Primera aproximación de un sistema de recuperación de información booleano con expansión semántica de consultas

Mireya Tovar Vidal, Ana Laura Lezama Sánchez, Darnes Vilariño Ayala, Beatriz Beltrán, Mauricio Castro Cardona

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla,
Mexico

{mtovar,darnes,bbeltran,mcastro}@cs.buap.mx
yumita1102@gmail.com

Resumen. En el presente trabajo se propone una aproximación que utiliza la expansión de consultas en un Sistema de Recuperación de Información Booleano (SRIB), con la finalidad de mejorar el nivel de precisión de un SRIB sin expansión. Las consultas están formadas por los conceptos y relaciones existentes en ontologías de dominio. El SRIB sin expansión y con expansión asocia a cada consulta la información relevante extraída desde el corpus de dominio. En base a los resultados experimentales obtenidos, se observa que la precisión del SRIB con expansión mejora al SRIB sin expansión, al recuperar más información, incluso al identificar más conceptos con información en el corpus, que el sistema tradicional sin expansión. Se analizaron cuatro ontologías de dominio y los resultados experimentales obtenidos resultan ser satisfactorios con esta aproximación.

Palabras clave: Sistema de recuperación de información, expansión semántica de consultas, ontologías.

1. Introducción

La Recuperacón de Información*(RI)* es el área de la ciencia y la tecnología que trata de adquirir, representar, almacenar, organizar y acceder a elementos de información. Desde el punto de vista práctico, dada una necesidad de información del usuario, un sistema de RI produce como salida un conjunto de documentos cuyo contenido satisface potencialmente esa necesidad. Esta última puntualización es de suma importancia, ya que la función de un sistema de RI no es la de devolver la información deseada por el usuario, sino únicamente la de indicar qué documentos son potencialmente relevantes para dicha necesidad de información.

Hoy en día la búsqueda de información es el eje central de cualquier investigación. Las búsquedas son proporcionadas por el usuario en su lenguaje natural y se

espera que los documentos recuperados sean aquellos que satisfagan la consulta realizada.

Esta investigación parte de un sistema de recuperación de información que permite recuperar documentos de un corpus de dominio, asociados a cada concepto y relaciones de una ontología de dominio. Tales conceptos y relaciones son utilizados como consultas que se emplean en la entrada a dicho sistema. En [16] se emplea un Sistema de Recuperación de Información Booleano y la información recuperada por cada concepto y relación es utilizada posteriormente para la evaluación automática de ontologías de dominio. Con la finalidad de mejorar la precisión de este sistema, se propone la extensión al mismo. En este caso se añade únicamente la expansión semántica de los términos que forman la consulta, en este caso la consulta está formada por los sinónimos exactos de los conceptos de la ontología extraídos desde WordNet [9].

Esta investigación está estructurada de la siguiente manera: en la sección 2 se describe la información general sobre sistemas de recuperación de información, en la sección 3 se presentan algunas propuestas por diversos autores para la expansión de consultas, en la sección 4 se describe la aproximación propuesta, en la sección 5 se presentan los experimentos y el conjunto de datos y finalmente en la sección 6 se discuten las conclusiones y el trabajo a futuro.

2. Sistemas de recuperación de información

La Recuperación de Información (RI) ha sido interpretada por diversos autores. En el caso de Ricardo Baeza-Yates et al. [1] " la Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información". Salton [12] propuso una definición que plantea que el área de RI "es un campo relacionado con la estructura, ánalisis, organización, almacenamiento, búsqueda y recuperación de información". Croft [15] estima que la recuperación de información es el "conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado". Korfhage [7] definió RI como "la localización expresada como una pregunta". De manera más genéral, se puede plantear que la recuperación de información intenta resolver el problema de "encontrar y ordenar documentos relevantes que satisfagan la necesidad de información de un usuario, expresada en un determinado lenguaje de consulta" [15].

