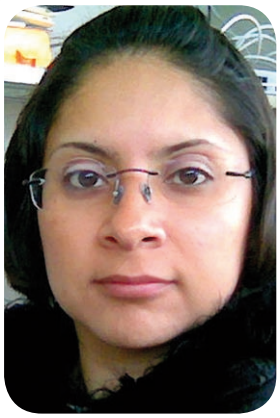
Uso de ontologías para la mejora de resultados de motores de búsqueda web

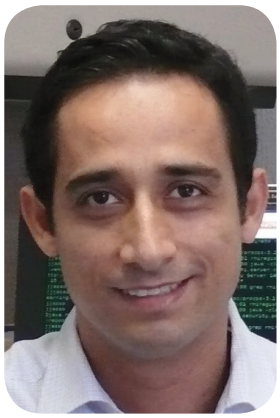
Por **Dulce Aguilar-Lopez, Ivan Lopez-Arevalo** y **Victor Sosa-Sosa**

Resumen: *Con el aumento del número de webs, el tiempo que un usuario invierte en la revisión de los resultados ofrecidos por los motores de búsqueda se incrementa de manera considerable. La naturaleza del contenido de estas páginas es semánticamente heterogénea y orientada al humano que sabe interpretarla correctamente. Es importante que el resultado de la búsqueda realmente corresponda a la información deseada. Una propuesta para lograrlo es comparar el contenido de la página web con el vocabulario formal del tema (ontología) y con el vocabulario informal (términos comunes del tema pero ajenos a la ontología). Se describe un tipo de búsqueda web que aprovecha las ontologías para reducir el espacio de búsqueda de ciertos temas. Con esta propuesta se mejora la relevancia de los resultados de los buscadores utilizando ontologías de dominio, el tesauro WordNet y una medida de similitud jerárquica. El aumento en la relevancia de los resultados se traduce en la disminución en el tiempo de revisión de los mismos.*



***Dulce Aguilar-López**, máster en ciencias de la computación en el Laboratorio de Tecnologías de Información del Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav-IPN) de México. Estudió ingeniería en sistemas computacionales en el Instituto Tecnológico de Pachuca (México). Sus áreas de interés son las ontologías y la recuperación de información.*

***Iván López-Arévalo** es investigador titular en el Laboratorio de Tecnologías de Información del Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav-IPN) de México. Doctor en computación por la Universitat Politècnica de Catalunya. Ha sido asistente de investigador en la Universitat Autònoma de Barcelona y asistente de profesor en la Universitat Rovira i Virgili. Sus áreas de interés son la minería de datos y la representación del conocimiento.*



***Víctor Sosa-Sosa** es investigador titular en el Laboratorio de Tecnologías de Información del Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional (Cinvestav-IPN) de México. Es doctor en ingeniería informática por la Universitat Politècnica de Catalunya y máster en ciencias computacionales por el Cenidet, México. Sus áreas de interés son los sistemas de información distribuida, en particular tecnologías web, bases de datos y minería de datos.*

Palabras clave: *Ontologías, Búsqueda semántica, WordNet.*

Title: Use of ontologies to enhance the results of web search engines

Abstract: *With the increasing number of web sites, the time spent by users reviewing the results also increases. In addition, the nature of web content is semantically heterogeneous and oriented to people who will be able to understand it. Frequently the results from search engines do not correspond to the expected topic. One approach to improve the results is to match the content of the web pages with a formal vocabulary on the topic (ontology) and with the informal vocabulary (common terms of the topic but not in the ontology). This paper describes a web search method that takes advantage of ontologies to reduce the search area of certain topics. With this approach the relevance of search engine results is enhanced by filtering the content through the integration of domain ontologies, the WordNet thesaurus, and a hierarchical similarity measure. Thus, the improvement on the relevance of results reduces the time required to review such results.*

Keywords: *Ontologies, Semantic search, WordNet.*

Aguilar-Lopez, Dulce; **Lopez-Arevalo, Ivan**; **Sosa-Sosa, Victor**. "Uso de ontologías para la mejora de resultados de motores de búsqueda web". *El profesional de la información*, 2009, enero-febrero, v. 18, n. 1, pp. 34-40.

