Aproximaciones para la expansión semántica de consultas de un Sistema de Recuperación de Información Booleano

Ana Laura Lezama, Mireya Tovar, Darnes Vilariño, David Pinto

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

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

Resumen. En el presente trabajo se proponen dos aproximaciones para la expansión de consultas de un Sistema de Recuperación de Información Booleano (SRIB), con la finalidad de mejorar los niveles de precisión de un SRIB sin expansión. Las consultas están formadas por las palabras que integran a los conceptos y las relaciones semánticas de cuatro ontologías de dominio. El propósito de estas dos aproximaciones consiste en recuperar información relevante del corpus de dominio de cada concepto y relación de la ontología de dominio. Analizando los resultados que se obtuvieron en los experimentos, se observa que la precisión del SRIB con la segunda aproximación mejora los resultados de la primera aproximación del mismo SRIB y también al SRIB sin expansión.

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

Approaches to the Semantic Expansion of Query in a Boolean Information Retrieval System

Abstract. In this research work we propose two approaches for query expansion aiming to improve the performance of a Boolean Information Retrieval System (BIRS). Queries are made up of words and integrate the concepts and semantic relationships retrieved from four ontologies of restricted domain. The final purpose is to retrieve relevant information from the domain corpus for each concept and relationship of the ontology of restricted domain. By analyzing the results obtained in the experiments, it can be observed that the precision of the second approach proposed improves the results of the first approximation both, with and without the expansion process.

Keywords: Information retrieval system, semantic expansion of query, ontology.

1. Introducción

La Recuperación de Información *(RI)* es un campo relacionado y se centra con la estructura, almacenamiento, organización y búsqueda de elementos de información [3,12]. Tal proceso debería dar al usuario la información relevante a su necesidad de información, sin embargo, existen problemas demasiado significativos en cuanto a las consultas ingresadas por el usuario, que dificultan a un SRI recuperar toda la información relevante en cuanto a la consulta ingresada.

La función de un SRI no es la de devolver la información solicitada por el usuario, sino sólo indicar qué documentos son potencialmente relevantes para la consulta ingresada [3]. 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 Tovar [14] emplean un Sistema de Recuperación de Información Booleano y la información recuperada es utilizada para la evaluación automática de ontologías de dominio. Con la finalidad de mejorar la precisión del sistema mencionado, proponen la extensión del mismo. En la primera aproximación realizada para el SRIB mencionado, se extrajeron los sinónimos asociados a las consultas o conceptos completos que entran al sistema [16]. Los sinónimos son extraídos desde WordNet [7]. En la segunda aproximación las consultas están formadas por los sinónimos de cada palabra que integra a los conceptos o consultas que también son extraídos desde WordNet [7].

Esta investigación está estructurada de la siguiente manera: en la subsección 1.1 se describe la información general sobre sistemas de recuperación de información, en la sección 2 se presentan algunas propuestas por diversos autores para la expansión de consultas, en la sección 3 se puede visualizar un algoritmo general que contiene a la primera aproximación 3.1, y la segunda aproximación 3.2, en la sección 4 se presentan los experimentos y el conjunto de datos y finalmente en la sección 5 se describen las conclusiones.

1.1. Sistemas de Recuperación de Información

Los sistemas de recuperación de información, a menudo son comparados con las bases de datos relacionales. Tradicionalmente, los sistemas de recuperación de información, tienen información recuperada de textos no estructurados, lo que quiere decir que es texto en lenguaje natural. La diferencia fundamental entre bases de datos y sistemas de recuperación, es que las bases de datos son diseñadas para consultas de datos relacionales y que tienen conjuntos de archivos predefinidos, y los sistemas de recuperación de información, poseen un modelo de recuperación, un índice invertido o *postings lists*, es decir, un diccionario de términos, que nos indica el número de línea donde se encuentra el término, entre otros [6].

1.2. Sistemas de Recuperación de Información con Expansión de Consultas

La expansión de consultas o *(Query Expansion)* es la técnica comúnmente usada en Recuperación de Información, para mejorar el desempeño de los resultados por reformulación de la consulta original, ya sea añadiendo nuevos términos o reponderación de los términos originales [13].

Los términos de la expansión de consultas pueden ser automáticamente extraídos de los documentos, o tomándolos de recursos de conocimiento, como tesauros, ontologías léxicas como WordNet [7], algoritmos genéticos, etc. La ventaja de dichas técnicas es la expansión de términos que son extraídos de la colección [13], como los descritos en la sección 2.

