EXPANSIÓN DE CONSULTAS BASADA EN ONTOLOGÍAS PARA UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

H. Kuna[1], **M. Rey**[1], **L. Podkowa**[1], **E. Martini**[1], **L. Solonezen**[1]

1. Programa de investigación en Computación. Depto. de Informática, Facultad de Ciencias Exactas Quím. y Naturales Universidad Nacional de Misiones.

hdkuna@gmail.com

RESUMEN

Cuando un usuario ingresa una consulta para buscar información en bibliotecas digitales, éste espera que los resultados respondan a sus expectativas de la mejor manera. Los Sistemas de Recuperación de Información buscan optimizar el proceso de búsqueda y recuperación de información utilizando diversas herramientas. Entre las variantes de este tipo de sistemas se encuentran los meta-buscadores, que cuentan con la capacidad de ampliar el espectro de cobertura en una búsqueda, a partir de la posibilidad de utilizar las bases de datos de varios buscadores en simultáneo; además de poder incorporar diversos métodos para mejorar el espectro de documentos a ser recuperados. En este trabajo se presenta el desarrollo del componente para la expansión de consultas, basado en una ontología de dominio específico, en un sistema de recuperación de información para la búsqueda de documentos científicos en el área de las ciencias de la computación.

Palabras clave: *recuperación de información, ontología, expansión de consultas, inteligencia artificial.*

CONTEXTO

Está línea de investigación articula el "Programa de Investigación en Computación" de la Facultad de Ciencias Exactas Químicas y Naturales (FCEQyN) de la Universidad Nacional de Misiones (UNaM); el Grupo de Investigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España.

1 INTRODUCCION

1.1 Sistemas de Recuperación de Información

Un Sistema de Recuperación de Información (SRI) es un proceso capaz de almacenar, recuperar y mantener información [1], [2]. En la actualidad los principales modelos de SRI que operan sobre internet son: los directorios, los buscadores y los meta-buscadores [3]. En el contexto de la presente investigación cobran mayor relevancia los meta-buscadores, ya que los mismos poseen una estructura modular en la cual cada componente puede ser desarrollado en forma específica para un dominio particular [4]. Uno de los componentes que incluye un meta-buscador es aquel realiza expansiones de las consultas (QE por su sigla en inglés) ingresadas por el usuario en base a diferentes métodos, generando una serie de consultas extra que tienen por objetivo ampliar el espectro de búsqueda para recuperar documentos de mayor relevancia para el usuario [5], [6].

1.2 Ontologías

Una ontología define los términos y las relaciones que conforman el vocabulario básico de un dominio en particular, incorporando reglas para combinar conceptos y relaciones para extender el vocabulario [7], [8]. En un sentido más general, se puede considerar a una ontología como un sistema de representación del conocimiento en un ámbito específico, por lo tanto, puede organizarse en forma jerárquica para facilitar la representación de conocimiento [9].

Desde la óptica de las Ciencias de la C omputación (CS por su sigla en inglés), una ontología se refiere a la formulación de un esquema conceptual en un dominio acotado, facilitando la comunicación y pasaje de información entre sistemas [10]. Permitiendo, a través de una estructura de clases y subclases, el tratamiento y análisis del conocimiento a la luz de las relaciones, las propiedades y las reglas definidas entre las instancias de tales clases, sirviendo como herramienta para la recuperación automática de información [11].

1.2 Expansiones de las consultas en un SRI

En un SRI se suelen utilizar diferentes métodos para la optimización del proceso de búsqueda de información, uno de ellos es el de QE [5], [12]. En general se trata de un proceso por el que se toma la consulta original ingresada por el usuario y se la amplía a partir de la inclusión de diversos términos, obtenidos a partir de fuentes de información externas, que guardan relación con el dominio de la consulta. De esta manera con los términos originales y los obtenidos en foram externa se genera una serie de consultas alternativas, denominadas expansiones [6], [13]. A partir de éstas se busca que el SRI adquiera la capacidad de acceder a nuevos documentos que sean relevantes para el usuario, pudiendo ejecutar en paralelo las consultas a las diferentes fuentes de datos definidas y sus resultados posteriormente ser mezclados y ordenados generando el listado final a ser presentado al usuario [4], [5].

