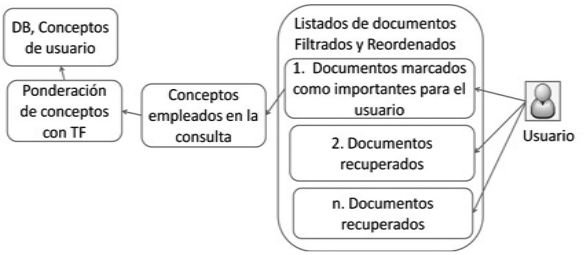


**Figura 12.** *Relación de conceptos en la consulta para formar el perfil de usuario.*

- **Clasificación de conceptos.** Se toma los conceptos de mayor ponderación, teniendo en cuenta la importancia relativa del concepto en el documento marcado como importante para el usuario. Cada uno de estos conceptos y su ponderación (valor de TF) se guardan en la base de datos personal del usuario, acompañados de la url, el snipet o texto resumen y el título del documento con lo que se forma el perfil. Este perfil es un historial de conceptos, que serán utilizados en las consultas futuras que realice el usuario.

  Este historial es utilizado cuando el usuario (que ya ha marcado documentos en consultas anteriores como relevantes) realiza una nueva consulta, en este proceso se toman los conceptos de la consulta digitada y se comparan con los que se encuentran en el historial (base de datos de perfil de usuario), si existen conceptos que están en un mismo contexto (tratan de la misma temática) se recuperan los documentos que contienen dichos conceptos. Estos documentos son desplegados en los primeros lugares de la lista debido a la importancia representativa de estos para el usuario.

### 3.6 Metodología del proyecto

A continuación se describe un modelo y una herramienta (meta-buscador) para validar dicho modelo, denominada XGhobi, Para obtener estos productos se orientó un proceso de investigación basado en la propuesta de Nunamaker et al (1991), denominada *Systems Development Method* (SDM) para el desarrollo de sistemas de información en investigación científica [32] atendiendo los siguientes pasos:

- **Concepto de diseño.** En esta etapa se realizó una documentación detallada del tema de interés, en este caso recuperación de información, indexación, pre procesamiento, algoritmos de stemming, meta-buscadores, lo que permitió identificar limitaciones y deficiencias en los sistemas existentes e identificar el objetivo central de investigación. En esta etapa se definió el modelo inicial del meta-buscador que fue refinado en cada una de las etapas posteriores de la investigación, vinculando de esta forma nuevas características y/o mejoras a las características previamente definidas.

- **Construcción de la arquitectura del sistema.** Para esta etapa se tomaron los conocimientos adquiridos en la etapa previa, conocimientos relacionados con patrones de arquitectura, en especial el desarrollo de sistemas multi-nivel o en capas, y con ellos se creó la arquitectura inicial del modelo. Este paso de SDM se baso en las actividades, tareas y productos más importantes del Proceso Unificado (UP).

- **Prototipos.** En esta etapa se desarrolló el sistema con base en los resultados de las etapas anteriores, aplicando análisis, diseño, implementación y pruebas en forma iterativo e incremental conforme a UP, incluyendo además revisiones con expertos y evaluaciones con usuarios, lo que brindó la posibilidad de mejorar las funcionales disponibles e incorporar nuevas opciones que fueron consideradas como necesarias por los expertos o los usuarios.

- **Desarrollo del producto.** En esta etapa se adopta el nuevo sistema producido como un producto que es usado en el ambiente de producción. La meta-principal de XGhobi es ahorrar tiempo en las búsquedas y posibilitar la apropiación más rápida de conocimiento por parte de los usuarios. A nivel del grupo de investigación, en este momento se reflexionó sobre el conocimiento y la experiencia adquirida en la creación, evaluación y uso de XGhobi. Con esto se definieron nuevos requisitos para la nueva versión de XGhobi, en el que se espera incluir el manejo de dos lenguajes (inglés y español), el manejo de una ontología de dominio general, entre otros.

## 4. XGHOBI

XGhobi cuenta con una interfaz web centrada en el usuario final, procurando incorporar los principales atributos que componen la Usabilidad. Atributos objetivos como facilidad de aprendizaje, facilidad de memorización, eficacia, eficiencia o tiempo empleado para completar una tarea, operabilidad, y facilidad de comprensión; y atributos subjetivos orientados a la satisfacción del usuario [33-35] como, Accesibilidad, Funcionalidad, Utilidad, Estética y Credibilidad.