Uno de los modelos existentes para la recuperación de información es el modelo booleano que representa la colección de documentos como una matriz binaria documento-término. Los términos son extraídos de los documentos y representan el contenido de los mismos. Se utilizan operadores lógicos: AND, OR y NOT, y los resultados son referencias a documentos, donde la representación de la consulta satisface las restricciones lógicas de la expresión de búsqueda. En el modelo original no hay orden de relevancia sobre el conjunto de respuestas a la consulta, todos los documentos poseen la misma relevancia [15].

La precisión de los sistemas de recuperación de información depende mucho de los términos que se encuentran en la consulta, es por ello que intentar de

manera eficiente expandir la consulta, puede aunmentar la cantidad y calidad de los documentos recuperados y satisfacer la necesidad de información dada por el usuario.

3. Trabajos relacionados

En el caso de la expansión de consultas por sinónimos, algunos autores han recurrido a diferentes técnicas de expansión, así como diferentes modelos de recuperación de información. A continuación se describen algunos trabajos relacionados con esta investigación.

En Cotelo et al. [3] el problema principal consiste en definir un lenguaje de consulta que sea utilizado para recibir consultas con información semántica y un algoritmo de ordenamiento que permita ordenar los documentos. Dentro de las características deseables del lenguaje de consulta se encuentran: identificar objetos y atributos de los mismos, permitir al usuario indicar el significado de una palabra polisémica, incluir semántica temporal en las consultas, expandir la consulta con sinónimos y permitir operadores sobre los predicados.

Kuna et al. [8], utiliza una ontología de dominio específico para la expansión de consultas, además de un sistema de recuperación de información para la búsqueda de documentos científicos.

En Valbuena et al. [18] se propone el uso de ontologías para garantizar que los resultados en una búsqueda hecha por el usuario, correspondan al dominio de la misma.

En Muñoz et al.[4] se propone el desarrollo de un sistema de recuperación de información en Inteligencia Artificial enfocado a textos médicos, con el objetivo de conseguir un sistema destinado a introducirse en el campo de la Medicina Personalizada y en el campo turístico.

En Hernández-Aranda et al.[6] se desarrolló un prototipo que consta de una interfaz web que permite la búsqueda y visualización de resultados a partir de una consulta dada.

Shabanzadeh et al.[14], proponen un algoritmo para la expansión de consultas basado en relaciones semánticas, utilizan Wordnet para extraer las relaciones semánticas entre palabras. Se demostró que las relaciones semánticas pueden mejorar la expansión de consultas, que las palabras vagas reducen el rendimiento de la recuperación de información.

Chauhan et al.[2], proponen la técnica de expansión de consulta semántica que incluye un modelo matemático para calcular la similitud semántica entre conceptos y un algoritmo para la expansión de consultas basado en una ontología de dominio.

En Moreno et al. [13], se implementó una búsqueda textual sobre una ontología, permitiendo obtener los conceptos de la ontología en función de una búsqueda expresada en lenguaje natural.

Neha et al. [10], proponen un algoritmo genético para la expansión de consultas hechas en lenguaje natural, se utiliza el coeficiente de Czekanowski durante el proceso de expansión, para que la recuperación de documentos sea más eficiente.

Finalmente, en Hany et al. [5] se emplea el modelo espacio vectorial que se adaptó en su propuesta de trabajo para la representación de documentos, retira palabras vacías, etc. La consulta es expandida por sinónimos extraídos de Wordnet.

En esta investigación se propone el uso de los sinónimos recuperados desde WordNet de los conceptos que integran a la ontología, para la expansión de consultas. Las consultas están formadas por las palabras de cada concepto de la ontología y por otro lado por los sinónimos de estos conceptos. También se presenta un algoritmo que realiza la unión de los documentos recuperados por el Sistema de Recuperación de Información Booleno con los conceptos y sus sinónimos correspondientes. La finalidad de esta investigación es la de incorporar información adicional, como los documentos que contienen al sinónimo del concepto y al concepto mismo, para la evaluación posterior de los mismos y las relaciones semánticas existentes en la ontología de dominio. A continuación se presenta la aproximación propuesta.

