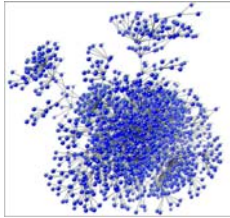


Generación semiautomática de ontologías

Seminari IULA
INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA
Universitat Pompeu Fabra



Rafael Pedraza Jiménez
Àrea de Coneixement de Biblioteconomia i Documentació
Universitat Pompeu Fabra
rafael.pedraza@upf.edu

Seminari IULA - Grupo IULATERM
[23/02/2007]

1

Índice

1. ¿Qué es una ontología?
 1. ¿Por qué están de moda?
 2. Áreas de aplicación.
 3. Inconveniente de las ontologías.
2. La ingeniería de ontologías.
 1. Características de estos sistemas
 2. El aprendizaje de ontologías
 3. Etapas en el aprendizaje de ontologías.
4. Aprendizaje de ontologías:
 1. Importación y reutilización.
 2. Extracción.
 3. Poda.
 4. Refinamiento.
5. Arquitectura de un sistema semiautomático para el aprendizaje de ontologías.
3. Extracción de ontologías mediante técnicas de Análisis Formal de Conceptos
 1. FCA vs. Técnicas de agrupamiento
 2. Definición del problema
 3. Tareas
 4. Diseño experimental
 5. Resultado esperado
4. Conclusiones.
5. Bibliografía.

Seminari IULA - Grupo IULATERM
[23/02/2007]

2

¿Qué es una ontología?

- Una ontología es “una especificación **explícita** y **formal** de una **conceptualización compartida**” [Studer, 1998].

```
...  
<owl:Class rdf:ID="WineDescriptor" />  
<owl:Class rdf:ID="WineColor">  
<rdfs:subClassOf rdf:resource="#WineDescriptor" />  
...  
<owl:Class>  
<owl:ObjectProperty rdf:ID="hasWineDescriptor">  
<rdfs:domain rdf:resource="#Wine" />  
<rdfs:range rdf:resource="#WineDescriptor" />  
<owl:ObjectProperty>  
<owl:ObjectProperty rdf:ID="hasColor">  
<rdfs:subPropertyOf rdf:resource="#hasWineDescriptor" />  
<rdfs:range rdf:resource="#WineColor" />  
...  
<owl:ObjectProperty>  
...
```

Seminari IULA - Grupo IULATERM
[23/02/2007]

3

¿Por qué están de moda?

- La popularidad actual de las ontologías se debe a que el avance en el estudio de estas herramientas presagia la tan ansiada **comprensión de un dominio tanto para personas como para aplicaciones.**

Seminari IULA - Grupo IULATERM
[23/02/2007]

4

Áreas de aplicación

- Web Semántica.
- Gestión del conocimiento.
- Sistemas de recomendación de consultas.
- Hipertexto.
- Teleeducación (e-learning).
- Comercio electrónico.
- etc.

Seminari IULA - Grupo IULATERM
[23/02/2007]

5

Inconveniente de las ontologías

- En muchas ocasiones la generación de ontologías de forma manual supone un elevado coste en tiempo y dinero.
- ¿Existe alguna solución?...

Seminari IULA - Grupo IULATERM
[23/02/2007]

6

La ingeniería de ontologías

- Es el área dedicada al estudio y diseño de entornos y/o aplicaciones que ayuden a elaborar, mantener y utilizar ontologías.
- Su objetivo es, por tanto, automatizar los procesos de construcción, mantenimiento y uso de una ontología.

Seminario IULA - Grupo IULATERM
[23/02/2007]

7

Características de un entorno para la ingeniería de ontologías

- Un sistema para la generación automatizada de ontologías debe [Mizoguchi, 2004] :
 1. Permitir la gestión de todo el proceso de desarrollo de la ontología.
 2. Facilitar el desarrollo colaborativo.
 3. Poseer una metodología fundamentada en la teoría de ontologías.
 4. Representar formalmente la ontología resultante mediante alguna norma o sintaxis (normalmente las recomendaciones del W3C).
 5. Disponer de un motor de inferencias.
 6. Ser usable
 7. Ser extensible

Seminario IULA - Grupo IULATERM
[23/02/2007]

8

Sistemas para la ingeniería de ontologías

- KAON
<http://kaon.semanticweb.org/>
- Hozo
<http://www.hozo.jp/>
- WebODE
<http://webode.dia.fi.upm.es/WebODEWeb/index.html>
- Protégé
<http://protege.stanford.edu/>

