OntOAIr: UN MÉTODO DE CONSTRUCCIÓN DE ONTOLOGÍAS

Ma. Auxilio Medina
mauxmedina@gmail.com

Resumen: Las ontologías son componentes clave en el desarrollo de la web semántica porque representan un dominio de interés común. Su construcción consume tiempo y recursos, principalmente la fase de adquisición de conocimiento. Este documento describe el método OntOAIr, un método de construcción parcialmente automático de ontologías ligeras a partir de colecciones de documentos. El método emplea representaciones simplificadas de documentos y una adaptación del algoritmo de agrupamiento FIHC (acrónimo de "*Frequent Itemset-based Hierarchical Clustering* algo-rithm").

Palabras clave: ontologías, métodos de aprendizaje de ontologías, agrupamiento de documentos.

1. Introducción

Las ontologías son componentes clave en el desarrollo de la web semántica porque proveen una representación no ambigua de un dominio. En bibliotecas digitales, apoyan tareas como interoperabilidad, descripción de recursos, integración, búsqueda y navegación de información.

La fase de adquisición de conocimiento en la construcción de ontologías es la más costosa. Cuando esta fase se automatiza, se emplean métodos de aprendizaje de ontologías ("*ontology learning methods*"). Estos métodos se basan con frecuencia en técnicas de aprendizaje de máquina, procesamiento de lenguaje natural o algoritmos de agrupamiento. Este documento describe el método OntOAIr (abreviación de "*Ontologies from Open Archives Initiative Repositories to Support Information Retrieval*", ontologías para repositorios de la Iniciativa de Archivos Abiertos que apoyan la recuperación de información). OntOAIr es un método de construcción de ontologías ligeras, lo que implica que la participación de un experto del dominio se limita a la asignación de valores a parámetros de entrada. Originalmente, el método consideró las colecciones de la Iniciativa de Archivos Abiertos (OAI) [Lagoze y Sompel 2001], organización que promueve estándares de interoperabilidad, aunque puede aplicarse a cualquier colección que permita acceder al título y resumen de sus documentos.

Las ontologías construidas por el método OntOAIr son estructuras jerárquicas de grupos disjuntos de documentos similares; representan un modelo de datos accesible a usuarios humanos y de software que facilitan la organización y la recuperación de la información de una colección de documentos.

2. Trabajo relacionado

Existen diferentes métodos de aprendizaje de ontologías. Esta sección describe algunos de los relevantes.

El método OntoLearn [Navigli y Velardi 2004] aplica un algoritmo jerárquico de agrupamiento a un conjunto documentos almacenados en sitios web o bancos de datos. Los documentos pertenecen a un dominio predeterminado como medicina, turismo o música. Tanto la salida del algoritmo como la red de palabras WordNet (Disponible en: *http://wordnet.princeton.edu/*) se emplean para construir las ontologías. A diferencia del método OntOAIr, los conceptos descritos en OntoLearn pueden estar compuestos por dos o más términos. Sin embargo, requiere que los documentos estén previamente clasificados.

Los documentos [Ljubic et al. 2005] y [Plisson et al. 2005] describen métodos para construir ontologías a partir de un conjunto de documentos que describen las competencias de una compañía. El primero usa el algoritmo de bisección, el segundo el de k-partes. Su desventaja principal es el criterio de paro que requieren los algoritmos. En general, asignar un valor como criterio de paro es una tarea difícil en escenarios reales. En el algoritmo FIHC (utilizado en el método OntOAIr) este criterio es opcional.

Diederich y Balke 2007 describen un algoritmo para crear ontologías ligeras denominadas *sistemas de categorización de temas*. El algoritmo emplea las palabras clave provistas por los autores para calcular una métrica de ocurrencia, encontrar relaciones y construir grafos de vecindad. La aplicación del algoritmo requiere que

los documentos cuenten con anotaciones digitales.

3. El método OntOAIr

El método OntOAIr usa representaciones simplificadas de documentos, una adaptación del algoritmo FIHC (propuesto en [Fung et al. 2003]) y técnicas de ingeniería ontológica. El método es universal si se considera que puede ser usado en diferentes dominios, lenguajes y aplicaciones.