XGhobi cuenta con una interfaz web sencilla y usable (ver la Figura 13). Esta interfaz cuenta con una caja de texto para capturar las palabras clave que conforman la consulta, en la que se encuentra incorporada la opción de autocompletar y un botón que despliega la interfaz de inicio de la búsqueda.


**Figura 13.** *Interfaz web de XGhobi*

La Figura 14 muestra la forma en que los resultados son desplegados una vez el meta-buscador ha realizado la recuperación, procesamiento, filtrado y ordenamiento de los documentos obtenidos. Esta interfaz despliega los documentos clasificados por idioma, cada uno en una pestaña. En cada pestaña se resalta el título de los documentos con un color de fácil identificación, el *snippet* o texto resumen se muestra de una manera clara, la dirección URL cuenta con un color que la resalta, el buscador o buscadores en donde se recuperó cada documento se muestra en la parte final entre corchetes y una imagen en forma de estrella que el usuario puede seleccionar para marcar los documentos que considera como relevantes y con ello gestionar el perfil de cada usuario.

XGhobi cuenta con una barra con información del número de documentos recuperados, número de documentos en la página actual, la consulta realizada por el usuario, una barra para la navegación entre páginas de resultados y la opción de reformulación de consulta. Finalmente, con el propósito de evaluar el proceso de recuperación de información del meta-buscador se incorporó en XGhobi una encuesta a los usuarios, acerca de la satisfacción de los resultados entregados en cada consulta.

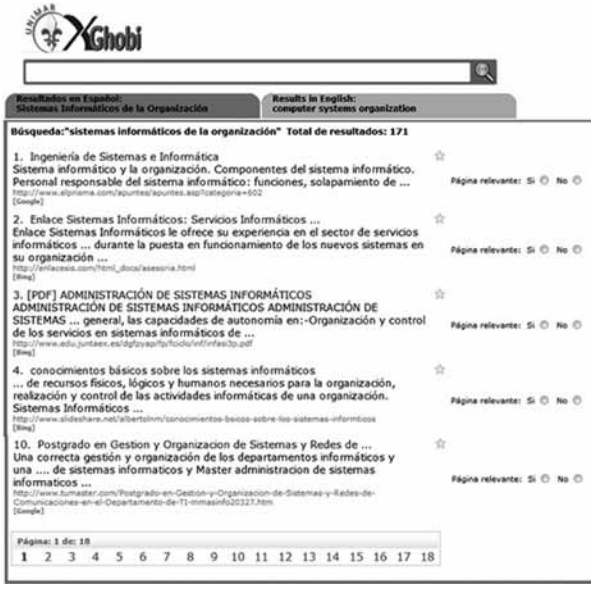


**Figura 14.** *Despliegue de resultados en XGhobi*

## 5. EVALUACIÓN DEL MODELO

La evaluación de un Sistema de Búsqueda Web involucra múltiples aspectos [36-38], entre los cuales se tienen en cuenta el rendimiento, la funcionalidad, la usabilidad, entre otros, pero uno de especial importancia es la satisfacción del usuario en cuanto a la relevancia de los resultados reportados por el sistema [39, 40].

Para muchas aplicaciones de RI, en especial en búsqueda web, a los usuarios lo que más les importa es que tan relevantes o útiles son los resultados en las primeras páginas (ya que normalmente ellos no revisan toda la lista de resultados). Esto motiva a que la evaluación involucre una medición de los valores de precisión en ciertos puntos de los resultados recuperados, como por ejemplo los primeros 10 o 20 documentos. Esta medida se conoce como la "precisión en K" (Precisión at K) y tiene la ventaja de no requerir ninguna estimación del conjunto total de resultados relevantes (factor clave en la evaluación de buscadores web); por lo anterior, esta medida fue usada para evaluar XGhobi.

El proceso de evaluación se realizó comparando los resultados del sistema propuesto con los entregados por buscadores tradicionales (Google, Yahoo!, Bing). La estrategia de evaluación se basó en una evaluación a ciegas, donde los usuarios evaluaron los resultados de las búsquedas sin saber cual sistema estaban utilizando. Para ello, se desarrolló una interfaz para cada sistema donde se elimino todo tipo de logo, imagen, icono o mensaje que hiciera alusión al nombre del sistema. Con

este proceso de evaluación se disminuye la posibilidad de sesgar los resultados que los usuarios pudieran reportar.