4. Aproximación para la expansión de consultas

En este artículo se plantea la expansión de consultas por sinónimos, la cual se utiliza para recuperar documentos relevantes a la misma, por medio de un sistema de recuperación de información booleano. Las consultas están formadas por las palabras que integran los conceptos extraídos de ontologías de dominio.

A continuación se presentan las etapas de la aproximación propuesta:

1. Extracción de conceptos y relaciones de las ontologías de dominio.
2. Extracción de los sinónimos de los conceptos desde WordNet.
3. Preprocesamiento del corpus de dominio, de los conceptos, de las relaciones y de los sinónimos. Esta etapa incluye las siguientes acciones:
   a) División del corpus en líneas.
   b) Eliminación de símbolos especiales, números y palabras cerradas.
   c) Aplicación de un lematizador, en particular se utiliza el algoritmo de Porter [11].
4. Formación de consultas. Existen tres tipos de consultas:
   a) Consultas formadas con las palabras del concepto.
   b) Consultas formadas con los sinónimos del concepto.
   c) Consultas formadas con los dos conceptos que forman la relación semántica.
5. Aplicación del Sistema de Recuperación de Información Booleno (SRIB) para conceptos.
6. Aplicación del Sistema de Recuperación de Información Booleno (SRIB) para los sinónimos de los conceptos.
7. Mezcla de los resultados obtenidos (posting) por el SRIB de los dos pasos anteriores. La mezcla consiste en la unión de postings sin repetir información.
8. Aplicación del operador AND para la consulta que incluye los dos conceptos que forman la relación semántica. El operador AND realiza la intersección de las líneas que integran los posting de ambos conceptos que forman la relación semántica.

9. Evaluación de resultados obtenidos tanto para los conceptos como para las relaciones. La medida de evaluación que se utiliza en este caso es la de precisión.

$$P_C = \frac{Conceptos\ recuperados}{Total\ conceptos} \tag{1}$$

$$P_R = \frac{Relaciones\ recuperadas}{Total\ relaciones} \tag{2}$$

Donde: *Conceptos recuperados* es el total de conceptos obtenidos por el SRIB, y el *Total conceptos* es el total de conceptos existentes en la ontología de dominio. En el caso de *Relaciones recuperadas* se evalúa por separado las relaciones tipo class-inclusion y las relaciones no taxonómicas (para más información ver [17]). El *Total relaciones* corresponden a las relaciones de cada tipo recuperadas de la ontología de dominio evaluadas de manera independiente.

La Figura 1 muestra el comportamiento de manera gráfica de este algoritmo.

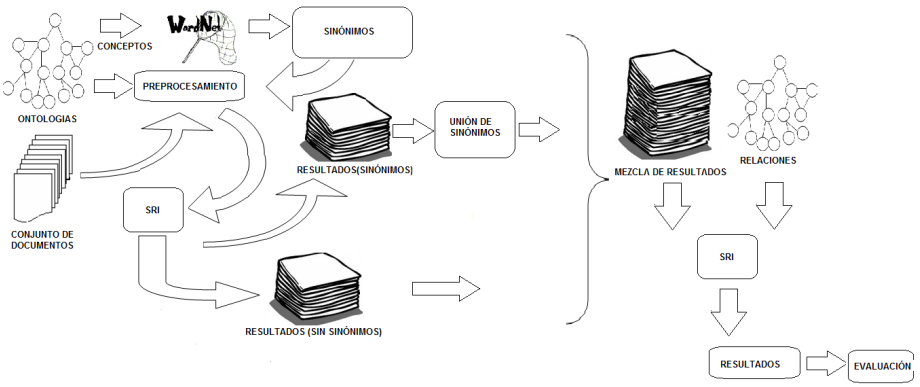

Fig. 1. Primera aproximación para la expansión de consultas en un SRIB.

