Ciencias de la Información Vol. 45, No. 3, septiembre-diciembre, pp. 55 - 60, 2014

Desarrollo de un software para la detección automática de tópicos en documentos textuales basada en taxonomía

Patrick Pedreira Silva
Christian Freitas
Elvio Gilberto da Silva

En este trabajo propone un método que utiliza una taxonomía para la identificación automática de tópicos. La idea principal que subyace este artículo es la de aprovechar una estructura jerárquica taxonómica para encontrar tópicos de un texto. Las palabras clave que se extraen de un texto son mapeadas según sus correspondientes conceptos de la taxonomía. El método selecciona nodos de la taxonomía entre los conceptos nodos que son posibles en los tópicos del texto. Pese un vocabulario limitado, nuestra investigación ha obtenido una buena tasa de detección.

Palabras clave: taxonomía; IA; software; PLN

RESUMEN

ABSTRACT

This paper proposes a method of using taxonomy hierarchy in automatic topic identification. The fundamental idea behind this work is to exploit an taxonomy hierarchical structure in order to find a topic of a text. The keywords that are extracted from a given text will be mapped onto their corresponding concepts in the taxonomy. The method pick a single node among the concepts nodes that it believe is the topic of the target text. Despite, a limited vocabulary our topic identification has a good detection rate.

Keywords: taxonomy; AI; software; NLP

Introducción

La World Wide Web contiene miles de millones de documentos de las más diversas fuentes, que cubren todos los temas esenciales para el ser humano. Con esta gran cantidad de datos, muchas veces no estructurados, no es fácil acceder a la información, siendo necesarias técnicas de extracción automática de información para entender el contenido de los documentos. Conocer los tópicos puede, ayudar en la comprensión del contenido del documento. Un tópico (Villarreal et al, 2009), en un sentido abstracto, puede ser pensado como un objeto, una categoría o simplemente un tema. Algunas de las técnicas de detección automática de tópicos están basadas en la agrupación de palabras clave/documentos, considerando básicamente el contaje de palabras (Chakrabarti, 2003)(Yang y Qiu, 2011)(Kong et al, 2006). Otras técnicas explotan una estructura jerárquica ontológica/taxonómica con el fin de encontrar un tópico de un texto. El proceso asigna palabras clave, en las que conceptos considerados importantes son elegidos (Fang et al, 2007)(Coursey y Moen, 2009)(He et al, 2001).

Se puede definir una taxonomía como una especificación explícita de una conceptualización (Gruber, 1993). Habitualmente, una taxonomía es una jerarquía de conceptos en la cual en el nivel más alto se usan categorías principales que están asociadas con palabras o términos clave relacionados, los cuales ofrecen un refinamiento de términos de nivel superior. En el segundo nivel, se usan subcategorías, las cuales podrán tener por debajo un conjunto de conceptos que les permitan tener relaciones cruzadas con otros niveles de la jerarquía (Gruber, 1993).

El procesamiento semántico automático a través de las taxonomías es cada vez más común en los últimos años, ya que propician los medios para representar y utilizar el conocimiento del mundo. En aplicaciones que involucran el procesamiento de contenido textual, este conocimiento de mundo puede significar, por ejemplo, entender sobre el contenido de un texto. Taxonomías pueden servir para identificar

en un texto original sus tópicos principales con el fin de subsidiar a las más diversas áreas.

A causa de este potencial, este proyecto propone adoptar el uso de una taxonomía, como una de las estrategias para identificar tópicos y como apoyo para el desarrollo de un software que utiliza ese conocimiento. En este contexto, la taxonomía utilizada en este proyecto describe una jerarquía de palabras, que consta de vocabularios de representación de conceptos, que procuran condiciones potenciales para describir el conocimiento de un dominio, que se utilizan como indicadores de información relevante. Los conceptos se van a expresar por medio del lenguaje natural, utilizando términos específicos, es decir, palabras que, cuando se encuentran, indican la presencia del concepto (Loh, 2001) en los documentos.

La herramienta que se desarrolló en esta investigación, se le ha llamado EXTRATOP (EX-TRA acción automática de tópicos), la cual combina características lingüísticas y estadísticas para la detección automática de los tópicos de un texto en portugués. La hipótesis que orienta este trabajo es el hecho de que la información semántica recuperada de una taxonomía permite que el sistema determine qué tópicos son relevantes para indicar el contenido que caracteriza un documento. El sistema posibilita la identificación de tópicos por el contaje de conceptos, utilizando la taxonomía de Yahoo enriquecida y adaptada de (Pedreira-Silva, 2006).