La evaluación incluyó 25 búsquedas de usuarios seleccionados de la siguiente forma, del personal docente se contó con 10 personas que en el momento son candidatos a magister en diferentes aéreas, 15 estudiantes de décimo semestre y 10 de noveno semestre todos vinculados al programa de ingeniería de sistemas de la Universidad Mariana de la Ciudad de San Juan de Pasto, en lo cual para cada grupo se evaluó en la sesión grupal un total de 50 consultas, evaluando las 10 primeros documentos de la primer página recuperada. La Figura 15 muestra como resultado, la precisión promedio en diferentes valores de K, con resultados comprendidos entre el 72% y 100% para XGhobi, mostrando de esta forma el alto grado de precisión del modelo propuesto, además en la gráfica se puede observar que los resultados de XGhobi siempre fueron mejores que los obtenidos con los buscadores tradicionales por separado.

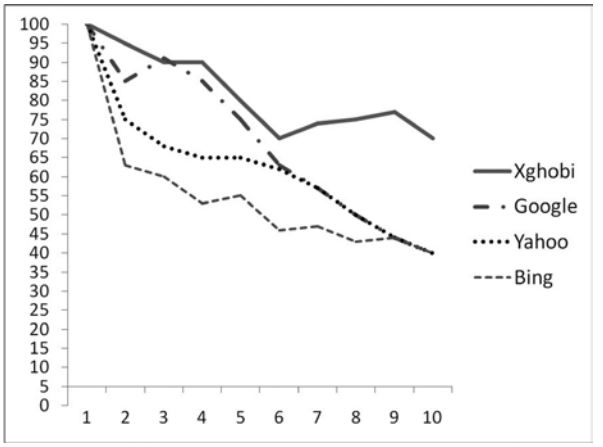


**Figura 15.** *Precisión en K resultados (1..10)*

El proceso de evaluación también incluyó la comparación de XGhobi con otros meta-buscadores, en este caso Excite (el más preciso según [41]) y Dogpile (el más popular según [9]). La evaluación se realizó con 30 usuarios entre estudiantes y docentes del programa de Ingeniería de Sistemas de la Universidad Mariana, aplicando las mismas consultas en ciencias de la Computación conforme a la evaluación anterior, adicionando consultas en los dos idiomas (español, inglés). Los resultados de esta evaluación se muestran a continuación.

La Figura 16 muestra gráficamente los resultados de la precisión promedio en diferentes valores de K, con resultados comprendidos entre el 100% y 94,02% para XGhobi. En la gráfica se puede observar que los resultados de XGhobi fueron siempre ligeramente mejores que los obtenidos con los otros dos meta-buscadores, para la mayoría de los valores de K.

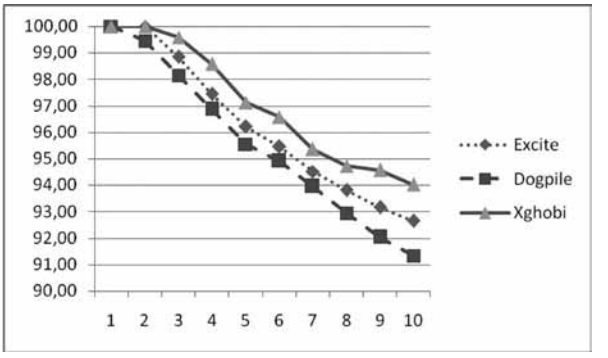


**Figura 16.** *Precisión de los meta-buscadores en K resultados*

El resultado general de la Precisión Media Promedio o Mean Average Precision fue de 97,58% para XGhobi, 96,21% para Excite y 95,52% para Dogpile. Este resultado muestra que en promedio, XGhobi fue un 1,37% mejor que Exite y 2,6% mejor que DogPile.

## 6. CONCLUSIONES Y TRABAJO FUTURO

Se diseñó, desarrolló y evaluó un modelo y un meta-buscador web que entrega resultados más relevantes que los obtenidos con los buscadores tradicionales Google, Yahoo! y Bing. Este meta-buscador además está diseñado para recuperar documentos y expandir las consultas de los usuarios en dos idiomas (inglés, español). El modelo y el meta-buscador integra sinérgicamente una taxonomía general de conocimiento, una ontología de dominio general, múltiples ontologías de dominio específico y un perfil del usuario con el objetivo de asistir el proceso de la expansión de la consulta, y filtrar y reordenar de una mejor forma los resultados presentados a los usuarios.