1.3. Sistemas de Recuperación de Información sin Expansión de Consultas

Los SRI sin expansión de consultas consisten en que el usuario plasma su necesidad de información en una consulta aceptada por un SRI, por su parte el SRI transformará dicha consulta en una representación interna que permita su comparación con los documentos indexados.

La consulta supone un intento por parte del usuario de especificar las condiciones que permitan acotar dentro de la colección aquel subconjunto de documentos que contienen la información que desea. Por lo tanto, el SRI parte de la consulta formulada por el usuario, no de la necesidad de información original, por lo que una formulación incorrecta o insuficiente no podrá guiar adecuadamente al SRI durante el proceso de búsqueda. A este respecto los mayores problemas a los que ha de hacer frente el SRI son, por una parte, la escasa habilidad del usuario a la hora de formular su necesidad en forma de consulta y, por otra parte, que a la hora de describir un mismo concepto los términos empleados por el usuario y los autores de los documentos suelen diferir, impidiendo el establecimiento de correspondencias [17].

Uno de los modelos existentes, para la recuperación de información, es el modelo booleano que es uno de los métodos más utilizados para la recuperación de información [6]. Este modelo se basa en la agrupación de documentos, los cuales están compuestos por conjuntos de términos y en la concepción de las preguntas como expresiones booleanas, de ahí deriva el nombre de modelo de recuperación booleano. Es un modelo de recuperación simple, basado en la teoría de conjuntos y el álgebra booleana. Se denomina Álgebra de Boole o álgebra booleana a las reglas algebraicas, basadas en la teoría de conjuntos, para manejar ecuaciones de lógica matemática. Se denomina así en honor de George Boole, famoso matemático, que la introdujo en 1847. Dado su inherente simplicidad y su pulcro formalismo ha recibido gran atención y ha sido adoptado por muchos de los primeros sistemas bibliográficos comerciales. Su estrategia de recuperación está basada en un criterio de decisión binario (pertinente o no pertinente) sin ninguna noción de escala de medida, sin noción de un emparejamiento parcial en las condiciones de la consulta.

2. Trabajos relacionados

En el caso de la expansión de consultas, 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 Schneider et. al. [10] plantean el uso de una ontología para mejorar los resultados de búsqueda en un SRI en un dominio en particular. Fue desarrollada dentro del marco de su investigación, que se centró en el dominio financiero. En el que se distinguen dos capas, la primera que relaciona todas la entidades presentes en el mercado bursátil y la segunda que asigna metadatos a cada una de las entidades. La expansión de la consulta fue abordada desde dos puntos de vista, lanzando la consulta completa del usuario en lenguaje natural y el análisis semántico de la consulta para expandir únicamente las entidades. El análisis semántico de la consulta se realiza utilizando Textalytics[1]. La búsqueda en lenguaje natural se basa en la utilización conjunta de la ontología y Lucene. Para la evaluación de la búsqueda en la ontología, diseñaron la prueba de Cranfield y la evaluación de cada una de las consultas es por medio de precisión y recuerdo.

En Soni et. al. [11] 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 sea más eficiente, ya que mide la similitud entre los documentos recuperados y la consulta dada. Utiliza un analizador de texto que ayuda a encontrar palabras claves en los documentos, que serán utilizadas para hacer el cromosoma que es la base del algoritmo genético. La expansión de la consulta, es realizada en base al documento que tenga el cromosoma con el mejor valor en su función de aptitud, para después hacer la expansión manualmente, usando una medida de similitud. Observaron que el uso de un algoritmo genético aumenta la relevancia de documentos recuperados, si la tasa de mutación es menor el cromosoma converge en una sola generación.

En Harb et. al. [4] usaron un rastreador que debe recorrer la WWW para obtener documentos en el dominio del cuidado de la salud, específicamente en enfermedades ictéricas. El modelo de espacio vectorial es adaptado en la propuesta de este trabajo para la representación de documentos, retira palabras vacías, etc. La consulta es expandida por sinónimos extraídos de WordNet, pero sólo con aquellos sentidos más comunes de cada término de la consulta. Con el método de recuperación de información semántico propuesto se han aprovechado las ventajas de la web semántica para recuperar documentos pertenecientes al dominio mencionado. Supera el método de recuperación de información clásica y demuestra mejoras en el rendimiento.