Materiales y métodos

El primer paso en el desarrollo de este trabajo ha consistido en un estudio de las tecnologías disponibles para la especificación, la representación de la taxonomía y utilización del algoritmo de selección de tópicos. El sistema posibilita la identificación de tópicos por el contaje de conceptos, utilizando la taxonomía de Yahoo enriquecida y adaptada de (Pedreira-Silva, 2006).

Las principales categorías o conceptos de Yahoo incluyen: Artes y cultura, deportes, educación, ciencia, fuentes de referencia Regionales, Business to Business, salud, compras y servicios, ocio, computadoras, Internet, noticias, finanzas, gobierno y sociedad. Cada concepto se describe por un conjunto de palabras clave que lo caracteriza. Las palabras clave, a su vez, esbozarán un camino que indica la posición del concepto en la jerarquía. En otras palabras, un subconcepto se describe mediante la adición de una palabra clave para el conjunto de palabras clave que caracterizan a su súper concepto. Teniendo en cuenta esta sucesión de asignación de conceptos a cada nodo de la jerarquía, todos los nodos de la jerarquía heredarán de forma decreciente los conceptos de sus sucesores. Un ejemplo de camino con cinco conceptos interrelacionados se indica mediante «>>». El superconcepto, es en este caso, artes y cultura y el más elemental subconcepto u hoja es Bibi Ferreira:

Arte y cultura >> Artes escénicas >> Artistas actores y actrices >> Bibi Ferreira

Estos son los conceptos que son utilizados por la herramienta implementada como posibles tópicos de un documento en esta investigación. Sin embargo, como (Pedreira-Silva, 2006) describe la incorporación de la taxonomía de Yahoo para un sistema computacional no se da directamente, justificando el proceso de enriquecimiento realizado por el autor. Dicho proceso, que se describirá abajo, ha culminado en la generación de una base de datos taxonómica que ha sido incorporada en el esquema propuesto en esta investigación.

Se define como enriquecimiento, en el contexto de esta investigación, el proceso de describir un concepto de taxonomía de Yahoo a través de palabras del lenguaje natural en foco. La metodología aplicada por (Pedreira-Silva, 2006) ha envuelto la recopilación manual de vocabulario externo, usando como descriptores de palabras conceptos que tienen algún tipo de relación semántica con los conceptos taxonómicos. Estas relaciones pueden ser de varios tipos, por ejemplo, sinonimia, hiponimia y hiperonimia. Así que, intentamos garantizar un conjunto mínimo de 26.300 descriptores que se han añadido a esta taxonomía (Chakrabarti, 2003). Es necesario destacar que estos descriptores podrían estar compuestos por más de una palabra, como «Bibi Ferreira».

Actualmente, la taxonomía enriquecida por este proceso cuenta con aproximadamente 5.500 conceptos y unos 26.300 descriptores asociados con conceptos, utilizando la técnica descrita enriquecimiento (Pedreira-Silva, 2006). Cabe mencionar que el enriquecimiento se proyectó para 2.500 conceptos originales de la taxonomía de Yahoo (aproximadamente la mitad de la colección). Además, ha sido agregado manualmente un pequeño conjunto de nuevos descriptores para la taxonomía con el fin de actualizarla y hacerla más rica. Sin embargo, como el enfoque de esta investigación reside en la verificación del uso potencial de la taxonomía en tareas de detección automática de tópicos, no ha habido ninguna preocupación en completar esta colección a causa del esfuerzo para llevar a cabo manualmente este enriquecimiento (que se puede hacer en un segundo paso de esta investigación). Este repositorio se ha añadido a EXTRATOP.

Desarrollo

Después de recoger todo el conocimiento y la información relacionada con la ejecución de los trabajos relativos a la taxonomía, se ha comenzado la fase de desarrollo del software. Este desarrollo ha consistido principalmente en la implementación de una base de datos que contiene información de naturaleza taxonómica, el desarrollo del contorno del algoritmo puntuación e implementación de una interfaz Web (fig. 1).