DOI: 10.3145/epi.2009.ene.05

1. Introducción

La mayoría de los buscadores utilizan mecanismos que revisan sólo el contenido formal de las páginas web para saber si contienen o no las palabras clave que se les indican. Las técnicas que emplean están basadas en el número de veces que los usuarios acceden a esas páginas, el número de enlaces que tienen o la posición

Artículo recibido el 25-09-08
Aceptación definitiva: 25-11-08

que ocupan. En general, los motores de búsqueda tradicionales no prestan atención al contenido semántico. Esta es la causa de que a veces devuelvan resultados poco relevantes para el tema que se está buscando, además de que el análisis de los mismos supone una gran pérdida de tiempo. Una propuesta para abordar este problema es convertir la información en conocimiento por medio de ontologías.

La definición comúnmente aceptada de ontología es la propuesta por **Gruber**[1]: una especificación explícita de una conceptualización. Más concretamente, una ontología es un vocabulario común para personas y aplicaciones en un área determinada, independientemente del comportamiento y dominio de la aplicación que las use[2].

El objetivo de este trabajo es obtener un método de búsqueda de páginas web utilizando ontologías de dominio, por medio del cual los resultados realmente corresponderán al dominio elegido. Se toma como premisa que las ontologías utilizadas están bien construidas y están escritas en inglés. La propuesta además incluye al tesauro *WordNet*[3], los buscadores *Excite*[4], *Google*[5], *HotBot*[6], *Metacrawler*[7], *MSN*[8] y una medida de similitud jerárquica[9], tal como hemos mostrado en trabajos anteriores[10, 11].

2. Trabajo previo

Entre los sistemas más recientes para realizar una búsqueda semántica mediante ontologías se encuentra la propuesta por **Bocio** et al.[12], quienes utilizaron ontologías de dominio y algunos parámetros que el usuario debía indicar (nombre del buscador, máximo número de páginas resultantes, lenguaje, etc.). **Gao** et al.[13] usaron como estructura semántica una ontología y el algoritmo de "cruce de vectores de pesos" para analizar la información inicial y almacenar en un conjunto de conceptos los resultados preliminares. Dicha propuesta construía un vector de pesos acorde con la influencia del conjunto de conceptos dentro de la ontología. Por su parte, **Ramachandran** et al.[14] elaboraron una herramienta que utiliza la ontología del proyecto *Linked Environments for Atmospheric Discovery* (*LEAD*)[15], y contiene conceptos de ciencias atmosféricas además de definiciones y relaciones entre fenómenos atmosféricos, datos y servicios. El uso de esta ontología extiende las capacidades de búsqueda a un catálogo de metadatos y recursos web. **Sánchez-Ruenes**[16] implementó una aplicación para construir ontologías a partir de un dominio usando las palabras clave que obtenía de la Web para dicho dominio. Posteriormente los conceptos de la ontología se usaban para realizar nuevas búsquedas y así seleccionar las páginas web que pertenecían al dominio.

"Como punto de partida se requiere la palabra clave a buscar, así como el dominio en el que debe realizarse la búsqueda"

3. Propuesta ontológica

La propuesta que presentamos utiliza ontologías de dominio, el tesauro *WordNet*, cinco motores de búsqueda y una medida de similitud jerárquica con cosenos para obtener de la Web los resultados que realmente sean de utilidad para el usuario al realizar sus búsquedas dentro de un dominio dado.

El enfoque sigue la arquitectura mostrada en la figura 1, que se compone de los módulos descritos a continuación:

Entrada de datos

Como punto de partida se requiere la palabra clave a buscar, así como el dominio en el que debe realizarse la búsqueda, predefinido en una colección de ontologías de dominios específicos (libres y en lenguaje *owl-ontology web language* y *rdf-resource description framework*).

Búsqueda

Se realiza la búsqueda de la palabra clave deseada usando los buscadores indicados: *Excite*, *Google*, *HotBot*, *Metacrawler* y *MSN*.

Filtrado

Se construye una lista única con las páginas resultantes de cada buscador considerado, eliminando los duplicados y/o no disponibles en el momento de la búsqueda.

"La propuesta presentada está basada en ontologías de dominio, un tesauro y una medida de similitud jerárquica"

Procesamiento

Por cada página html recogida se lleva a cabo el siguiente proceso:

– Se le extraen sus descriptores en un archivo *xml*. A fin de tener un contenido más limpio se elimina la publicidad que contenga la página así como el contenido *flash*, imágenes, sonido y enlaces ajenos.