5. Resultados experimentales

En esta sección, se presentan los datos utilizados (5.1) y los resultados obtenidos en los experimentos (5.2).

5.1. Conjunto de datos

En la Tabla 1 se presenta el número de conceptos (C), el total de relaciones class-inclusion (CI) y el total de relaciones no taxonómicas (NT) de las ontologías evaluadas. También se incluye el número de documentos (D), número de tokens

(T), cantidad de vocabulario (V), y el número de oraciones. Los dominios utilizados en los experimentos son Inteligencia Artificial (IA), Aprendizaje e-Learning (SCORM) [19], ontología del dominio de Petróleo (OIL), y Turismo (Turismo).

Tabla 1. Conjunto de datos.

Dominio	Ontología			Corpus de referencia				
	C	SC	NT	D	T	V	O	S
AI	276	205	61	8	11,370	1,510	475	415
SCORM	1,461	1,038	759	36	1,621	34,497	1,325	1,606
OIL	48	37	-	577	546,118	10,290,107	168,554	157,276
Turismo	963	1,016	-	1,801	877,519	32,931	36,505	31,418

5.2. Resultados obtenidos

A continuación se presentan los resultados experimentales obtenidos por los dos algoritmos desarrollados y su comparación, es decir, resultados del Sistema de Recuperación de Información Booleno (SRIB) sin expansión de consultas y del Sistema de Recuperación Información Booleano (SRIB) con expansión de consultas.

Los resultados obtenidos por ambos algoritmos, para el caso de los conceptos, se muestran en la Tabla 2 para cada ontología revisada (Dominio). En la Tabla 2 también se muestra el total de conceptos extraídos de la ontología (CO), los conceptos recuperados por el SRIB sin expansión (C), los conceptos que no obtuvieron líneas asociadas (F) y la precisión (P); los conceptos recuperados por el SRIB con expansión (CA), los conceptos que no logró recuperar el SRIB con expansión (FA) y la precisión obtenida (PA).

Además, en la tabla se incluye la cantidad de oraciones obtenidas por el SRIB sin expandir (OC), con expansión (OCA), la diferencia del número de líneas recuperadas con expansión y sin ella (OCE) y el porcentaje de incremento (%). En base a los resultados obtenidos para los conceptos, se observa que en los casos de los dominios de SCORM y Turismo principalmente, se incrementó el número de conceptos recuperados que los que se recuperan con el SRIB sin expansión. Además, la cantidad de oraciones que contienen los sinónimos del concepto incrementa la cantidad de líneas u oraciones asociadas a cada concepto de las ontologías, esto ocurre para cada dominio. El porcentaje de incremento de la información recuperada por el SRIB con expansión es mayor al 27 %, lo que indica que el concepto puede ser representado en el corpus por su sinónimo correspondiente y que esta información es adicional a la presentada por el SRIB sin expanción.

En la Tabla 3 se presentan los resultados obtenidos por ambos Sistemas de Recuperación de Información con expansión y sin ella, para relaciones de tipo class-inclusion de cada ontología de dominio. La columna OSC corresponde al

Tabla 2. Resultados del Sistema de Recuperación Booleano con expansión para el caso de los conceptos de cada ontología de dominio.