El Software ha sido desarrollado a partir del uso del lenguaje de programación PHP, JavaScript y la base de datos MySQL. Se ha utilizado el software "sublime Text" como ambiente de desarrollo. La herramienta abarca la parte de interfaz de usuario y la definición del algoritmo de detección de tópicos a ser utilizado. Para su desempeño, una fuente de texto se da como entrada al sistema y para la detección de los tópicos, se realizan los siguientes pasos:

1. primero, el texto fuente es pre procesado, siendo eliminado la puntuación tradicional (por ejemplo, punto, signo de exclamación y el signo de interrogación);

2. se detectan conceptos subyacentes al texto, basados en taxonomía.

3. se eliminan en el texto palabras vacías (stopwords) y, entonces, se seleccionan los tópicos y se clasifican según su proximidad a los conceptos de taxonomía.

4. por último, aparecen como respuesta final los tópicos que indican el contenido principal del documento.

El primer paso del método de detección de

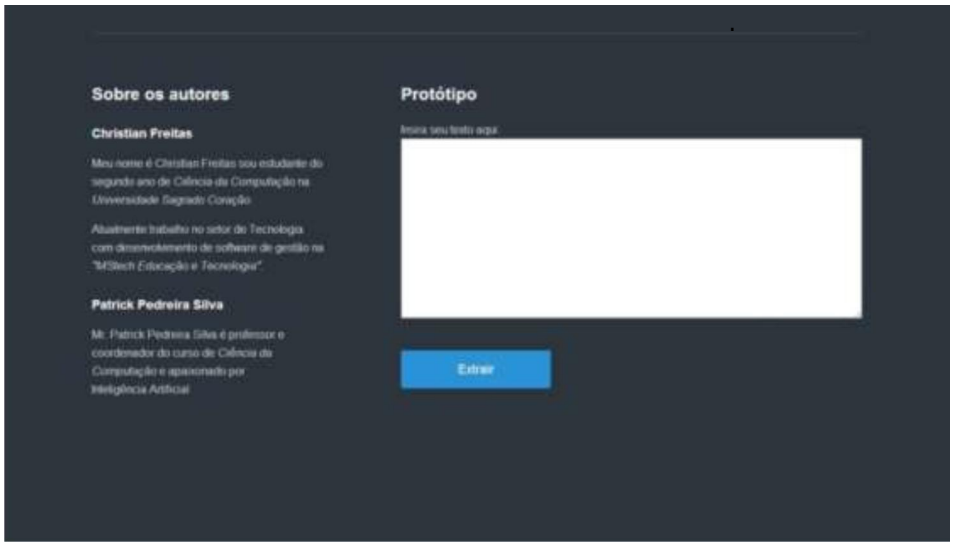


Figura 1. Pantalla de entrada de datos del prototipo

El primer paso del método de detección de tópicos es determinar aquellos que son más importantes en el documento, estimando su relevancia. Esto se hace verificando si las palabras presentes en el texto corresponden a las que describen los conceptos taxonómicos. Siempre que se produce esta coincidencia, se asume que ese concepto está en el texto y representa, por lo tanto, uno de sus tópicos. El proceso de asignación se produce después de que el pre procesamiento del texto - que incluye la eliminación de las supuestas palabras vacías que corresponden a las palabras del portugués consideradas irrelevantes (a, o, de, da,...). Específicamente para esta investigación se ha elaborado una lista con 255 palabras.

Como un factor de discriminación de la importancia de los conceptos, se calcula, inicialmente, el peso de todos ellos. Para calcular el peso de un concepto, se basa en la frecuencia de palabras en el documento, que se asignan con el concepto, o sea, que se corresponden con los descriptores de los conceptos. La frecuencia se identifica mediante el recuento absoluto de las ocurrencias de la palabra en el documento. Calcular el peso de un concepto, basado en la frecuencia de palabras, sigue la suposición de que la repetición de palabras en un texto se hace con el fin de destacar algún tópico y puede ser un indicador de la significación de las palabras (Tiun et al, 2001).

En EXTRATOP, donde una palabra del texto corresponde a un descriptor de un concepto, el peso de este concepto se incrementa en 1 unidad. Este proceso es acumulativo, es decir, cuando la misma palabra aparece en el texto, se debe agregar 1 unidad. Teniendo en cuenta la estructura jerárquica de la taxonomía y las relaciones entre los conceptos, la detección de un concepto subyacente del documento implica indirectamente la presencia de su "concepto-padre" también en el documento (Tiun et al, 2001).