– Se extrae una lista de los términos importantes a partir de los descriptores siguiendo el "modelo espacio

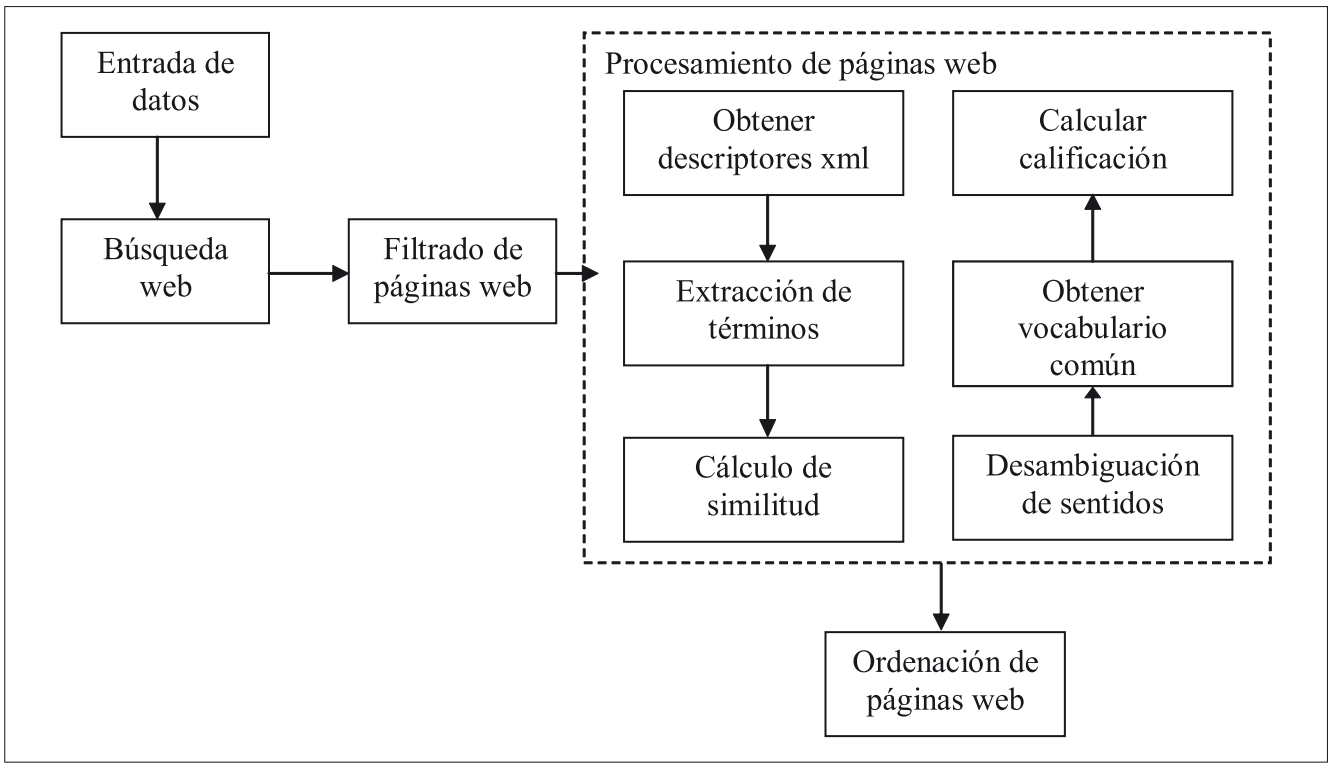


Figura 1. Arquitectura de búsqueda web semántica

vectorial"[17]. Se eliminan las palabras poco relevantes o *stop-words* (artículos, preposiciones, adverbios, conjunciones y pronombres personales) –extraídas de las listas utilizadas por *WVTool*[18] y *Bow*[19]–.

– Se calcula la similitud entre cada página y la ontología utilizando la "medida de similitud generalizada de cosenos"[9]. Esta medida es útil para conocer el grado de similitud entre dos colecciones agrupadas en forma de árbol. En esta medida se consideran dos colecciones: A (la ontología) y B (la página web). Estas colecciones se representan mediante vectores según la ecuación 1:

$$\vec{A} = \sum a_i l_i , \vec{B} = \sum b_j l_j \qquad (1)$$

donde $a_i = W(l_i) * Count(l_i)$ para $i = 1... n$ y $b_j = W(l_j) * Count(l_j)$ para $j = 1... n$;

l representa las hojas de las colecciones,

n es el número de nodos de la colección,

$W(l_i)$ es el peso del nodo l_i (en este caso $W(l_i) = 1$ para $i = 1... n$) y

$Count_A(l_i)$ es el número de veces que el nodo l_i aparece en la colección A.