En Mahgoub et. al. [5] introducen una aproximación de expansión de consultas usando una ontología construída con páginas de Wikipedia, además de otros tesauros para mejorar la precisión en la búsqueda del idioma árabe. Su aproximación, depende de tres recursos árabes que son Wikipedia en árabe, como el recurso con mayor información semántica, el diccionario Al Raed, que es un dic-

[1] http://textalytics.com

cionario monolingüe para palabras modernas, y el diccionario Google_WordNet que es una colección de todas las palabras en WordNet y traducidas con el traductor de Google. La indexación y recuperación de su sistema depende de Lucene. Para expandir la consulta, primero localizan el nombre de las entidades o conceptos que aparecen en la consulta , si el nombre de una entidad o concepto es localizado, agregan el título de redirigir la página que conduce al concepto similar agregando una subcategoría del sistema. Para sus experimentos, usaron el conjunto de datos construídos desde el libro "Zad Al Ma'ad", dicho conjunto de datos contiene 25 consultas y 2,730 documentos.

En Fernández et. al. [2] muestran aspectos relacionados con la integración de la tecnología disponible del tratamiento del lenguaje natural en el desarrollo de un metabuscador que alcance un mayor grado de acierto en la recuperación de información realizada por un buscador tradicional así como en el tratamiento posterior de los documentos recuperados. Describen su proceso realizado, para la expansión de las consultas de los usuarios, con información lingüística empleando dos recursos léxicos para el castellano: ARIES que es un léxico morfológico desarrollado por la Universidad Politécnica de Madrid y la Universidad Autónoma de Madrid para el tratamiento de la morfología y EuroWordNet [18] para el tratamiento de la semántica. La generación de la consulta está compuesta por dos tareas principales, la primera consiste en transformar la consulta del usuario en lenguaje natural (LN) en una consulta formal que el buscador pueda ejecutar. La segunda funcionalidad consiste en extender los términos significativos de la consulta (formal) utilizando conocimiento lingüístico; para ello se añaden a los términos significativos de la consulta (enlazados con AND) las variantes morfológicas y semánticas mediante OR con el fin de construir una consulta en forma normal conjuntiva. Su trabajo forma parte del sistema MESIA, modelo computacional para extracción selectiva de información de textos cortos, que amplía la búsqueda habitual (consulta y presentación de resultados) con nuevas capacidades morfológicas y semánticas y analiza otros aspectos obtenidos a partir de la estructura de las páginas, del tratamiento lingüístico de algunas de las unidades de texto seleccionadas automáticamente y de la experiencia de uso.

En Cruanes et. al. [1] proponen una aproximación de mapeo de información en lenguaje natural del dominio de enfermería, utilizando métodos de similitud léxica. Los autores generan expansión por sinónimos y buscan antonímia. No usaron recursos como EuroWordNet, ya que de acuerdo a los autores, no se ajustaba a las necesidades del dominio estudiado.

En Deco et. al. [8] proponen un refinamiento semántico, que guiará al usuario a desambiguar los términos ingresados por el. Realizaron expansión semántica de consultas por sinónimos, usaron WordNet, y en la generación de la estrategia, ocuparon operadores lógicos.

En la Tabla 1 se presenta un resumen de los trabajos revisados anteriormente. En esta tabla se obsevan los recursos léxicos, los dominios, el tipo de expansión y el sistema de recuperación de información que cada autor usó en su investigación.

En esta investigación se propone la expansión de las consultas ingresadas a un SRIB, dicha expansión a diferencia de algunos trabajos del estado del arte, que

Tabla 1. Estado del arte de Sistemas de Recuperación de Información con expansión de consultas

Autores	Recursos léxicos	Dominios	Expansión de consultas	Tipo SRI
[10]	Textalytics	Financiero	Ontología	Lucene
[11]	Analizador de texto	-	Algoritmo genético	MEV
[4]	Analizador de texto	Cuidado de salud	Sinónimos	MEV
[5]	WordNet	Idioma Árabe	Ontología de dominio	Lucene
[2]	ARIES/EuroWordNet	Festivales	Sinonimia/Hiponimia	Lucene
[1]	Métodos de similitud léxica	Enfermería	Sinónimos	-
[8]	WordNet	Cuidado de la salud	Sinónimos	Booleano

realizan la expansión de consultas por medio de algoritmos genéticos, ontologías, etc. se realiza extrayendo los sinónimos de WordNet, y después usando los sinónimos extraídos, se lleva a cabo la expansión de las consultas por medio de las dos aproximaciones presentadas en este documento.