Existen diferentes métodos para realizar procesos de expansión de consultas, como ser: uso de tesauros, diccionarios, sistemas expertos, entre otros [6], [13], [14], [15].

En este trabajo se presenta un método de QE para un SRI, específicamente un meta-buscador, basado en la utilización de una ontología de dominio específica para un sub-área temática dentro de las CS, se detalla el proceso de construcción de la ontología y el método de QE basado en el uso de la misma.

2 LÍNEAS DE INVESTIGACION, DESARROLLO E INNOVACIÓN

El desarrollo de un SRI que haga uso de diversos métodos de búsqueda para su aplicación específica en bases de datos de documentos científicos del área de CS, implica la generación de un conjunto de componentes que permitan mejorar la relevancia de los resultados a retornar al usuario. Uno de los componentes que tiene implicancia directa en la calidad de los resultados es aquel que efectúa la transformación de la consulta original del usuario con el objetivo de ampliar el espectro de búsqueda del SRI. En este contexto, las ontologías se presentan como una alternativa más que válida para la generación de un método de QE en un ámbito como es la recuperación de documentos científicos.

3 RESULTADOS Y OBJETIVOS

3.1 Construcción de la ontología

Para el diseño de la ontología se siguieron los siguientes pasos [16]:

1. Definición del dominio de la ontología
2. Definición de los términos a incluir en la ontología
3. Definición de clases y jerarquía de clases
4. Definición de las propiedades de las clases
5. Creación de las instancias internas a cada clase

En el caso del presente trabajo la defnición del dominio consistió en seleccionar cuál sub-área, dentro de las CS, sería con la que se trabajaría inicialmente. Se determinó comenzar por el sub-área de Inteligencia Artificial (IA).

Para el segundo de paso se analizó el estado del arte de la disciplina y se generó un listado de términos conformado por conceptos que defninen y conforman las sub-áreas en las que se puede subdividir la IA, además de diversos sinónimos de tales términos que luego pudieran ser utilizados como propiedades de los elementos de la ontología [17]–[19].

En cuanto al desarrollo del tercer paso, se utilizó el método top-down para la generación de la jerarquía de clases, dado que tal método se considera más natural para el contexto de la ontología. Se definieron como clases aquellos conceptos que representan categorías en las que se subdivide la disciplina y, en algunos casos, dentro de ellas se han definido instancias que se convirtieron en subclases para que la ontología refleje de mejor manera la taxonomía inherente a los conceptos a utilizar.

Para el cuarto paso se definieron las propiedades de cada clase de la ontología, considerando como propiedad a todo atributo que pudiera describir a un concepto, incluyendo las relaciones que pudiera tener con otros elementos. En cuanto a las relaciones entre los conceptos, se definieron dos tipos: las relaciones implicitas y las explicitas. Las primeras son aquellas que están dadas por la estructura propia de la ontología, las mismas no se definen formalmente sino que se crean a partir de las conexiones existentes entre las clases, resultando en tres posibles subtipos:

- "Es un (padre)", simbolizando que una clase determinada "contiene" a otra clase.
- "Es un (hijo)", simbolizando que una clase determinada "es contenida por" otra clase.
- "Es un (hermano)", simbolizando que n clases "son contenidas por" una misma clase "padre".

Mientras que las explicitas son las que se definen específicamente según el dominio de la ontología, en este caso se definió la relación de sinónimos, donde cada elemento de la ontología puede tener asociado un término con igual significado.

Para finalizar el diseño de la ontología se determinaron aquellos conceptos que conformarían las instancias de cada clase definida, para lo que fueron utilizados los términos de mayor atomicidad obtenidos del listado conformado previamente.

Una vez completo el diseño de la ontología se prosiguió con su implementación, para ello se utilizó una herramienta software específica para operar con ontologías como es Protègè [20], en la misma no se ha hecho más que reproducir la estructura generada previamente para luego exportarla y posiblitar su uso en la QE.