Las tareas principales en el método OntOAIr son: (1) recolección, (2) representación, (3) agrupamiento y (4) formalización. Estas tareas se describen a continuación:

Recolección.- Consiste en obtener los documentos de una o más colecciones. Típicamente, las colecciones cuentan con cientos o miles de documentos cuyos datos descriptivos deben transmitirse en la red, (esta tarea se realiza fuera de línea, "*off-line*").

Representación.- Construye *vectores característicos* para cada uno de los documentos recolectados. El nombre se adopta de [Fung et al. 2003], quien los define como representaciones simplificadas formadas por palabras clave (términos diferentes de proposiciones y artículos) y pesos (valores numéricos asociados a las palabras clave que intentan reflejar su relevancia).

Agrupamiento.- Aplica el algoritmo FIHC a los vectores característicos para producir un árbol de grupos. FIHC es un algoritmo aglomerado basado en la siguiente hipótesis: "si un grupo de documentos tratan temas similares, compartirán un conjunto de términos". A los términos compartidos se les denomina *conjuntos de términos frecuentes* ("*frequent item sets*").

Formalización.- La tarea de formalización transforma el árbol de grupos producido en el agrupamiento en una ontología ligera. Se ha explorado el uso de los lenguajes XML, RDF y OWL para representar las ontologías construidas.

El algoritmo FIHC produce una estructura jerárquica de grupos disjuntos, requiere de tres parámetros de entrada: 1) soporte mínimo, 2) soporte global y 3) soporte de grupo. El soporte mínimo es un valor asociado con la extracción de conjuntos de términos frecuentes; el soporte global mide el portentaje de documentos en una colección que contiene un conjunto de términos frecuentes y el *soporte de grupo* es el porcentaje de documentos en un grupo que contiene un conjunto de términos frecuente. Los valores de los parámetros de entrada se determinan experimentalmente.

La representación en XML de la ontología se muestra en la Tabla 1, la cual puede interpretarse como sigue:

1. El atributo fecha ("*date*") y el elemento algoritmo ("*algorithm*") describen los datos descriptivos de la ontología (los meta-datos)
2. La ontología está formada por uno o más grupos (elementos "*cluster*")
3. El elemento "*record*" representa un documento
4. Cada grupo tiene una etiqueta ("*label*"), un nivel ("*level*") y uno o más documentos ("*record*")
5. Los elementos "*subject*" y "*description*" representan el tema y resumen de un documento, respectivamente. Estos elementos son opcionales, los restantes son obligatorios.

Tabla 1. Estructura de una ontología producida por el método OntOAIr.

```
<!ELEMENT ontologyofrecords (algorithm, cluster+)>
<!ATTLIST ontologyofrecords date CDATA #REQUIRED>
<!ELEMENT algorithm EMPTY>

<!ATTLIST algorithm name CDATA #FIXED "FIHC"
                    globalsupport CDATA #REQUIRED
                    clustersupport CDATA #REQUIRED>

<!ELEMENT cluster (label, level, record*, cluster*)>
<!ELEMENT label (#PCDATA)>
<!ELEMENT level (#PCDATA)>

<!ELEMENT record (title, subject?, description?,
identifier, url, dataprovider, metadataformat,
datestamp)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT subject (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT identifier (#PCDATA)>
<!ELEMENT url (#PCDATA)>
<!ELEMENT dataprovider (#PCDATA)>
<!ELEMENT metadataformat (#PCDATA)>
<!ELEMENT datestamp (#PCDATA)>

<!ENTITY generatedBy "OntoSIR 2.1" >
```

Las ontologías construidas con el método OntOAIr son estructuras jerárquicas que agrupan documentos similares de forma que los documentos de un grupo en el nivel *k* comparten los *k* términos de la etiqueta.