Para tratar de dar forma a este proceso de generalización en la que se detecta un concepto, partir de sus "conceptoshijos", se ha implementado un proceso de calificación de tópicos que considera que cada vez se incrementa el peso de un concepto debido a la asignación de una palabra, también se incrementa el peso de su concepto-padre. Cabe hacer hincapié en el contexto que se ha considerado en esta investigación, ya que el proceso de generalización es solamente entre el concepto-hijo y su concepto-padre (contexto inmediato). Así, cada vez que un peso se propaga del conceptohijo al concepto-padre, se reduce su valor.

En EXTRATOP, la propagación de puntos para el padre del concepto coincidirá con el 50% del peso que se ha obtenido por concepto de hijo. Este valor ha sido elegido empíricamente y esta reducción ha sido incluida en el algoritmo de la herramienta.

Resultados y discusión

Como forma de probar el funcionamiento del prototipo en la tarea de detección de tópicos relevantes se ha desarrollado un experimento que contó con la participación de 40 voluntarios. El experimento consistió en presentar, para cada voluntario, un conjunto de 5 textos cortos, cada uno sobre un tema específico (fútbol, música, educación, economía y política). Asociado a cada texto había una lista con 5 ó 6 tópicos (definidos automáticamente por la herramienta como relevantes) que se debía analizar y marcar por los voluntarios si los mismos entendieran que representaban tópicos relacionados con los textos.

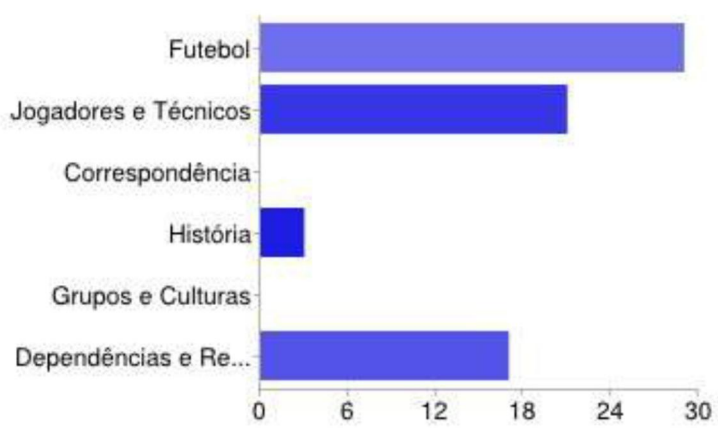
Para evaluar las respuestas del prototipo se ha comprobado, para cada texto, cuál es el nivel de acuerdo con los tópicos presentados. A pesar de ser un simple experimento, se hace útil para mostrar, en un primer momento, el potencial del algoritmo implementado. Para realizar este análisis se definió definida como medida de precisión del sistema, la relación entre los 6 temas identificados por el EXTRATOP para cada texto y los que efectivamente han sido referenciados por los voluntarios. La fórmula siguiente muestra cómo se ha calculado esta relación:

Precisión = (número de tópicos valorados como relevantes por el voluntario) / (número de tópicos considerados relevante por el prototipo).

En este sentido, cuanto mayor sea la precisión (valor más cercano a 1) mayor será la posibilidad de que el sistema indique buenos tópicos para el texto.

La Fig. 2 muestra los datos que se han recogido sobre un texto relacionado con el tema del fútbol. En la figura se han registrado las indicaciones recibidas cada uno de los temas que se presentaron. Considerando que, por lo tanto, en el primer texto se observa que cuatro de los seis tópicos han sido también indicados por los voluntarios como relevantes. De ese modo, la precisión del sistema ha sido alrededor del 67%. Una observación interesante es que «fútbol» y «los jugadores y entrenadores» han sido los tópicos considerados más relevantes por los voluntarios (41% y 30%, respectivamente) y estos son los tópicos más relevantes señalados por EXTRATOP.

La Fig. 3 trae datos acerca de un texto sobre música. Según la figura, cuatro de los seis tópicos fueron señalados también por los voluntarios como relevantes. De esta manera, al igual que ha pasado con el texto 1, la precisión del sistema ha sido del 67%.