3.2 Desarrollo del algoritmo de expansiones de las consultas

Con la ontología implementada se prosiguió con el desarrollo del método que realiza la expansión de las consultas. El objetivo principal del algoritmo es buscar en la ontología aquel término más similar a la consulta ingresada por el usuario al SRI, en adelante "*consulta_original*", para utilizar tal elemento y obtener, a través de sus relaciones y propiedades, a su concepto "padre", los conceptos que se encuentran en su mismo nivel, es decir sus "hermanos" y, de existir, sus sinónimos para realizar la expansión en sí.

El algoritmo se compone de dos etapas, por un lado una primera instancia de selección del concepto de la ontología más similar a la *consulta_original* y luego la generación de las consultas a partir de la expansión para hacer las búsquedas.

La etapa de selección del algoritmo se compone de los siguientes pasos:

1. Para cada término de la consulta ingresada: se recorre la ontología (clases, instancias y propiedades) buscando el / los conceptos con mayor cantidad de coincidencias con el término, el resultado se almacena en una colección temporal.
2. Se analiza el contenido de la colección generada en el paso anterior:
    a. En caso de no contener ningún elemento se da por finalizada la operación de expansión sin resultados.
    b. Si la colección contiene únicamente un elemento se selecciona al mismo para continuar con la expansión y se lo pasa a denominar "*término_candidato*".
    c. Si existe más de un elemento, se debe analizar cada uno de ellos

para determinar cuál contiene la mayor cantidad de coincidencias sintácticas con el término en evaluación de la *consulta_original*. En caso de empate se hace uso de las relaciones ontológicas definidas entre los elementos de la siguiente manera:

i. Si todos poseen el mismo "padre" se selecciona a ese concepto como el *término_candidato*.
ii. Si los conceptos no comparten el mismo "padre" se toma a aquel que referencie más instancias.
iii. En caso de presentarse un empate en la cantidad de instancias de los "padres" evaluados se consideran todos ellos como término_candidato a través de una nueva colección.

Finalizada la etapa de selección se cuenta con uno o varios candidatos para dar inicio al proceso de expansión. Éste se descompone en los siguientes pasos, en caso de existir más de un término_candidato se repite el mismo proceso para cada uno de ellos:

1. Se obtiene el concepto "padre" del *término_candidato*, denominado "*concepto_padre*".
2. Se obtienen aquellos conceptos del mismo nivel que el *término_candidato*, sus "hermanos", generando la colección "*[conceptos_hermanos]*".
3. Se obtienen los sinónimos del *término_candidato*, generando la colección "*[sinónimos _concepto]*".
4. Se generan las consultas expandidas a partir de los resultados de los pasos anteriores y la consulta original, las nuevas consultas se construyen de la siguiente manera:
   - Expansión_1 = *consulta_original* AND *término_candidato*
   - Expansión_2 = *término_candidato* AND *concepto_padre*
   - Expansión_3 = *término_candidato* OR *[conceptos_hermanos]*
   - Expansión_4 = *término_candidato* OR *[sinónimos_concepto]*

Como resultado de la ejecución de ambas fases del algoritmo se cuenta con una serie de cadenas de caracteres correspondientes a las expansiones de la *consulta_original* ingresada por el usuario al SRI. Cada una de las consultas será ejecutada sobre cada fuente de datos definida en el SRI obteniendo listados de resultados que posteriormente serán unificados, ordenados y presentados al usuario.

3.3 Trabajos Previstos en la Próxima Etapa

Para el año 2014 se tiene previsto:

- Validar el desarrollo de la ontología con expertos en el dominio.
- Generalizar la ontología a otras sub-áreas de las CS.
- Incorporar elementos que permitan al SRI operar con múltiples lenguajes, adaptando en cada caso al método desarrollado para la QE.

4 FORMACION DE RECURSOS HUMANOS

Este proyecto es parte de las líneas de investigación del "Programa de Investigación en Computación" de la FCEQyN de la UNaM, con siete integrantes (todos ellos alumnos, docentes y egresados de la carrera de Licenciatura en Sistemas de Información de la FCEQyN – UNaM) de los cuales tres están realizando su tesis de grado, uno se encuentra realizando una maestría y dos están realizando un doctorado. Esta línea de investigación vincula al "Programa de Investigación en Computación" del Departamento de Informática de la FCEQyN de la UNaM, al Grupo de Iventigación Soft Management of Internet and Learning (SMILe) de la Universidad de Castilla-La Mancha, España.