Dominio	Ontología							SRI			
	CO	C	F	P	CA	FA	PA	OC	OCA	OCE	%
IA	276	274	2	0.992	274	2	0.992	1,992	3,110	1,118	56.12 %
SCORM	1,461	1,443	18	0.987	1,444	17	0.988 %	23,479	31,833	8,354	35.58 %
OIL	48	48	0	1.00	48	0	1.00	232,603	297,234	64,631	27.78 %
Turismo	963	683	280	0.709	711	252	0.736	86,077	232,855	146,778	170.51 %

total de relaciones tipo class-inclusion incluidas en la ontología de dominio correspondiente. La columna SC es el total de conceptos recuperados con información del SRI sin expansión. La columna correspondiente a F es la diferencia de las relaciones recuperadas por el SRI booleano sin expansión y con expansión (FA). La precisión del sistema sin expansión (P) y con expansión (PA). También se incluye la cantidad de oraciones recuperadas en total por el SRIB sin expansión (OSC) y con expansión ($OSCA$) para este tipo de relaciones, la diferencia obtenida (OE) y el porcentaje de la diferencia (%). En base a los resultados obtenidos se observa que el número de relaciones de tipo class-inclusion de las tres primeras ontologías se mantienen por los dos algoritmos diseñados, pero en el caso de la ontología de Turismo el número de conceptos se incrementa de 292 a 387 esto indica que existen conceptos en el corpus que sólo se pueden encontrar por su correspondiente sinónimo y al SRIB sin expansión no le es posible encontrarlo exactamente. También, la cantidad de oraciones asociadas a los SRIB con expansión se incrementa para las cuatro ontologías y más aún para la ontología de Turismo, reforzando nuevamente la existencia de los sinónimos de los conceptos encontrados en el corpus.

Tabla 3. Resultados del Sistema de Recuperación Booleano con expansión para el caso de las relaciones tipo class-inclusion de cada ontología de dominio.

Dominio	Ontología							SRI			
	OSC	SC	F	P	SCA	FA	PA	OSC	OSCA	OE	%
IA	205	205	0	1.00	205	0	1.00	782	824	42	5.37
SCORM	1,038	1,006	32	0.969	1,006	32	0.969	10,624	10,784	160	1.50
OIL	37	32	5	0.864	32	5	0.864	12,691	12,699	8	0.063
Turismo	1,016	292	724	0.287	387	629	0.380	4,886	19,520	14,634	299.5

En el caso de las relaciones tipo no taxonómicas, que sólo las ontologías IA y SCORM tienen, se observa que la cantidad de relaciones recuperadas es la misma para ambos sistemas. Sólo se incrementaron algunas oraciones en las cuales existen el sinónimo correspondiente a cada concepto que forma la relación (ver Tabla 4).

Tabla 4. Relaciones no taxonómicas.

Dominio	Ontología							SRI			
	ONT	NT	F	P	NTA	FA	PA	ONT	ONTA	OE	%
IA	61	61	0	1.000	61	0	1.000	106	121	15	14.15 %
SCORM	759	744	15	0.980	744	15	0.980	8,752	9,589	837	9.56 %

5.3. Análisis de resultados

La aproximación propuesta, sistema de recuperación booleano con expansión semántica por sinónimos, recupera más información que lo que se obtiene con el sistema de recuperación booleano tradicional (ver columna % de cada tabla). La necesidad de incorporar sinónimos en la expansión se debe a que estos son considerados en una de las etapas de diseño de ontologías y el SRIB tradicional no logra identificar los conceptos exactos en el corpus, pero en base a los resultados se observa que el sinónimo correspondiente mantiene una relación semántica con evidencia en el corpus, dando la posibilidad de encontrar más relaciones existentes en la ontología y en el corpus de dominio.

Una de las limitaciones que se identifica en la aproximación es que el recurso semántico (WordNet) no es heterogeneo, es decir, no se obtienen sinónimos para cualquier tipo de dominio. Por lo tanto, se considera el uso de otras alternativas para la extración de sinónimos en el corpus, como es el caso del uso de patrones léxico-sintácticos.