Para dos términos (l_i y l_j) tomados de las colecciones antes mencionadas, el producto punto entre ellos está dado por la ecuación 2:

$$\vec{l}_i \cdot \vec{l}_j = \frac{2 * depth(LCA_U(l_i, l_j))}{depth(l_i) + depth(l_j)} \qquad (2)$$

El término *depth* de un nodo es el número de arcos que hay en la ruta de la raíz al nodo y *LCA* es el menor antecesor común (*lowest common ancestor*), es decir, el nodo con mayor *depth* que es antecesor tanto de l_i como de l_j. Para obtener esta medida de similitud de cosenos es necesario obtener primero el producto punto de las colecciones utilizando la ecuación 3:

$$\vec{A} \cdot \vec{B} = \sum \sum a_i b_j \vec{l}_i \cdot \vec{l}_j \qquad (3)$$

A continuación se muestra la fórmula para obtener la medida de similitud de cosenos entre 2 colecciones. Los detalles de esta ecuación se plantearon anteriormente[9]:

$$sim(A, B) = \frac{\vec{A} \cdot \vec{B}}{\sqrt{\vec{A} \cdot \vec{A}} \sqrt{\vec{B} \cdot \vec{B}}} \qquad (4)$$

Una vez calculada la similitud entre la ontología y cada página web, se realiza el proceso de desambiguación de sentidos[20] utilizando la lista de términos de cada página. Para esto se aplica el "método de vectores de contexto de segundo orden"[21]. Un vector de contexto indica todas las co-ocurrencias que tiene una palabra encontrada en un texto. Para obtener dichos vectores se utilizan las glosas o sentidos de *WordNet*. Una glosa es una definición y/o sentencia de ejemplo de una palabra. Por ejemplo, para la palabra *lamp, WordNet* devuelve las siguientes glosas:

(noun) lamp (an artificial source of visible illumination)
(noun) lamp (a piece of furniture holding one or more electric light bulbs)

Los *vectores de contexto* se obtienen usando los sentidos para cada palabra de cada página web. El primer conjunto de sentidos se denomina de "primer orden". El conjunto de sentidos que se obtiene a partir de las palabras relevantes que este conjunto contiene se llama de "segundo orden". Este proceso se realiza para la ontología usada en la búsqueda y cada página obtenida. Los conjuntos de sentidos se comparan con el fin de obtener las coincidencias entre ellos.

El proceso de desambiguación tiene por objetivo conocer las semejanzas entre el tema del que trata una página y el de la ontología. Para ello se utilizan los 10 sentidos con mayor número de repeticiones obtenidos a partir de las palabras que involucran. A cada uno de los sentidos se le asigna una calificación comenzando con 1,0 para el sentido con más repeticiones, y para los siguientes sentidos se irá disminuyendo proporcionalmente la calificación hasta llegar a 0,01. Así, la medida de desambiguación de una página web está dada por la ecuación 5:

$$disamTotal = \frac{\sum_{i=1}^{n} disamGloss_i}{n} \quad (5)$$

donde *n* es el número de sentidos obtenidos para dicha página (máximo 10 sentidos) y $disamGloss_i$ es la calificación asignada a la glosa o sentido *i*.

Como siguiente paso se construye un vocabulario de términos comunes a partir de los términos de cada página web obtenida en una búsqueda dada, eliminando los que ya se encuentren dentro de la ontología. El porcentaje del vocabulario que corresponda a cada página se asigna con la ecuación 6:

$$pCommVoc = \frac{totalTermsPage}{VocabularySize} \quad (6)$$

donde *totalTermsPage* es el número de términos que contiene la página y que no están en la ontología y *VocabularySize* es el tamaño del vocabulario de términos comunes.

A cada página se le asigna una calificación tomando en cuenta su similitud con la ontología, la calificación en la desambiguación de sentidos y su porcentaje del vocabulario común. Por el momento este cálculo es heurístico teniendo en cuenta la importancia de cada parte, para ello se utiliza la ecuación 7. Este cálculo podría mejorarse con una regresión lineal simple.

$$score = 0.3 * sim(A,B) + 0.5 * dis(B) + 0.2 * pCommVoc \quad (7)$$

donde *sim(A,B)* es la similitud entre la ontología A y la página web B, *dis(B)* es la calificación para la desambiguación de la página B y *pCommVoc* es el porcentaje del vocabulario común "informal". **Chignell**[22] propuso una asignación de pesos para la precisión de la web que podría utilizarse posteriormente ya que quita subjetividad en la identificación de las páginas web relevantes.