3. Propuesta

En esta sección se presenta un algoritmo general, que contiene a las dos aproximaciones de expansión propuestas, para el SRIB.

1. Extracción de conceptos y relaciones de las ontologías de dominio.
2. Extracción de los sinónimos con WordNet, esta etapa se encuentra dividida en dos procesos diferentes:
   a) Primera Aproximación
      1) Extracción de los sinónimos, de los conceptos completos con WordNet.
   b) Segunda Aproximación
      1) Se retiran palabras cerradas.
      2) Extracción de los sinónimos, de cada palabra que integran al concepto, sin palabras cerradas con 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 [9].
4. Formación de consultas. Existen tres tipos de consultas para las tres aproximaciones propuestas:
   a) Primera Aproximación
      1) Consultas formadas con las palabras del concepto.
      2) Consultas formadas con los sinónimos del concepto.
      3) Consultas formadas con las palabras del concepto que forman la relación semántica.
   b) Segunda Aproximación

   1) Consultas formadas con las palabras del concepto.
   2) Consultas formadas con los sinónimos de cada palabra que integra al concepto, y los conceptos originales.
   3) Consultas formadas con las palabras del concepto que forman la relación semántica.
5. Aplicación del Sistema de Recuperación de Información Booleno (SRIB) para conceptos, sin expansión.
6. Aplicación del Sistema de Recuperación de Información Booleno (SRIB) para los sinónimos de los conceptos, con expansión.
7. Mezcla o unión de los resultados obtenidos (posting) por el SRIB para resultados sin sinónimos y con sinónimos de los dos pasos anteriores.
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.

En el caso de la evaluación de los resultados obtenidos, se utilizan las Ecuaciones (1) y (2) para medir la precisión a nivel de conceptos y relaciones:

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

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

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 taxonómicas y las relaciones no taxonómicas (para más información ver [15]). El *Total relaciones* corresponden al total de relaciones de cada tipo existentes en la ontología de dominio evaluadas de manera independiente. A continuación se presenta brevemente cada aproximación propuesta.

3.1. Primera aproximación

La primera aproximación, realiza la extracción de sinónimos de los conceptos o consultas, completos de cada ontología de dominio, y después se hace uso del algoritmo de unión o mezcla de los documentos recuperados por el SRIB sin expansión y con expansión, y posteriormente la evaluación del mismo. En la Figura 1 se muestra el comportamiento de manera gráfica del algoritmo general, incluyendo únicamente la primera aproximación.

3.2. Segunda aproximación para la expansión de consultas

En la segunda aproximación se plantea la expansión de la consulta, al incorporar los sinónimos correspondientes a cada palabra que forman al concepto de la ontología. En la Figura 2 se observa de manera gráfica los pasos que

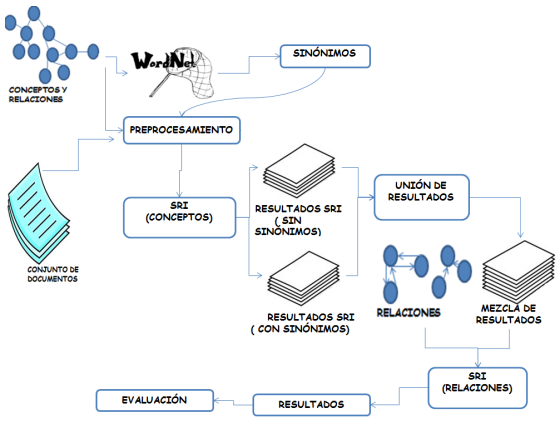

Fig. 1. Primera aproximación para la expansión de consultas en un SRIB

se siguieron en el algoritmo general incluyendo sólo la segunda aproximación. La figura incluye la búsqueda de los sinónimos de cada palabra que integran al concepto (sin cerradas en WordNet), así como la generación de las nuevas consultas procesadas por el SRIB, la unión de los resultados del SRIB con expansión y sin expansión.

