Hacia la construcción automática de ontologías

Isidra Ocampo-Guzman[1], Ivan Lopez-Arevalo[1], Edgar Tello-Leal[2], Victor Sosa-Sosa[1]

[1]Laboratorio de Tecnologías de Información
Cinvestav - Tamaulipas. Victoria, Tamaulipas, Mexico

[2]Universidad Autónoma de Tamaulipas.
Victoria, Tamaulipas, Mexico

{iocampo,ilopez,vsosa}@tamps.cinvestav.mx, etello@uat.edu.mx

Abstract. *Este artículo propone una metodología para la construcción automática de ontologías a partir de un corpus de texto. Primero se identifican los conceptos (temas) de los documentos del corpus utilizando Latent Dirichlet Allocation (LDA). Con base en el conjunto de temas identificados se construye para cada uno de éstos su taxonomía utilizando sus términos de mayor probabilidad. En la construcción de la taxonomía de los temas se utiliza el tesauro WordNet, mediante el cual se obtiene la similaridad y relación entre los términos que forman al tema. Las taxonomías resultantes se unen para formar la ontología final. Se evalúa la metodología propuesta con el corpus Lonely Planet.*

1. Introducción

Las ontologías tienen un papel clave para estructurar, manejar, compartir y procesar la información en la Web Semántica, corpus digitales, comercio electrónico, aplicaciones médicas, etc. ya que proveen conocimiento compartido y común de un determinado dominio [1]. El conocimiento compartido permite a las personas y aplicaciones de cómputo comunicarse de manera efectiva.

Las ontologías se pueden construir de forma manual o automática. Los seres humanos somos capaces de construir taxonomías semánticas de manera eficiente. Sin embargo, la construcción manual representa un trabajo intensivo, debido a que el conocimiento crece y es difícil generar, mantener y expandir con adecuada calidad ontologías de gran tamaño [2]. Así pues, se hace necesario automatizar la construcción de ontologías por medio de la aplicación de métodos que permitan minimizar el costo en tiempo y esfuerzo para su desarrollo y mantenimiento.

Por lo general la construcción de ontologías consiste en extenderlas o enriquecerlas con nuevos conceptos obtenidos a partir de un corpus relacionado con el dominio de la ontología inicial [3, 4], se parte de un conjunto de conceptos previamente formalizados. Tal es el caso de Alfonseca et al. [5] quienes extienden una ontología con nuevos conceptos, los cuales son obtenidos de términos que co-ocurren con los conceptos existentes en la ontología *semilla*. El método requiere que haya varias ocurrencias de los conceptos para que las términos sean considerados. Otro de los enfoques es construir

la ontología utilizando modelos probabilísticos como el trabajo de Jian-hua Yeh [2] que utiliza LDA para extraer los temas, con los cuales se construye aplicando el algoritmo de clustering aglomerativo jerárquico [6]. Ellos utilizan la medida de similaridad de cosenos y generan temas de alto nivel llamados *super temas*. Zavitsanos et al. [7] aplican LDA de manera repetida con diferente número de temas. Ellos parten de la idea de que un número pequeño de temas debe capturar todo el conocimiento que el corpus contiene, es decir, deben de ser más generales con respecto al muestreo de un número mayor de temas. Los temas aprendidos son directamente usados como conceptos que forman la estructura de la ontología. Este modelo es incapaz de obtener relaciones cuando un tema de nivel superior subsume a un solo tema del nivel inferior. Además no realizan el etiquetado de la ontología. Fortuna et al. [8] utilizan el modelo TF/IDF [9] y Latent Semantic Indexing (LSI), para asignar términos a un tema, haciendo cada tema una distribución sobre términos. LSI tiene el problema de ser sobre-especializado.

En el enfoque propuesto en este artćulo se construye la ontología sin previo conocimiento de los temas (conceptos) abordados por los documentos del corpus. Se identifica el conjunto de temas utilizando el modelo Latent Dirichlet Allocation (LDA) [10]. Posteriormente se construye para cada uno de ellos su taxonomía con sus términos de mayor probabilidad. Se utiliza el tesauro de WordNet [11] para identificar el término de mayor similaridad[1] con el resto para ser la raíz de la taxonomía y la relación[2] semántica entre los términos que forman al tema. Para determinar la similaridad y relación semántica se asocia cada término del tema con un número de glosa[3] de WordNet, la cual se determina aplicando TF-IDF del modelo espacio vectorial (MEV) [9]. Finalmente, se unen las ontologías cuya similaridad de sus nodos raíces sea mayor que un determinado *umbral*.

El resto de este artículo se encuentra estructurado de la siguiente manera. En la sección 2 se presenta una breve descripción del modelo utilizado. En la sección 3 se describe la metodología propuesta. La sección 4 contiene los experimentos, resultados y evaluación. Finalmente, en la sección 5 se presentan algunas conclusiones.