Ordenamiento de páginas

Como paso final en el proceso de búsqueda, la lista de páginas resultante se ordena de manera descendente de acuerdo con la calificación (*score*) que se le asignó en el paso anterior.

```
for i = 1 to totalEngines
   <links> = <links> + GETLINKSXENGINE(keyword)
   <taxomomy> = GETTAXONOMY(domain)
for i = 1 to size(<links>)
   <termsxLink> = GETTERMSXLINK(link_i)
   <sensesOnt> = GETSENSESWITHWORDNET(<taxonomy>)
for i = 1 to size(<links>)
   similarityPage(i) = GETSIMILARITY(<taxonomy>, <termsxLink>)
   <sensesxPage> = GETSENSESWITHWORDNET(link_i)
   disPage(i) = GETCOMMONSENSES(<sensesOnt>, <sensesxPage>)
   <ListVocabulary> = ADDCOMMONvOCABULARY(termsxLink)
   <LisVocabulary> = ELIMINATE(<taxonomy>)
for i= 1 to size(<links>)
   percent(i) = CALCULATEPERCENTOFVOCABULARY(link_i)
   CALCULATESCORE(i, similarityPage(i), disPage(i), percent(i))
ORDERBYSCORE(<links>)
Return(<links>)
```

Figura 2. Algoritmo de búsqueda web semántica

4. Resultados preliminares

La figura 2 presenta el algoritmo propuesto para la búsqueda semántica. Para probar el método han sido recogidas 80 ontologías de dominio de 20 temas de repositorios como *Protégé*[23], *Dumontier Lab*[24] y *SchemaWeb*[25] y se ha desarrollado un prototipo. Aunque el estudio sigue en proceso, se han realizado ya algunas pruebas (un análisis global y una búsqueda específica) y se han obtenido resultados alentadores que se muestran en la siguiente sección. Ambos experimentos fueron realizados en los motores de búsqueda *Excite*, *Google*, *HotBot*, *Metacrawler* y *MSN*.

Las pruebas se llevaron a cabo en un ordenador con procesador *AMD* 64 *Dual Core*, 2,3 *GHz* y 2 *GB* de *RAM*. El prototipo ha sido implementado con *Java server pages* (*JSP*) y *JavaBeans*, y es accesible en:

http://lsd.tamps.cinvestav.mx:8080/sws/

4.1. Desempeño general

Se han diseñado 30 casos de prueba, utilizando diferentes dominios y palabras clave, y realizando 31 consultas por cada caso para garantizar una distribución normal de los resultados (generalización del experimento), según el "Teorema del límite central"[26]. En la tabla 1 se muestran los dominios y palabras clave utilizados en cada conjunto de pruebas y el correspondiente

tiempo de ejecución promedio. Las primeras 7 pruebas son de usuarios no expertos en el tema y el resto fueron realizadas por expertos. El tiempo promedio de espera fue de 294 seg (4,9 min).

Los buscadores y el enfoque propuesto han sido comparados considerando estos cuatro aspectos:

– Capacidad de búsqueda. Se refiere a los beneficios que un buscador ofrece para recuperar información, así como los operadores de búsqueda que utiliza. Son considerados únicamente los operadores booleanos (*and*, *not*, *or*) y la búsqueda por frase literal. Un buscador que acepta todo lo anterior tendrá un 100% de capacidad de búsqueda.

– Cobertura. Cantidad de recursos que un buscador indexa para una búsqueda específica.

– Tiempo de respuesta. Tiempo transcurrido para presentar los resultados al usuario.

– *Recall*. Porcentaje de recursos recuperados que realmente corresponden al dominio elegido para la búsqueda (exhaustividad, utilidad).

Dominio	Palabra clave	Tiempo de ejecución promedio (seg)
Animals classification	Chihuahua dog	296
Animals classification	Rodent	195
Environment	Vegetation species	307
Types of plants	Salvia	384
Types of plants	Types of rose	152
Illnesses	Prostate cancer	384
Wines	Dessert wine	283
Animals classification	Giraffe characteristics	720
Wines	Breast structure	232
Wines	White wine	167
Catholic bible	Religious belief	472
Computer components	Entertainment software	336
Famous cars	Corvette	232
Famous cars	Jaguar	148
Famous cars	Mercedes Benz	293
Famous cars	Passengers cars	314
Food	Dessert	183
Illnesses	Appendectomy	131
Illnesses	Bone fracture	266
Illnesses	Sunburn	309
Movies	Drama movies	235
Movies	Horror movies	318
Music styles	Jazz	434
Music vocabulary	Saxophone	232
Music vocabulary	Soprano clarinet	540
Pizza	Vegetarian pizza	166
Plants	White rock rose	158
Types of food	Seafood	188
Types of publicity	Computer magazine	283
Wines	White wine	384