5 BIBLIOGRAFIA

[1] G. Salton y M. Mcgill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.

[2] G. Kowalski, *Information Retrieval Systems: Theory and Implementation*, 1st ed. Norwell, MA, USA: Kluwer Academic Publishers, 1997.

[3] J. A. Olivas, *Búsqueda Eficaz de Información en la Web*. La Plata, Buenos Aires, Argentina: Editorial de la Universidad Nacional de La Plata (EDUNLP), 2011.

[4] J. Serrano-Guerrero, F. P. Romero, J. A. Olivas, y J. de la Mata, «BUDI: Architecture for fuzzy search in documental repositories», *Mathware & Soft Computing*, vol. 16, n.º 1, pp. 71–85, 2009.

[5] J. Ruiz-Morilla, J. Serrano-Guerrero, J. Olivas, y E. Viñas, «Representación Múltiple de Consultas: Una alternativa a la Expansión de Consultas en Sistemas de Recuperación de Información», en *Actas del XV Congreso Español sobre Tecnologías y Lógica Fuzzy. ESTYLF*, 2010, pp. 531–536.

[6] M. de la Villa, S. García, y M. J. Maña, «¿De verdad sabes lo que quieres buscar? Expansión guiada visualmente de la cadena de búsqueda usando ontologías y grafos de conceptos», *Procesamiento del Lenguaje Natural*, vol. 47, n.º 0, pp. 21-29, sep. 2011.

[7] F. E. Grubbs, «Procedures for Detecting Outlying Observations in Samples», abr. 1974.

[8] J. Hendler, «Agents and the semantic web», *IEEE Intelligent systems*, vol. 16, n.º 2, pp. 30–37, 2001.

[9] S. Delisle, «Towards a better integration of data mining and decision support via computational intelligence», en *Sixteenth International Workshop on Database and Expert Systems Applications, 2005. Proceedings*, 2005, pp. 720 - 724.

[10] A. Muñoz y J. Aguilar, «Ontología para bases de datos orientadas a objetos y multimedia», *Avances en Sistemas e Informática*, vol. 6, n.º 2, pp. 167–184, 2009.

[11] S. E. S. Sánchez López, «Modelo de indexacion de formas en sistemas VIR basado en ontologias», Maestría, Universidad de las Américas Puebla, México, 2007.

[12] A. H. Alsaffar, J. S. Deogun, V. V. Raghavan, y H. Sever, «Enhancing Concept-Based Retrieval Based on Minimal Term Sets», *Journal of Intelligent Information Systems*, vol. 14, n.º 2-3, pp. 155-173, mar. 2000.

[13] Y. Chang, I. Ounis, y M. Kim, «Query reformulation using automatically generated query concepts from a document space», *Information Processing & Management*, vol. 42, n.º 2, pp. 453-468, mar. 2006.

[14] S. Gauch y J. B. Smith, «An expert system for automatic query reformation», *Journal of the American Society for Information Science*, vol. 44, n.º 3, pp. 124-136.

[15] J. C. French, D. E. Brown, y N.-H. Kim, «A classification approach to Boolean query reformulation», *J. Am. Soc. Inf. Sci.*, vol. 48, n.º 8, pp. 694-706, ago. 1997.

[16] N. F. Noy y D. L. Mcguinness, «Ontology Development 101: A Guide to Creating Your First Ontology», 2001.

[17] E. A. Feigenbaum, A. Barr, y P. R. Cohen, *The handbook of artificial intelligence*. Addison-Wesley New York, 1989.

[18] E. Rich y K. Knight, «Artificial intelligence», *McGraw-Hill, New*, 1991.

[19] N. J. Nilsson, *Principles of artificial intelligence*. Springer, 1982.

[20] Stanford Center for Biomedical Informatics Research, *Protègè*. Stanford University, 2014.