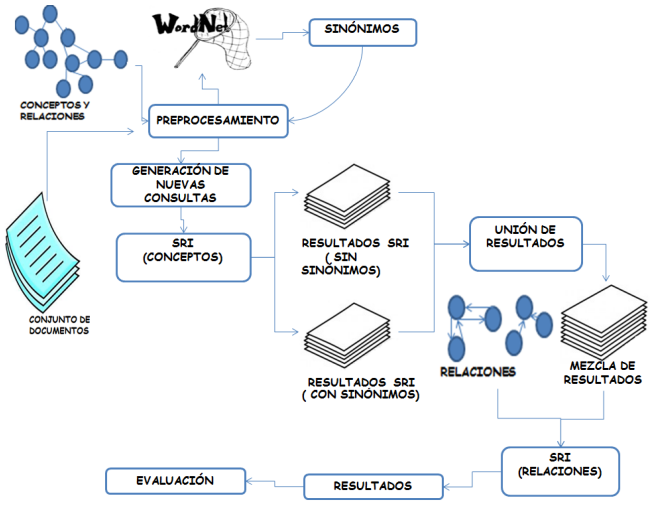

Fig. 2. Segunda aproximación para la expansión de consultas en un SRIB

4. Resultados experimentales

En esta sección, se presentan los datos utilizados (4.1) y los resultados obtenidos en los experimentos (4.2).

4.1. Conjunto de datos

En la Tabla 2 se presenta el número de conceptos (C), el total de relaciones taxonómicas (T) y el total de relaciones no taxonómicas (NT) de las ontologías evaluadas. También se incluye el número de documentos (D), el número de tokens (T), la 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 2. Conjunto de datos

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

4.2. 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, para las dos aproximaciones presentadas.

Primera aproximación 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 3 para cada ontología revisada (Dominio). En la Tabla 3 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 (CE), los conceptos que no logró recuperar el SRIB con expansión (FE) y la precisión obtenida (PE).

Además, en la tabla se incluye la cantidad de oraciones obtenidas por el SRIB sin expandir (O), con expansión (OE), la diferencia del número de líneas recuperadas con expansión y sin ella (DI) 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 en número de líneas, 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 expansión.

Tabla 3. Resultados de la primera aproximación, del Sistema de Recuperación Booleano con expansión para el caso de los conceptos de cada ontología de dominio

Dominio	Ontología							Corpus			
	CO	C	F	P	CE	FE	PE	O	OE	DI	%
IA	276	274	2	0.992	274	2	0.992	1,994	3,057	1,063	53.30
SCORM	1,461	1,434	27	0.981	1,435	26	0.982	23,406	31,093	7,687	32.84
OIL	48	48	0	1.00	48	0	1.00	232,603	295,986	63,383	27.24
Turismo	963	682	281	0.708	711	252	0.738	86,353	224,764	138,411	160.28

En la Tabla 4 se presentan los resultados obtenidos por ambos Sistemas de Recuperación de Información con expansión y sin ella, para relaciones de tipo taxonómicas de cada ontología de dominio. La columna RT corresponde al total de relaciones taxonómicas incluidas en la ontología de dominio correspondiente. La columna RR es el total de relaciones recuperadas con información del SRI sin expansión y con expansión RRE. La columna correspondiente a F es la diferencia de las relaciones recuperadas por el SRI booleano sin expansión y con expansión (FE). La precisión del sistema sin expansión (P) y con expansión (PE). También se incluye la cantidad de oraciones recuperadas en total por el SRIB sin expansión (O) y con expansión (OE) para este tipo de relaciones, la diferencia obtenida (DI) y el porcentaje de la diferencia (%).

En base a los resultados obtenidos se observa que el número de relaciones de tipo taxonómicas de las tres primeras ontologás se mantienen por los dos algoritmos diseñados, pero en el caso de la ontología de Turismo el número de relaciones se incrementa de 291 a 386 esto indica que existen relaciones 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 las relaciones encontradas en el corpus.