Tabla 1. Tiempos de ejecución por dominio y palabra clave

"Los buscadores seleccionados y el enfoque propuesto se comparan considerando cuatro aspectos: capacidad de búsqueda, cobertura, tiempo de respuesta y *recall*"

Sólo se consideraron los primeros 20 resultados por cada buscador. En la tabla 2 se muestra el promedio del resultado. Es importante señalar que el tiempo promedio dedicado por un usuario a elegir las páginas relevantes en un buscador tradicional ronda los 312 segundos (5,2 min)[27, 28], tiempo mayor al promedio obtenido en este experimento.

Buscador	Capacidad de búsqueda %	Cobertura	Tiempo de respuesta (seg)	Recall %
Excite	100	Bajo	2,453	72,17
HotBot	100	Medio	1,842	77,67
Google	100	Alto	1,769	75,50
Metacrawler	100	Bajo	3,455	66,00
MSN	100	Medio	3,861	62,50
Enfoque propuesto	100	Bajo	294,196	97,28

Tabla 2. Comparativa de los buscadores seleccionados y el enfoque propuesto para los conjuntos de prueba del experimento global

4.2. Búsqueda específica

Búsqueda puntual utilizando la palabra clave *white wine*. Las páginas web resultantes obtenidas por cada buscador se muestran en la tabla 3. En la tabla 4 se observa la lista final de páginas ordenadas por calificación descendente. En ella no se espera obtener una página con una calificación de 1,0 ya que esto significaría que está construida totalmente a partir de la ontología y contiene todos sus términos. La tabla 5 muestra los tiempos de ejecución y la 6 una comparativa entre los buscadores y el enfoque propuesto para este experimento.

Un usuario que utiliza un buscador tradicional debe invertir un tiempo en el análisis de las páginas resultantes, que comúnmente excede los 5 minutos. Utilizando el prototipo desarrollado sólo tendrá que visitar las páginas propuestas ya que los resultados obtenidos por los buscadores se han filtrado y ordenado por calificación, encabezando la lista los más precisos. Como puede verse en la tabla 6, nuestro método obtiene el mejor *recall* (exhaustividad).

A pesar de que nuestro prototipo puede tardar más tiempo en responder que un buscador tradicional, los resultados obtenidos tendrán mayor calidad ya que son

URL	Buscador
www.vinography.com/archives/white wine/	MSN
en.wikipedia.org/wiki/White wine#Red or white wine	HotBot
mustlovewine.com/group.php?group id=1	Metacrawler
www.thewinedoctor.com/advisory/tasteclassicgrapeswhite.shtml	Google
en.wikipedia.org/wiki/White wine	MSN, Excite, Metacrawler
www.thefreedictionary.com/white+wine	MSN
www.wine.com/wineshop/product list.asp?N=7155+125	Excite, HotBot
www.chiff.com/wine/white.htm	Google, MSN, Excite, Metacrawler
interviews.slashdot.org/article.pl?sid=04/05/17/0057241	Google
www.french-wine-online.com/	MSN
wine.about.com/od/howwineismade/r/Whitelightsangr.htm	MSN
wine.about.com/od/whitewines/A Guide to White Wines.htm	Google, HotBot, MSN, Excite
en.wikipedia.org/wiki/Wine	Google, MSN, Excite, Metacrawler
www.wine.com/	Metacrawler
www.whitewinegame.com/	Google, MSN, Excite, HotBot
www.recipezaar.com/library/getentry.zsp?id=184	Google, Excite
www.wine.com/wineshop/product list.asp?n=7155+125	Google, Metacrawler, HotBot
www.murrayscheese.com/	Excite
www.finorestaurant.com/wine-list/white-wine	MSN

Tabla 3. Resultados preliminares por buscador para la palabra clave "white wine" y el dominio "wines"

Calificación	URL
0,353	www.vinography.com/archives/white wine/
0,317	en.wikipedia.org/wiki/White wine
0,317	en.wikipedia.org/wiki/Wine
0,293	en.wikipedia.org/wiki/White wine#Red or white wine
0,226	interviews.slashdot.org/article.pl?sid=04/05/17/0057241
0,214	mustlovewine.com/group.php?group id=1
0,213	www.thewinedoctor.com/advisory/tasteclassicgrapeswhite.shtml
0,182	www.wine.com/wineshop/product list.asp?n=7155+125
0,172	www.chiff.com/wine/white.htm
0,138	www.french-wine-online.com/
0,133	wine.about.com/od/howwineismade/r/Whitelightsangr.htm
0,109	www.thefreedictionary.com/white+wine
0,058	wine.about.com/od/whitewines/A Guide to White Wines.htm