6. Conclusiones

En este artículo se presenta una aproximación que realiza la expansión de consultas con el uso de sinónimos. Las consultas están formadas por los conceptos extraídos de las ontologías de dominio, la aproximación propuesta utiliza un SRIB. En base a los resultados experimentales se observa que la expansión permite recuperar más información del corpus de dominio. En algunos casos el SRIB con expansión permite recuperar más conceptos e información asociada a estos conceptos desde el corpus, al añadir los sinónimos correspondientes obtenidos desde WordNet. En algunas ontologías la cantidad de oraciones recuperadas supera significativamente al SRIB sin expansión. Como trabajo a futuro se propone el diseño de otro algoritmo de expansión que considere el uso de sinónimos por cada palabra que integra al concepto. Se considera que esa propuesta facilitará la incorporación de más información a procesar por cada concepto. También como consecuencia de este tipo de expansión consideramos la propuesta de extensión de las ontologías de dominio al incluir la relación semántica de tipo sinonimía.

Referencias

1. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)

2. Chauhan, R., Goudar, R., Rathore, R., Singh, P., Rao, S.: Ontology based automatic query expansion for semantic information retrieval in sports domain. In: Eco-friendly Computing and Communication Systems, pp. 422–433. Springer (2012)
3. Cotelo, S., Makowski, A., Chiruzzo, L., Wonsever, D.: Búsqueda de documentos utilizando criterios semánticos (2012)
4. Gil, R.M.n., Aparicio, F., de Buenaga, M.: Sistema de acceso a la información basado en conceptos utilizando freebase en español-inglés sobre el dominio médico y turístico. Procesamiento del lenguaje natural 49, 29–38 (2012)
5. Hany, M.H., Khaled, M.F., Nagdy, M.N.: Recuperación semántica enfocada en documentos web. International Journal of Advanced Computer Science and Applications (2011)
6. Hernández-Aranda, D., Granados, R., García-Serrano, A.: Servicios de anotación y búsqueda para corpus multimedia. Procesamiento del Lenguaje Natural 49, 213–216 (2012)
7. Korfhage, R.R.: Information storage and retrieval (2008)
8. Kuna, H.D., Rey, M., Podkowa, L., Martini, E., Solonezen, L.: Expansión de consultas basada en ontologías para un sistema de recuperación de información. In: XVI Workshop de Investigadores en Ciencias de la Computación (2014)
9. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
10. Neha, S., others: Mejora de la consulta con coeficiente de czekanowski por expansión usando algoritmos genéticos. International Journal of Computer Science and Information Technologies (2014)
11. Porter, M.F.: Readings in information retrieval. chap. An Algorithm for Suffix Stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
12. Salton, G., McGill, M.J.: Introduction to modern information retrieval (1986)
13. Schneider, J.M., Declerck, T., Fernández, J.L.M., Martínez, P.: Prueba de concepto de expansión de consultas basada en ontologías de dominio financiero. Procesamiento del lenguaje natural 51, 109–116 (2013)
14. Shabanzadeh, M., Nematbakhsh, M.A., Nematbakhsh, N.: A semantic based query expansion to search. In: 2010 International Conference on Intelligent Control and Information Processing (ICICIP). pp. 523–528. IEEE (2010)
15. Tolosa, G.H., Bordignon, F.R.: Introducción a la recuperación de información (2008)
16. Tovar Vidal, M.: Evaluación automática de ontologías de dominio restringido. Ph.D. thesis, Cenidet (2015)
17. Tovar Vidal, M., Pinto Avendaño, D., Montes Rendón, A., González Serna, J.G., Vilariño Ayala, D.: Evaluation of ontological relations in corpora of restricted domain. Computación y Sistemas 19(1) (2015)
18. Valbuena, S.J., Londoño, J.M.: Búsqueda de documentos basada en el uso de índices ontológicos creados con mapreduce document search supported on an ontological indexing system created with mapreduce. Ciencia e Ingeniería Neogranadina 24(2), 57 (2014)
19. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Blomqvist, E., Gangemi, A., Hammar, K., del Carmen Suárez-Figueroa, M. (eds.) WOP. CEUR Workshop Proceedings, vol. 929. CEUR-WS.org (2012)