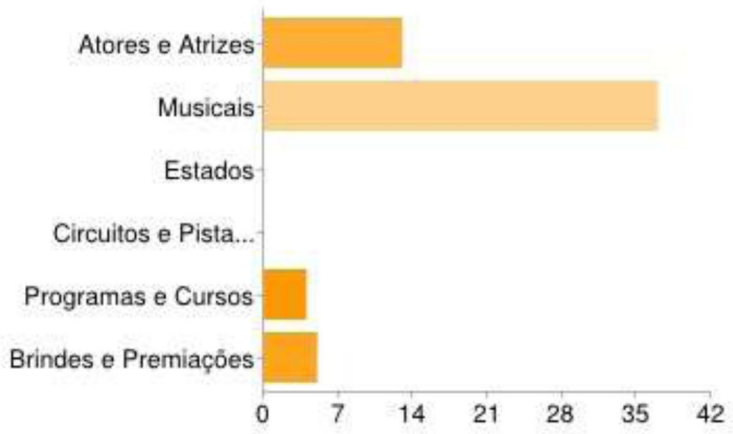
Futebol	29	41%
Jogadores e Técnicos	21	30%
Correspondência	0	0%
História	3	4%
Grupos e Culturas	0	0%
Dependências e Recuperação	17	24%

Figura 2. Datos de texto sobre fútbol

De la misma manera que ha ocurrido con el otro texto, había una correspondencia entre los dos tópicos más relevantes indicados por los voluntarios y aquellos relevantes apuntados por EXTRATOP, en el caso «actores y actrices» y «Musicales».

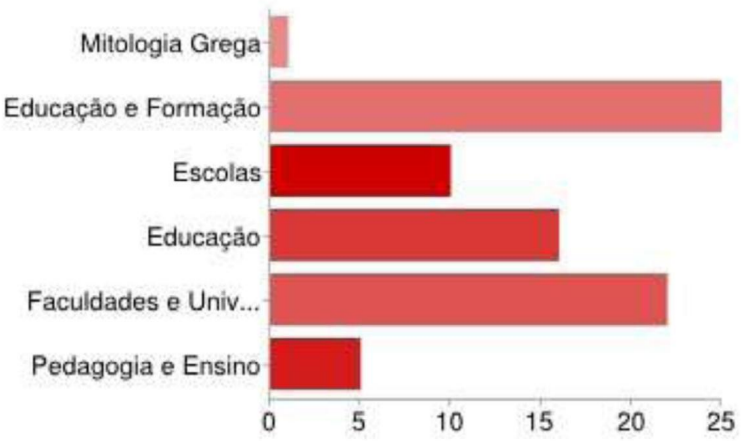
El análisis del resultado (fig. 4) para el texto 3, relacionado con el tema educación, muestra que todos los tópicos han sido marcados por los voluntarios. Así, la precisión específica para este texto fue del 100%. Diferente de lo ocurrido con otros textos, los resultados se centraron en dos o más alternativas, específicamente en cuatro alternativas (93% de las indicaciones) («educación y formación» con 32%, seguido por «Facultad y Universidad» con «educación» con 20% y «Escuelas» con 13%). Se observa que las alternativas más indicadas tienen como tema central la característica "educación", que también es el tópico más relevante por EXTRATOP.

La figura 5, refiriéndose al texto 4 sobre economía, demuestra que los resultados tenían una particularidad, concentrándose



Atores e Atrizes	13	22%
Musicais	37	63%
Estados	0	0%
Circuitos e Pistas de Corrida	0	0%
Programas e Cursos	4	7%
Brindes e Premiações	5	8%

Figura 3. Datos de texto sobre música

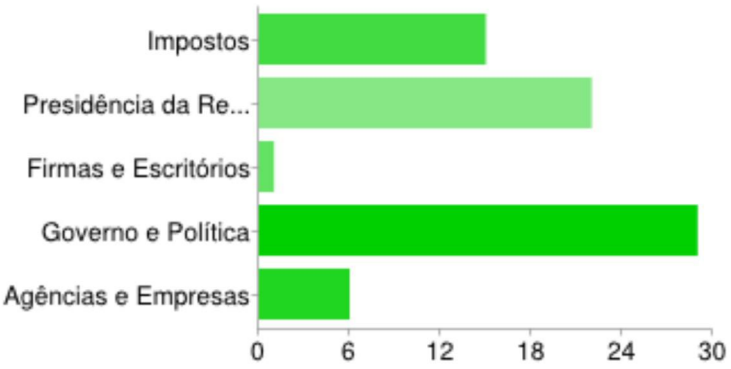


Mitologia Grega	1	1%
Educação e Formação	25	32%
Escolas	10	13%
Educação	16	20%
Faculdades e Universidades	22	28%
Pedagogia e Ensino	5	6%