Tabla 4. Webs resultantes ordenadas por calificación para la palabra clave "white wine" y el dominio "wines"

Mejor	282 seg
Peor	462 seg
Media	385 seg
Mediana	384 seg
Desviación	39 seg

Tabla 5. Estadísticas de los tiempos de ejecución de las 31 pruebas con "white wine" y el dominio "wines"

Buscador	Capacidad de búsqueda %	Cobertura	Tiempo de respuesta (seg)	*Recall* %
Excite	100	Baja	1,637	80
HotBot	100	Media	1,302	90
Google	100	Alta	1,899	80
Metacrawler	100	Baja	3,122	70
MSN	100	Media	2,671	85
Enfoque propuesto	100	Baja	385,253	100

Tabla 6. Comparativa para la palabra clave "white wine" y el dominio "wines" que incluye los buscadores seleccionados y el enfoque propuesto

eliminados los enlaces rotos y las páginas que no corresponden al dominio. Esto evita la tediosa tarea de visitarlos sin tener éxito, lo cual se refleja en el tiempo ahorrado en la revisión de los resultados.

5. Conclusiones

El procedimiento presentado en este artículo está basado en ontologías de dominio, un tesauro y una medida de similitud jerárquica a fin de obtener una herramienta útil para la búsqueda semántica de páginas web. Ha sido implementado un prototipo con el que se han llevado a cabo algunas pruebas con resultados aceptables. Las páginas web irrelevantes y/o no disponibles se eliminan de los resultados parciales ofrecidos por los buscadores seleccionados, lo que evita visitarlas sin éxito. Es importante destacar que el tiempo de análisis de los resultados se reduce debido a la calidad lograda.

Por el momento las restricciones del prototipo son: 1) sólo pueden ejecutarse búsquedas relacionadas con la colección de ontologías, a diferencia de un buscador tradicional que permite cualquier búsqueda en cualquier dominio; 2) el tiempo de ejecución para filtrar

los resultados relevantes es alto debido a los cálculos realizados por los algoritmos.

Los esfuerzos actuales se dirigen a mejorar el cálculo de la calificación final de cada página web usando una regresión lineal simple. Así mismo, el prototipo será adaptado para su ejecución en un cluster de 32 núcleos y 12 TB de almacenamiento para mejorar el tiempo de ejecución.

Referencias

1. **Gruber, Thomas**. "Toward principles for the design of ontologies used for knowledge sharing". *Intl. journal human-computer studies*, 1993, v. 43, pp. 907-929.

2. **Fernández-López, Mariano**; **Gómez-Pérez, Asunción**; **Juristo, Natalia**. "Methontology: From ontological art towards ontological engineering". En: *Proceedings of the AAAI97 spring symposium series on ontological engineering*, 1997, pp. 33-40.

3. **Morato, Jorge**; **Marzal, Miguel**; **Lloréns, Juan**; **Moreiro, José**. "*WordNet* applications". En: *Proceedings of the Second global WordNet conference*, 2007, pp. 270-278.

9. **Van-Haren, Mark**; **McIntyre, Ryan**; **Lutch, Ben**; **Kraus, Joe**; **Spencer, Graham**; **Reinfried, Martin**. *Excite*.
http://www.excite.com/

5. **Brin, Sergey**; **Page, Lawrence**. "The anatomy of a large-scale hypertextual web search engine". *Computer networks and ISDN systems*, 2008, v. 30, pp. 107-117.

6. **Brewer, Eric**; **Gauthier, Paul**. *HotBot*.
http://www.hotbot.com.

7. **Selberg, Erik**; **Etzioni, Oren**. "The *MetaCrawler* architecture for resource aggregation on the Web". *IEEE expert*, 1997, pp. 11-14.

8. *Microsoft Corporation*. *MSN*.
http://www.msn.com/.

9. **Ganesan, Prasanna**; **Garcia-Molina, Hector**; **Widom, Jennifer**. "Exploiting hierarchical domain structure to compute similarity". *ACM transactions on information syst*ems, 2003, v. 21, pp. 64-93.

10. **Aguilar-López, Dulce**; **López-Arévalo, Iván**; **Sosa, Víctor**. "Usage of domain ontologies for web search". En: *Proceedings of DCAI*, 2008, pp. 319-328.