Para mejorar los resultados de la búsqueda, se incorporó en el meta-buscador una taxonomía de conocimiento general, varias ontologías específicas en esta versión la clasificación ontológica de ACM (ontologías

específicas en ciencias de la computación) y una ontología de domino general (WordNet), con el fin de expandir y aportar semántica a cada uno de los términos de la consulta digitada por el usuario, además de la ontología se adiciona la opción de perfil de usuario con el propósito de definir consultas más precisas a medida que el usuario utiliza el meta-buscador.

Como trabajo futuro se espera incorporar en el modelo y en XGhobi la propuesta de Robertson and Sparck-Jones [42], para evaluar los conceptos relevantes y no relevantes en el proceso de expansión de consulta y filtrado de los resultados recuperados de los buscadores tradicionales. También se espera desarrollar un experimento que permita definir el impacto a largo plazo del perfil del usuario en la precisión de los resultados reportados por XGhobi a lo largo del tiempo. Finalmente, revisar otras medidas de evaluación y comparación de los resultados entregados por el meta-buscador, frente a los buscadores tradicionales.

## 7. AGRADECIMIENTOS

Nuestros agradecimientos a directivos de la facultad de ingeniería de la Universidad Mariana, al Grupo de I+D en Tecnologías de la Información (GTI) de la Universidad del Cauca y al Laboratorio de Investigación en Sistemas Inteligentes (LISI) de la Universidad Nacional de Colombia sede Bogotá.

## 8. BIBLIOGRAFÍA

[1] Baeza-Yates, R., A. and B. Ribeiro-Neto, *Modern Information Retrieval.* 1999: Addison-Wesley Longman Publishing Co., Inc. 513.

[2] Manning, C., P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval.* 2007, Cambridge University Press: Cambridge, England.

[3] Liaw, S.-S. and H.-M. Huang, *Information retrieval from the World Wide Web: a user-focused approach based on individual experience with search engines.* Computers in Human Behavior, 2006. 22(3): p. 501-517.

[4] Massimo, M., *A basis for information retrieval in context.* ACM Trans. Inf. Syst., 2008. 26(3): p. 1-41.

[5] Manning, C., P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* 2008, Cambridge University Press: Cambridge, England.

[6] Eui-Hong, H., et al., *Intelligent metasearch engine for knowledge management, in Proceedings of the twelfth international conference on Information and knowledge management %@ 1-58113-723-0.* 2003, ACM: New Orleans, LA, USA. p. 492-495.

[7] Mustafa, J., S. Khan, and K. Latif. *Ontology based semantic information retrieval.* in *Intelligent Systems, 2008. IS '08. 4th International IEEE Conference. 2008.*

[8] Susan, G., S. Mirco, and P. Alexander, *Ontology-Based User Profiles for Search and Browsing, in Ontologies,* S. US, Editor. 2007. p. 665-694.

[9] Karatzoglou, A. and I. Feinerer. *Text Clustering with String Kernels in {R}. in Advances in Data Analysis (Proceedings of the 30th Annual Conference of the Gesellschaft f{Ã¼}r Klassifikation e.V., Freie Universit{Ã¤}t Berlin, March 8--10, 2006).* 2007: Springer-Verlag.

[10] Etsioni, E.S.a.O. *Multi-service search and comparison using the MetaCrawler. in 4th International World Wide Web Conference.* 1995.

[11] Dogpile.com. *Different Engines, Different Results: Web Searchers Not Always Finding What They're Looking for Online.* 2007; Available from: http://www.infospaceinc.com/onlineprod/Overlap-DifferentEnginesDifferentResults.pdf.

[12] Carpineto, C., et al., *A survey of Web clustering engines.* ACM Comput. Surv., 2009. 41(3): p. 1-38.

[13] Barry, C.L., *User-Defined Relevance Criteria: An Exploratory Study.* Journal of the American Society for Information Science-A, 1994. 45(1): p. 149-159.

[14] Huang, A., et al. *Clustering Documents with Active Learning Using Wikipedia. in Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on.* 2008.

[15] Li, X. *Research on Text Clustering Algorithm Based on K_means and SOM. in Intelligent Information Technology Application Workshops, 2008. IITAW '08. International Symposium on.* 2008.

[16] Mao-Ting, G. and W. Zheng-Ou. *A New Algorithm for Text Clustering Based on Projection Pursuit. in Machine Learning and Cybernetics, 2007 International Conference on.* 2007.

[17] Fuzhi, Z., et al. *An Ant-Based Fast Text Clustering Approach Using Pheromone. in Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on.* 2008.