2. Latent Dirichlet Allocation

Latent Dirichlet Allocation [10] es un modelo temático probabilístico que captura las características *significativas* de los documentos del corpus. Consiste en un proceso generativo en donde los documentos de un corpus (considerados *bolsas de términos*) son generados como una mezcla temas (distribución multinomial de probabilidad sobre temas), donde los temas están formados por una mezcla de términos del vocabulario del corpus (distribución multinomial de probabilidad sobre el vocabulario). El proceso generativo para cada documento dado un número de temas K, es el siguiente:

1. Seleccionar N $\sim$ Poisson (ε)

[1] Cuantifica qué tanto dos conceptos son iguales con base en la información contenida en WordNet. Por ejemplo *autómovil* es más similar a *bote* que *árbol*.

[2] Indica los sentidos en que están relacionados dos conceptos, es información no jerárquica.

[3] Descripción y ejemplos de uso o aplicación de un *synset* (verbo, sustantivo, adjectivo o adverbio) contenido en WordNet.

2. Seleccionar $\theta \sim$ Dirichlet (α)
3. Para cada uno de los N términos w_n:
    - Seleccionar un tema $z_n \sim$ Multinomial (θ)
    - Seleccionar un término w_n de la distribución de probabilidad multinomial del tema z_n definida por $p(w_n|z_n,\beta)$.

$p(z_n=i)$ representa la probabilidad que el i^{th} tema sea muestreado para los *n* términos e indica los temas que son importantes (que reflejan el contenido) para un documento particular. Por su parte $p(w_n|z_n=i)$ representa la probabilidad de ocurrencia de un término w_n dado un tema *i* e indica la probabilidad de ocurrencia de un término para cada tema.

3. Metodología

La figura 1 muestra la metodología propuesta para construir de forma automática una ontología.

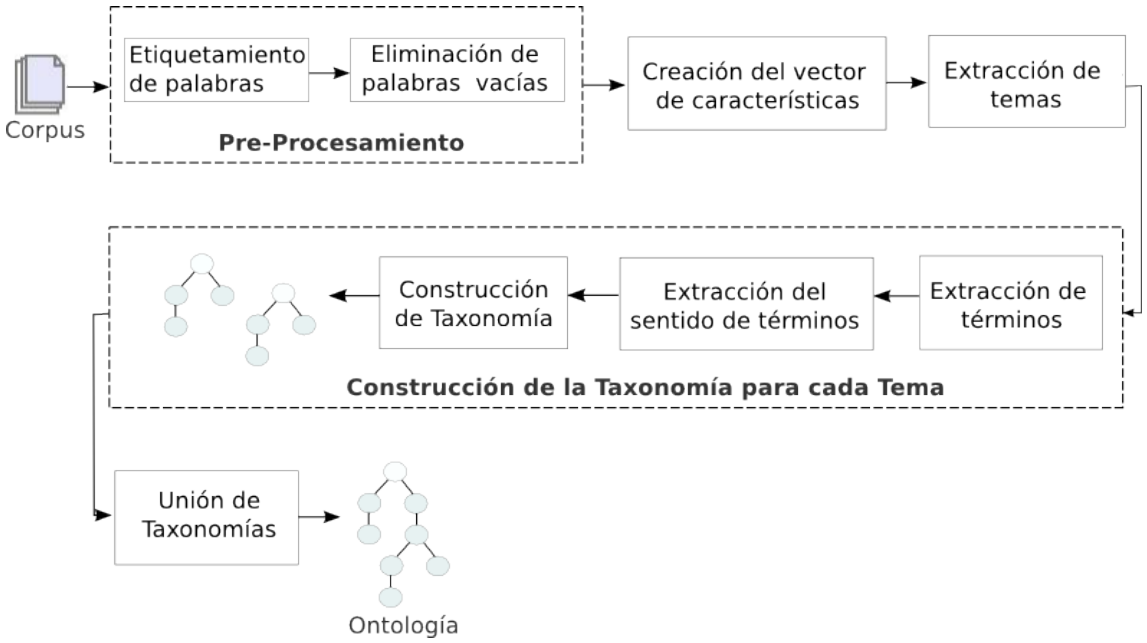

Figure 1. Metodología para la construcción de ontologías

La construcción comprende las siguientes fases:

1. *Pre-procesamiento*: Una ontología está formada por sustantivos, por lo que éstos se identifican mediante el siguiente proceso:
    - *Etiquetado de palabras*: permite obtener la categoría gramatical de las palabras de cada sentencia del documento. El etiquetado de las palabras considera el contexto en los que éstas ocurren. El objetivo es identificar los sustantivos.
    - *Eliminación de palabras vacías*: de acuerdo al resultado de la fase previa, se eliminan todos aquellas palabras que no sean sustantivos.