11. **Aguilar-López, Dulce**; **López-Arévalo, Iván**; **Sosa, Víctor**. "Web search based on domain ontologies". En: *15th Intl. multi-conference on advanced computer systems & computer information systems and industrial management applications (ACS)*, 2008.

12 **Bocio, Jaime**; **Isern, David**; **Moreno, Antonio**; **Riaño, David**. "Semantically grounded information search on the WWW". En: *Recent advances in artificial intelligence, research and development (Proceedings of Seté congrés català d'intel.ligència artificial (CCIA'04))*, 2004, pp. 349-356.

13. **Gao, Mingxia**; **Liu, Chunnian**; **Chen, Furong**. "An ontology search engine based on semantic analysis". En: *Icita 05: Proceedings of the Third intl. conf. on information technology and applications (Icita05)*, 2005, pp. 256-259.

14. **Ramachandran, Rahul**; **Movva, Sunil**; **Graves, Sara**; **Tanner, Steve**. "Ontology-based semantic search tool for atmospheric science". En: *22nd Intl. conf. on interactive information processing systems (IIPS), 86th American Meteorological Society annual meeting*, 2006.

15. **Droegemeier, Kevin**; **Gannon, Dennis**; **Reed, Daniel** et al. "Service-oriented environments in research and education for dynamically interacting with mesoscale weather". *IEEE computing in science & engineering*, 2005, v. 7, pp. 24-32.

16. **Sánchez-Ruenes, David**. *Domain ontology learning from the Web*. PhD tesis, *Universitat Politècnica de Catalunya*, *Departamento de lenguajes y sistemas informáticos*, 2007.

17. **Salton, Gerard**; **Wong, Anita**; **Yang, Chung-Shu**. "A vector space model for automatic indexing". *Commun. ACM*, 1975, v. 18, pp. 613-620.

18. **Wurst, Michael**. *The word vector tool user guide*.
http://nemoz.org/ joomla/mining/wvtool/wvtool.pdf.

19. **McCallum, Andrew Kachites**. *Bow: a toolkit for statistical language modeling, text retrieval, classification and clustering*.
http://www.cs.cmu.edu/mccallum/bow.

20. **Lesk, Michael**. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone". En: *Proceedings of the 5th annual intl. conf. on systems documentation (Sigdoc)*, pages 24-26, New York, NY, USA, 1986. ACM.

21. **Patwardhan, Siddharth**; **Pedersen, Ted**. "Using *WordNet* based context vectors to estimate the semantic relatedness of concepts". En: *Proceedings of the EACL 2006 workshop making sense of sense-bringing computational linguistics and psycholinguistics together*, 2006, pp. 1-8.

22. **Chignell, Mark**; **Gwizdka, Jacek**; **Bodner, Richard**. "Discriminating meta-search: a framework for evaluation". *Information processing and management*, 1999, v. 35, n. 3, pp. 337-362.

23. **Noy, Natalya**; **Ferguerson, Ray**; **Musen, Mark**. "The knowledge model of Protégé-2000: combining interoperability and flexibility". En: *12th Intl. conf. in knowledge engineering and knowledge management (EKAW00). Lecture notes in artificial intelligence*, 2000, v. 1937, pp. 17-32.

24. **Dumontier, Michel**; **Villanueva-Rosales, Natalia**. "Modeling life science knowledge with OWL 1.1. En: *Fourth intl. workshop OWL experiences and design*, *Owled 2008*, Washington, DC.

25. **Lindesay, Victor**. *SchemaWeb directory*.
http://www.schemaweb.info

26. **Wackerly, Dennis**; **Mendenhall, William**; **Scheaffer, Richard**. *Estadística matemática con aplicaciones*. Cengage Learning Editores, 2002.

27. **Silverstein, Craig**; **Marais, Hannes**; **Henzinger, Monika**; **Moricz, Michael**. "Analysis of a very large web search engine query log". *Sigir forum*, 1999, v. 33, n. 1, pp. 6-12.

28. **Jansen, Bernard**; **Spink, Amanda**; **Saracevic, Tefko**. "Real life, real users, and real needs: A study and analysis of user queries on the web". *Information processing and management*, 2000, v. 36, n. 2, pp. 207-227.

Dulce Aguilar-Lopez, Ivan Lopez-Arevalo, Victor Sosa-Sosa
Laboratorio de Tecnologías de Información, Cinvestav.
daguilar@tamps.cinvestav.mx
ilopez@tamps.cinvestav.mx
vjsosa@tamps.cinvestav.mx