Figura 4. Datos de texto sobre educación

Cinema e Filmes	0	0%
Educação e Formação	0	0%
Calendários	0	0%
Alberto Santos Dumont (1873-1932)	0	0%
Cédulas e Moedas	23	42%
Moeda	32	58%

Figura 5. Datos de texto sobre economía



Impostos	15	21%
Presidência da República	22	30%
Firmas e Escritórios	1	1%
Governo e Política	29	40%
Agências e Empresas	6	8%

Figura 6. Datos de texto sobre política

sólo en dos alternativas «Billetes y monedas» y «Monedas», respectivamente, con 42% y 58% de las indicaciones. Por lo tanto, la precisión del sistema ha llegado al 33%. Pese la baja precisión verificada, estos dos tópicos son los dos principales señalados por la herramienta EXTRATOP.

El último texto que se ha evaluado, se refiere a la política. A ejemplo de lo que ha sucedido en el texto 3, la precisión ha sido del 100%, una vez que todos los tópicos ha sido señalados en las respuestas de los voluntarios (Fig. 6). En este caso, los dos más citados, según los voluntarios, fueron «Política y gobierno» y «Presidencia y República», respectivamente con 40% y 30%, coincidiendo nuevamente con los dos tópicos más relevantes señalados por EXTRATOP.

Conclusiones

Un análisis global de todos los resultados, indica que el prototipo ha obtenido una precisión de un promedio del 73,4%. Este resultado se considera satisfactorio porque, en términos generales, la herramienta apunta a algunos tópicos que son, de hecho, relacionados con los textos; lo que puede indicar cierto potencial del enfoque que se ha sugerido para esta investigación.

Sin embargo, muchos ajustes con respecto al contenido de la taxonomía deben hacerse, una vez que, por el modo en que fue estructurada, cuando se considera un conjunto más grande (más de 4 ó 5 tópicos), los tópicos asociados por la herramienta aparentemente no están asociados al contenido de los textos como, por ejemplo, en el experimento que se ha llevado a cabo cuando uno de los temas relacionados con el texto sobre economía era «Cine y películas».

Esto se debe al hecho de que la lengua portuguesa contiene muchas palabras con más de un significado (que causan ambigüedad) y al trato sólo con el modo superficial del lenguaje (la herramienta no hace ningún tipo de desambiguación). Por otro lado, se observa también que hay palabras presentes en la taxonomía , porque son específicas para ciertos tópicos, que ayudan a la herramienta a definir con claridad los tópicos contenidos en los textos. Esto genera a su ves una diferencia de resultados, teniendo en cuenta los diferentes tópicos de los textos. Por ejemplo, en el experimento los textos concernientes a la «educación» y a la «política» han sido los que mostraron una mayor precisión. Esto ocurre por dos razones: en primer lugar, hay muchas palabras específicas que se refieren directamente y exclusivamente a los temas citados, o entonces, hay más palabras asociadas a estos tópicos en la taxonomía, incluso porque no había ningún tipo de preocupación para equilibrar el número de términos asociados a cada concepto.

Una limitación del EXTRATOP, como aplicación basada en taxonomía para procesar textos, se refiere a como se conceptualiza la propia taxonomía. Una vez que la lengua natural es dinámica, ello dificulta el proceso de levantamiento entre contenidos textuales y conceptos taxonómicos. El enriquecimiento de la ontología del Yahoo también ha sido un proceso de bastante limitado, ya sea desde

el punto de vista del número de conceptos descritos cómo desde el punto de vista de las fuentes que han sido utilizadas para el enriquecimiento. Esto puede, incluso, haber contribuido para los resultados obtenidos. Aun respecto a la taxonomía, su enriquecimiento ha sido influenciado directamente por la intervención humana, ya que no ha sufrido ningún tipo de refinamiento y ha sido hecho de modo totalmente subjetivo y ad hoc por un único ingeniero de conocimiento. Los propios ítem lexicales utilizados para describir los conceptos taxonómicos no han sido especificados hasta que se agotaran (a causa del elevado costo manual) en sus formas variantes, mínimamente, con relación a flexiones de género, número o grado. En esos casos, no se contemplan las variaciones lexicales, resultando en un levantamiento taxonómico frágil.