Tabla 4. Resultados de la primera aproximación, del Sistema de Recuperación Booleano con expansión para el caso de las relaciones taxonómicas de cada ontología de dominio

Dominio	Ontología							SRI			
	RT	RR	F	P	RRE	FE	PE	O	OE	DI	%
IA	205	205	0	1.00	205	0	1.00	782	876	94	12.02
SCORM	1,038	1,002	36	0.965	1,002	36	0.965	10,640	10,926	286	2.68
OIL	37	32	5	0.864	32	5	0.864	12,696	12,704	8	0.063
Turismo	1,016	291	725	0.286	386	630	0.379	5,606	20,198	14,592	260.29

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. La columna RNT corresponde, al número de relaciones no taxonómicas incluídas en cada ontología de dominio correspondiente. Las columnas R y RE, corresponde al número de relaciones obtenidas por el SRI sin expansión, y con expansión, respectivamente. Las columnas FE y F, corresponden a la diferencia de las relaciones recuperadas por el SRI con expansión y sin expansión, respectivamente. Las columnas P y PE, corresponde a la precisión del sistema sin expansión y con expansión respectivamente. Las columnas O y OE, corresponde a la cantidad de oraciones recuperadas por el SRI sin expansión y con expansión respectivamente y la columna DI, corresponde a la diferencia de oraciones recuperadas del sistema con expansión y el sistema sin expansión. En este caso, sólo se incrementaron algunas oraciones en las cuales existen el sinónimo correspondiente a cada concepto que forma la relación (ver Tabla 5).

Tabla 5. Resultados de la primera aproximación, para relaciones no taxonómicas

Dominio	Ontología							SRI			
	RNT	R	F	P	RE	FE	PE	O	OE	DI	%
IA	61	61	0	1.000	61	0	1.000	108	136	28	25.92 %
SCORM	759	738	21	0.972	738	21	0.972	8,728	9,655	927	10.62 %

Segunda aproximación 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, con la diferencia de la primera aproximación de que los conceptos o consultas ingresados en el SRIB, son ahora nuevas consultas creadas con los sinónimos extraídos de WordNet. Los resultados obtenidos por ambos algoritmos, para el caso de los conceptos, se muestran en la Tabla 6

para cada ontología revisada (Dominio), siguiendo la nomenclatura utilizada en la Tabla 3.

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 33 %, 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 expansión.

Tabla 6. Resultados de la segunda aproximación, del Sistema de Recuperación Booleano con expansión para el caso de los conceptos de cada ontología de dominio

Dominio	Ontología							Corpus			
	CO	C	F	P	CE	FE	PE	O	OE	DI	%
IA	276	274	2	0.992	274	2	0.992	1,994	3,325	1,331	66.75
SCORM	1,461	1,434	27	0.981	1,436	25	0.982	23,406	35,987	12,581	53.75
OIL	48	48	0	1.00	48	0	1.00	232,603	310,067	77,464	33.30
Turismo	963	683	281	0.708	784	179	0.814	86,353	227,451	141,098	163.39

En la Tabla 7 se presentan los resultados obtenidos por ambos Sistemas de Recuperación de Información con expansión y sin ella, para relaciones de tipo taxonómicas de cada ontología de dominio. En base a los resultados obtenidos se observa que el número de relaciones de tipo taxonómicas de las dos primeras ontologías se mantienen por los dos algoritmos diseñados, pero en el caso de la ontología de OIL el número de conceptos aumenta en uno, mientras que para la ontología de Turismo el número de conceptos se incrementa de 291 a 433 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.

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 oraciones en las cuales existen el sinónimo correspondiente a cada concepto que forma la relación (ver Tabla 8).

Tabla 7. Resultados de la segunda aproximación, del Sistema de Recuperación Booleano con expansión para el caso de las relaciones taxonómicas de cada ontología de dominio

Dominio	Ontología							SRI			
	RT	RR	F	P	RRE	FE	PE	O	OE	DI	%
IA	205	205	0	1.00	205	0	1.00	782	1089	307	39.25
SCORM	1,038	1,002	36	0.965	1,002	36	0.965	10,640	14,498	3,858	36.25
OIL	37	32	5	0.864	33	4	0.891	12,696	18,736	6,040	47.57
Turismo	1,016	291	725	0.286	433	583	0.426	5,606	22,823	17,217	307.11

Tabla 8. Resultados de la segunda aproximación, para relaciones no taxonómicas

Dominio	Ontología							SRI			
	RNT	R	F	P	RE	FE	PE	O	OE	DI	%
IA	61	61	0	1.000	61	0	1.000	108	149	41	37.96 %
SCORM	759	738	21	0.972	738	21	0.972	8,728	10,262	1,534	17.57 %