2. *Creación del vector de características*: se crea una matriz de documentos-palabras, donde cada vector representa un documento y las columnas a los términos del vocabulario. La matriz resultante tendrá un número de filas equivalente al número de documentos y un número de columnas equivalente al total de términos en el vocabulario. La matriz es la entrada del modelo LDA.
3. *Extracción de temas*: Se infiere un conjunto *T* de temas del corpus mediante la aplicación del proceso generativo inverso de los modelos probabilísticos. Es decir, a partir del conjunto de documentos ya conocidos se identifican los *temas* abordados por éstos aplicándoles el modelo LDA. La inferencia exacta de los temas es muy costosa, por lo que, se construye una cadena de Markov con técnicas de Monte Carlo utilizando un muestreador que siga el modelo LDA.
4. *Construcción de la taxonomía de cada tema*: Para cada tema *i* del conjunto de temas *T* se construye su taxonomía en tres fases:
    - *Extracción de términos*: los temas están compuestos por una mezcla de probabilidades sobre el vocabulario del corpus. La taxonomía de cada tema se construye utilizando un subconjunto de términos del vocabulario: para cada tema *i* del conjunto de temas *T* se extraen aquellos términos cuya probabilidad sea mayor que la mediana de su distribución de probabilidad T_i.

      $$W_{T_i}=P(w_j > \text{mediana de f}(T_i))$$

      Se considera que aquellos términos con mayor probabilidad en el tema son más representativos del contenido de éste y con mayor relación semántica entre ellos.
    - *Extracción del sentido de los términos*: para la construcción de la taxonomía de un tema se obtiene la similaridad entre los términos seleccionados en la fase anterior, por lo que se asocia a cada término w_i un sentido de éste en WordNet, el cual se determina utilizando el modelo espacio vectorial (MEV) [9]. Se obtienen todos los sentidos en WordNet del término en cuestión. Cada sentido se considera como un documento, el cual se modela como un vector de acuerdo al MVE. Emulando el funcionamiento del MEV en recuperación de información, se construye una *consulta* con los términos seleccionados del tema al que corresponde el término y se aplica la medida de similaridad de cosenos para obtener el número de *glosa* más similar a la consulta que se asociará al término.
    - *Construcción de taxonomía*: para construir la taxonomía de cada tema se calcula la similaridad y la relación semántica entre los términos seleccionados de cada tema, las cuales se obtienen mediante las medidas Gloss Vector[4] y Wu and Palmer[5] (WUP) de WordNet::Similarity[6]. El proceso es el siguiente:
        (a) La raíz de la taxonomía es el término *i* cuya sumatoria de similaridad sea mayor con respecto a los demás términos del tema. Para

[4]Medida de similaridad que forma vectores de segundo orden de co-ocurrencia de términos utilizando la información de un corpus o definiciones de los conceptos de WordNet. La similaridad de dos conceptos es determinada como el coseno del ángulo entre las glosas de los conceptos.

[5]Medida de relación que calcula la relación de dos conceptos considerando la profundidad de éstos en la jerarquía de WordNet.

[6]$http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi.$

cada término *i* se forman el total de combinaciones de pares de términos que contienen a *i* y otro término del resto que conforma al tema. Para cada combinación se obtiene la similaridad Gloss Vector, la cual se multiplica por la probabilidad del término *i* en el tema.

(b) Se agregan a la taxonomía aquellos términos cuya profundidad obtenida con la medida de relación WUP sea mayor que la del término raíz y su similaridad supere un determinado *umbral*. Para cada término a agregar se compara con los términos contenidos hasta el momento en la taxonomía, se agrega el término como nodo hijo del nodo con el cual tenga mayor similaridad.

Este proceso se repite para cada tema *i* del conjunto *T*. De esta manera se obtiene un conjunto *J* de taxonomías.

5. *Unión de taxonomías:* el conjunto de taxonomías *J* se une para obtener la ontología final del corpus. El proceso de unión se realiza obteniendo la similaridad entre pares de raíces del conjunto *J* de taxonomías, se unen aquellas que sobrepasen un *umbral* establecido. Se valida la profundidad de los términos raíces considerando dos casos:
   - Profundidad de taxonomías iguales: para unir las dos taxonomías se obtiene el concepto en WordNet que subsume a sus términos raíces. La taxonomía resultante se agrega al conjunto *J* de taxonomías y se eliminan las dos taxonomías semillas.
   - Profundidad de taxonomías diferentes: se identifica la taxonomía de menor profundidad (*TG*) y la de mayor profundidad (*TE*). Se considera a *TG* más general con respecto a *TE*, dado esto se agrega *TE* a *TG* en el nivel *K*, el cual es igual a la profundidad de *TE*. Si el conjunto de nodos en *TG* en el nivel *K-1* es mayor que 1, *TE* se agrega al nodo con mayor similaridad en dicho conjunto. Se elimina *TE* del conjunto *J* de taxonomías.