[18] Guo, Q.-l. and M. Zhang, *Semantic information integration and question answering based on pervasive agent ontology.* Expert Systems with Applications, 2009. 36: p. 10.

[19] Anil, K.J., *Data Clustering: 50 Years Beyond K-means, in Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I.* 2008, Springer-Verlag: Antwerp, Belgium.

[20] Jing, L., *Survey of Text Clustering.* 2008.

[21] Song, J.-f., et al., *Ontology-Based Information Retrieval Model for the Semantic Web, in Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05) on e-Technology, e-Commerce and e-Service.* 2005, IEEE Computer Society.

[22] Aufaure, M.A., R. Soussi, and H. Baazaoui. SIRO: *On-line semantic information retrieval using ontologies. in Digital Information Management, 2007. ICDIM '07. 2nd International Conference on.* 2007.

[23] Giannis, V., et al., *Semantic similarity methods in wordNet and their application to information retrieval on the web, in Proceedings of the 7th annual ACM international workshop on Web information and data management.* 2005, ACM: Bremen, Germany.

[24] Beck, H.W., T. Anwar, and S.B. Navathe, *A conceptual clustering algorithm for database schema design.* Knowledge and Data Engineering, IEEE Transactions on, 1994. 6(3): p. 396-411.

[25] Song, W., C.H. Li, and S.C. Park, *Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures.* Expert Systems with Applications, 2009. 36(5): p. 9095-9104.

[26] Bhatia, S.K. and J.S. Deogun, *Conceptual clustering in information retrieval.* Systems, Man, and Cybernetics, Part B, IEEE Transactions on, 1998. 28(3): p. 427-436.

[27] Liu, H. and H. Motoda, *Computational Methods of Feature Selection.* 2007: Chapman & Hall/ CRC.

[28] Salton, G. and C. Buckley, *Improving retrieval performance by relevance feedback.* Journal of the American Society for Information, 1999. 41(4): p. 288 - 297.

[29] Rich, E., *User modeling via stereotypes.* 1979: p. 329-354.

[30] Ordoñez, H. and C. Cobos. *Ghobi – Un Meta Buscador Web Optimizado Para Búsquedas En Español. in Quinto Congreso Colombiano de Computación.* 2010. Cartagena, Colombia.

[31] Salton, G. and C. Buckley, *Term-weighting approaches in automatic text retrieval.* Information Processing & Management, 1988. 24(5): p. 513-523.

[32] Song, W. and S.C. Park, *Genetic algorithm for text clustering based on latent semantic indexing.* Computers & Mathematics with Applications, 2009. 57(11-12): p. 1901-1907.

[33] Giugni O., M. and R. Loaiza B., *Metodología para el desarrollo de portales centrada en el usuario: una evaluación empírica.* Revista electrónica de estudios telemáticos, 2008. 7(3): p. 17.

[34] Fisher, D.H., *Knowledge acquisition via incremental conceptual clustering.* Machine Learning, 1987. 2(2): p. 139-172.

[35] Montero, Y.H., *Factores del Diseño Web Orientado a la Satisfacción y No-Frustración de Uso.* Revista Española de Documentación Científica, 2006: p. 239-257.

[36] Martínez, F., *Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet, in Información y Documentación.* 2002, Universidad de Murcia: Murcia, España. p. 283.

[37] Cacheda, F., V. Formoso, and V. Carneiro, *Performance Analysis of Distributed Web Information Retrieval Systems.* Latin America Transactions, IEEE (Revista IEEE America Latina), 2007. 5(6): p. 479-485.

[38] Can, F., R. Nuray, and A.B. Sevdik, *Automatic performance evaluation of Web search engines.* Information Processing & Management, 2004. 40(3): p. 495-514.

[39] Chen, S., D. Alahakoon, and M. Indrawan. *Building an Adaptive Hierarchy of Clusters for Text Data. in Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on.* 2005.

[40] Zhao, L., et al. *An improved measuring similarity for short text snippets and its application in clustering search engine. in Machine Learning and Cybernetics, 2008 International Conference on.* 2008.

[41] Forsati, R., et al. *Hybridization of K-Means and Harmony Search Methods for Web Page Clustering. in Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on.* 2008.

[42] Garcia, E. *RSJ-PM Tutorial: A Tutorial on the Robertson-Sparck Jones Probabilistic Model for Information Retrieval.* 2009; Available from: http://www.miislita.com/information-retrieval-tutorial/information-retrieval-probabilistic-model-tutorial.pdf.