Tampoco se ha tenido en consideración en la construcción de la taxonomía, la cuestión de la ambigüedad, sobretodo debido a la polisemia. El levantamiento que se ha adoptó en este caso, al considerarse todos los conceptos relacionados a una misma palabra, claramente introduce distorsiones al realizar el cómputo de los pesos de los componentes textuales. De ese modo, las informaciones contenidas en la taxonomía no pueden ser consideradas como un conjunto completo, a pesar de amplio. Los resultados deben ser interpretados dentro de esos límites. La forma de evaluación utilizada para medir el desempeño de la estrategia de detección de tópicos también constituye una limitación, pues toda tarea subjetiva, manual, involucra limitaciones de consideración, que comprometen la confiabilidad y escalabilidad de un sistema automático. Con el propio advenimiento de la Web Semántica, que visa dar significado semántico al contenido de las páginas Web, la tendencia es que el procesamiento semántico de documentos se convierta en algo cada vez más necesario, haciendo con que los recursos como, por ejemplo, las taxonomías, sean cada vez más importantes. Ese hecho genera buenas perspectivas para mejorías y futuras investigaciones del EXTRATOP.

Bibliografía

Villarreal, S. E. G., Elizalde, L. M. and Viveros, A. C. Clustering hyperlinks for topic extraction: an exploratory analysis. In Proceedings of the Eighth Mexican International Conference on Artificial Intelligence, 2009.

Chakrabarti, S. 2003. Mining the Web: Discovering the Knowledge from Hypertext Data. Elsevier Science, 48.

Yang, Y. , He, L. and Qiu, M. 2011. Exploration and Improvement in Keyword Extraction for News Based on TFIDF.In Proceedings of the ESEP, 2011.

Kong, H. , Hwang, M. , Hwang, G. Shim, J. and Kim, P. 2006. Topic Selection of Web Documents Using Specific Domain Ontology. In Proceedings of the MICAI 2006.

Fang, J. , Guo, L. , Wang, X. D. and Yang, N. 2007. Ontology-Based Automatic Classification and Ranking for Web Documents. In Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007).

Coursey, K. Mihalcea, R. and Moen, W. Using Encyclopedic Knowledge for Automatic Topic Identification. InProceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), 2009.

He, X. , Ding, C. H. Q. , Zha, H. and Simon, H. D. 2001. Automatic Topic Identification Using Webpage Clustering. In Proceedings of the International Conference on Data Mining (ICDM 2001).

GRUBER, T. R. A translation approach to portable ontologies. Knowledge Acquisition,v. 5, n. 2, p. 199-220, 1993.

LOH, Stanley. “Uma abordagem baseada em conceitos para descoberta de conhecimentoem textos”. Tese (Doutorado), Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

PEDREIRA-SILVA, Patrick. “ExtraWeb: um sumarizador de documentos Web baseadoem etiquetas HTML e Ontologia”. Dissertação (Mestrado), Departamento de Ciênciada Computação , Universidade Federal de São Carlos, 2006.

TIUN, Sabrina; ABDULLAH, Rosni; ENYA KONG, Tang. “Automatic topic identification using ontology hierarchy”. In: CICLING: CONFERENCE ON INTELLIGENT TEXTPROCESSING AND COMPUTATIONAL LINGUISTICS, 2., 2001, Mexico City.Proceedings... Heildelberg : Springer-Verlag, 2001. p. 444-453.

Recibido: 22 de agosto de 2014.
Aprobado en su forma definitiva:
21 de octubre de 2014

Patrick Pedreira Silva
Universidade Sagrado Coração. Brasil
Correo electrónico:
patrick.silva@usc.br

Christian Freitas
Universidade Sagrado Coração. Brasil
Correo electrónico:
chrfreitas@gmail.com

Elvio Gilberto da Silva
Universidade Sagrado Coração. Brasil
Correo electrónico:
egsilva@usc.br