La Figura 1 muestra las etiquetas de una ontología construida a partir de un conjunto de reportes técnicos extraídos de CORTUPP, la Colección de Reportes Técnicos de la Universidad Politécnica de Puebla (disponible en http://server3.uppuebla.edu.mx/cortupp/). La ontología está organizada en tres grupos con

etiquetas: *procesamiento*, ambientes y *robótica*. El grupo de procesamiento se divide en *procesamiento de lenguaje* y *procesamiento de imágenes*. A su vez, el grupo de *procesamiento de lenguaje* se subdivide (o se especializa) en el grupo *procesamiento de lenguaje natural*. La interpretación de los grupos restantes es similar.

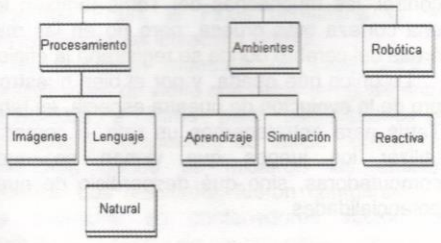

Figura 1. Ontología para un grupo de documentos de la colección TR.

La descripción detallada del método puede consultarse en [Medina y Sánchez 2008] y [Medina 2008].

4. Conclusiones

Este documento presentó al método OntOAIr, el cual es un método para construir ontologías ligeras. La construcción requiere de la implementación de cuatro tareas: recolección, representación, agrupamiento y formalización. Las ontologías construidas representan un modelo de datos que puede ser consultado por usuarios humanos y de software para apoyar tareas de bibliotecas digitales como organización de colecciones, recuperación de información e identificación de similitud entre documentos.

Actualmente, una ontología construida por el método OntOAIr representa la colección CORTUPP, ésta se usa en dos proyectos en desarrollo: el primero la utiliza para implementar un algoritmo de búsqueda del mejor grupo con el propósito de mantener la confiabilidad de la colección y el segundo asocia a la ontología una representación gráfica a manera de mecanismo de navegación.

5. Referencias

[Diederich y Balke 2007] Diederich J., Balke W. 2007. The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries'07* (ECDL'07, Budapest, Hungary, Septiembre. Lecture Notes in Computer Science. Springer. Volumen 4675. 33-40.

[Fung et al. 2003] Fung B.C.M. , Wang K., Ester M. 2003. Hierarchical Document Clustering Using Frequent Itemsets. *Proceedings of the Third SIAM International Conference on Data Mining*, (SDM'03, San Francisco, California, Mayo). SIAM Press. 59-70.

[Lagoze y Sompel 2001] Lagoze C. y Sompel Van de. 2001. The Open Archives Initiative: Building a Low-Barrier Interoperability Framework. *Conference Proceedings of Joint Conference on Digital Libraries* (JCDL'01, Roanoke, VA, USA). Junio, 54-62.

[Ljubic et al. 2005] Ljubic P., Lavrac N., Plisson J., Mladeniae D., Bollhalter S., Jermol M. 2005. Automated Structuring of Company Competencies in Virtual Organizations. *Proceedings of the Conference on Data Mining and Data Warehouses 2005* (SiKDD 2005, Ljubljana, Slovenia, October. 7th International Multi-conference on Information Society IS'05". Octubre. 190-193.

[Medina y Sánchez 2008]. Medina M. A., Sánchez A. 2008. OntOAIr: a method to construct lightweight ontologies from document collections. *Proceedings of the Ninth Mexican International Conference on Computer Science 2008*, (ENC 08, Baja California, México, Octubre), 2008. IEEE Computer Society". 62-73.

[Medina 2008] Medina M. A. 2008. OntOAIr: Construction of Lightweight Ontologies to Support Information Retrieval from Multiple Collections of Documents. Universidad de las Américas – Puebla. 2008. Tesis de Doctorado.

[Navigli y Velardi 2004]. Navigli R., Velardi P. 2004. Learning domain ontologies from document warehouses and dedicated web sites. Computational Lingüistics. Volumen 30. The MIT Press. 151-179.

[Plisson et al. 2005] Plisson J., Mladeniae D., Ljubic P., Lavrac N., Grobelnik M. 2005. Using Machine Learning to Structure the Expertise of Companies: Analysis of the Yahoo! Business Data. *Conference Proceedings on Data Mining and Data Warehouses* (SiKDD 2005). 7th International Multi-conference on Information Society IS'05. 186-189.