   El proceso anterior se repite mientras existan pares de taxonomías en el conjunto *J* que superen el *umbral*. Finalmente se unen las taxonomías del conjunto *J* bajo un nodo raíz *ROOT*.

4. Preliminares y evaluación

4.1. Experimentos

La metodología propuesta se ha evaluado empleando el corpus Lonely Planet[7]. La ontología de este corpus ha sido construida manualmente por los autores y contiene 96 conceptos extraídos de 1801 documentos de texto del dominio de turismo.

La metodología ha sido implementada en Java. El etiquetado del corpus se realizó con la herramienta Stanford POS Tagger[8]. Al obtener sólo las palabras etiquetadas como sustantivos se redujo en un 79% el total de palabras del corpus. Para la extracción de los temas del corpus se utilizó el muestreador Gibbs[9]. Los valores de los hiperparámetros de

[7] $http://olc.ijs.si/lpReadme.html$.
[8] $http://nlp.stanford.edu/software/tagger.shtml$.
[9] $http://gibbslda.sourceforge.net/$.

LDA α y β utilizados para el muestreo son de 50/*K* y 0.01 respectivamente de acuerdo a Griffiths y Steyver [12]. Se realizaron 31 experimentos (de acuerdo al Teorema del Límite Central) para el total de temas *K* a obtener del corpus y los mejores resultados se obtuvieron con un valor de *K*=25 y 10,000 iteraciones de muestreo. Por otro lado, también se desarrollaron diversas pruebas para determinar el valor del *umbral* de *similaridad* entre términos y éste se estableció en 0.7 ya que arrojó mejores resultados.

4.2. Resultados y Evaluación

Se identificaron 115 conceptos de los cuales 61 corresponden a los 96 contenidos en la ontología contruida de manera manual. La ontología resultante se evaluó utilizando la medida *F* [13]:

$$F = \frac{2*precision*cobertura}{precision+cobertura}$$

La medida *F* combina las métricas de precisión y cobertura. La precisión se refiere a la proporción de conceptos correctamente identificados con respecto al total de conceptos identificados en la ontología construida, mientras que la cobertura se refiere a la proporción de conceptos identificados correctamente con respecto al número de conceptos de la ontología de referencia. La tabla 1 muestra los resultados obtenidos.

Conceptos identificados		
Precisión	Cobertura	Medida F
53%	63%	58%

Table 1. Resultados

5. Conclusiones

Este artículo propone un enfoque para construir ontologías a partir de un corpus de manera autómatica haciendo uso del modelo probabilístico LDA, el cual permite identificar los temas abordados por los documentos del corpus. Se parte de la idea de que cada tema trata sobre un determinado concepto y que los términos que lo componen deben de tener una taxonomía que describa a éste. Se utiliza el teasuro WordNet para determinar la similaridad y relación semántica entre términos que conforman a cada tema.

Aunque WordNet es fundamental en la propuesta presentada, puede verse también como una desventaja ya que la construcción de la taxonomía de cada tema depende totalmente de ella. En el corto plazo se experimentará la utilización de técnicas de Procesamiento del Lenguaje Natural que permitan obtener la similaridad entre términos.

References

[1] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

[2] Jian-Hua Yeh and Naomi Yang. Ontology construction based on latent topic extraction in a digital library. In *ICADL 08: Proceedings of the 11th International Conference on Asian Digital Libraries*, pages 93–103, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] Eneko Agirre, Olatz Ansa, Eduard Hovy, and David Martínez. Enriching very large ontologies using the www. In *in Proceedings of the Ontology Learning Workshop*, 2000.

[4] Andreas Faatz and Ralf Steinmetz. Ontology enrichment with texts from the www. In *In Semantic Web Mining, WS02*, 2002.

[5] E. Alfonseca and S. Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *Poceedings of the First International Conference on General WordNet*, Mysore, India, 2002.

[6] Dwi H. Widyantoro, Thomas R. Ioerger, and John Yen. An incremental approach to building a cluster hierarchy. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining*, page 705, Washington, DC, USA, 2002. IEEE Computer Society.

[7] Elias Zavitsanos, Georgios Paliouras, George A. Vouros, and Sergios Petridis. Discovering subsumption hierarchies of ontology concepts from text corpora. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 402–408, Washington, DC, USA, 2007. IEEE Computer Society.

[8] B. Fortuna, D. Mladenic, and M. Grobelnik. Visualization of text document corpus. *Informatica Journal*, 29(4):497–502.

[9] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[11] Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro. Wordnet applications. In *In: Proc. of the 2nd Int. Conf. Global WordNet*, 2004.

[12] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004.

[13] C.J. van Rijsbergen and Ph. D. Information retrieval, 1979.