5. Conclusiones y trabajo futuro

En este artículo se presentan dos aproximaciones para la expansión de consultas en un Sistema de Recuperación de Información Booleano. La primera aproximación consiste en expandir utilizando los sinónimos del concepto exacto. La segunda aproximación realiza la expansión de consultas con el uso de sinónimos de cada palabra que integra a los conceptos. Las consultas están formadas por los conceptos extraídos de las ontologías de dominio. De acuerdo a los resultados experimentales se visualiza que la expansión de la segunda aproximación permite recuperar más información del corpus de dominio, en comparación con el SRIB sin expansión y con la primera aproximación realizada. En algunos casos el SRIB con expansión utilizando la segunda aproximación permite recuperar más conceptos, relaciones e información asociada a estos conceptos desde el corpus, al incorporar los sinónimos de las palabras que conforman a los conceptos desde WordNet. En algunas ontologías la cantidad de oraciones recuperadas supera significativamente al SRIB sin expansión, y a la primera aproximación. Como trabajo a futuro se propone el uso de patrones léxico-sintácticos para la extracción de relaciones tipo sinonimía desde el corpus de dominio, que permitan identificar los sinónimos de los conceptos de las ontologías de dominio.

Agradecimientos. Este trabajo de investigación esta parcialmente financiado por el número de proyecto VIEP 302 (2016), por el proyecto PRODEP (EXB-792) con número de convenio DSA/103.5/15/10854.

Referencias

1. Cruanes, J., Ferri, M.T.R., Pastor, E.L.: Análisis del uso de métodos de similitud léxica con conocimiento semántico superficial para mapear la información de enfermería en español. Procesamiento del lenguaje natural 49, 75–82 (2012)
2. Fernández, P.M., Serrano, A.G.: Utilizando recursos lingüísticos para mejora de la recuperación de información en la web. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial 6(16), 55–64 (2002)
3. Ferro, J.V., Nistal, J.L.F.: Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español. Procesamiento del Lenguaje Natural 36, 57–58 (2006)
4. Harb, H.M., Fouad, K.M., Nagdy, N.M.: Semantic retrieval approach for web documents. International Journal of Advanced Computer Science and Applications (IJACSA) 2(9), 67–76 (2011)
5. Mahgoub, A.Y., Rashwan, M.A., Raafat, H., Zahran, M.A., Fayek, M.B.: Semantic query expansion for arabic information retrieval. In: EMNLP: The Arabic Natural Language Processing Workshop, Conference on Empirical Methods in Natural Language Processing, Doha, Qatar. pp. 87–92 (2014)
6. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
7. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
8. Motz, R., Deco, C., Bender, C., Saer, J., Chiari, M.: Refinamiento semántico para recuperación de información desde la web. In: Proceedings Workshops on Artificial Intelligence, Iberamia. pp. 172–179 (2004)
9. 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)
10. 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)
11. Soni, N., Singh, J.: Relevancy enhancement of query with czekanowski coefficient by expanding it using genetic algorithm. International Journal of Computer Science and Information Technologies 5, 6106–6110 (2014)
12. Tolosa, G.H., Bordignon, F.R.: Introducción a la recuperación de información pp. 1–149 (2008)
13. Vechtomova, O., Wang, Y.: A study of the effect of term proximity on query expansion. Journal of Information Science 32(4), 324–333 (2006)
14. Vidal, M.T.: Evaluación automática de ontologías de dominio restringido. Ph.D. thesis, Cenidet (Febrero 2015)
15. Vidal, M.T., Avendaño, D.P., Rendón, A.M., Serna, J.G.G., Ayala, D.V.: Evaluation of ontological relations in corpora of restricted domain. Computación y Sistemas 19(1), 135–149 (2015)
16. Vidal, M.T., Sánchez, A.L.L., Ayala, D.V.n., Beltrán Martínez, B., Cardona, M.C.: Primera aproximación de un sistema de recuperación de información booleano con expansión semántica de consultas. Research in Computing Science 99, 55–63 (2015)
17. Vilares Ferro, J.: Aplicación del procesamiento del lenguaje natural en la recuperación de información en español. Ph.D. thesis, Universidad da Coruña, Departamento de Computación (Mayo 2005)

18. Vossen, P.: Introduction to eurowordnet. Computers and the Humanities 32(2), 73–89 (1998)
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)