Seminario IULA - Grupo IULATERM
[23/02/2007]

9

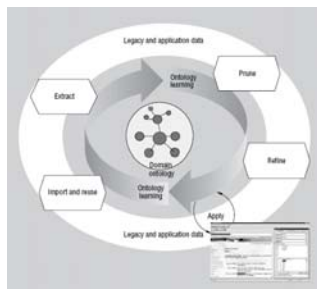
El aprendizaje de ontologías

- El aprendizaje de ontologías (Ontology learning) es la disciplina que investiga el desarrollo de métodos y herramientas que permitan la **creación de una ontología de forma semiautomática**.
- Estas aplicaciones estarán al servicio de un experto humano, el **"ingeniero ontológico"**, que supervisará todo el proceso de creación de la ontología.

Seminario IULA - Grupo IULATERM
[23/02/2007]

10

Etapas en el aprendizaje de ontologías



Fuente: [Maedche, 2001]

Seminario IULA - Grupo IULATERM
[23/02/2007]

11

Aprendizaje de ontologías: 1. Importación y reutilización

- Su objetivo es desarrollar mecanismos y estrategias para importar y reutilizar conceptualizaciones de un dominio a partir de estructuras o esquemas preexistentes. Combina técnicas automáticas y manuales.
- Etapas:
 1. Identificación de las estructuras o esquemas, y discusión con los expertos en el dominio.
 2. Fusión de los esquemas y estructuras seleccionadas para constituir una única base sobre la que elaborar la ontología.

Seminario IULA - Grupo IULATERM
[23/02/2007]

12

Aprendizaje de ontologías: 2. Extracción de la ontología

- La extracción de ontologías conlleva las siguientes fases:
 1. Selección de los conceptos del dominio y sus entradas léxicas.
 2. Generación de la taxonomía de conceptos: mediante técnicas de agrupamiento.
- Estas dos fases principales pueden ser complementadas con:
 3. Ampliación de las relaciones entre conceptos mediante diccionarios y reglas de asociación.

Seminario IULA - Grupo IULATERM
[23/02/2007]

13

Aprendizaje de ontologías: 3. Poda

- El sistema debe permitir al ingeniero ontológico eliminar un concepto cuando éste lo estime conveniente, pero mostrando las consecuencias de tal eliminación sobre el resto de la ontología.
- El sistema debe intentar mostrar la importancia de los conceptos dentro de la ontología (por ejemplo, en función de su frecuencia en un corpus), para que el ingeniero ontológico decida si debe mantenerlos o eliminarlos.

Seminario IULA - Grupo IULATERM
[23/02/2007]

14

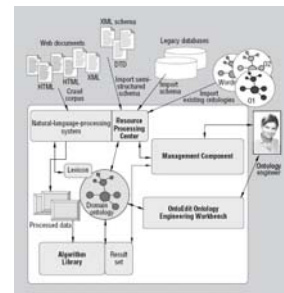
Aprendizaje de ontologías: 4. Refinamiento

- Consiste en la incorporación de nuevas entradas léxicas y/o conceptos como consecuencia de:
 1. Las necesidades específicas de los usuarios de un dominio.
 2. La actualización del dominio.

Seminario IULA - Grupo IULATERM
[23/02/2007]

15

Arquitectura de un sistema semiautomático para el aprendizaje de ontologías



Seminario IULA - Grupo IULATERM
[23/02/2007]

16

Herramientas para el aprendizaje de ontologías

- Text-to-Onto
<http://sourceforge.net/projects/texttoonto>
- Terminae
<http://www-lipn.univ-paris13.fr/~szulman/TERMINAE.html>
- Ontolearn
http://www.dbg-metzingen.de/Menschen/Lehrer/Q-T/Rittershofer/E-Learning/nach_Technik/OntoLearn/

Seminario IULA - Grupo IULATERM
[23/02/2007]

17

Extracción de ontologías mediante técnicas de Análisis Formal de Conceptos (FCA)

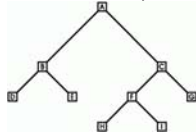
- Planteamos la extracción de ontologías (taxonomías) como una tarea de categorización [Pedraza-Jimenez, 2006].
- Tratamos de resolver el problema combinando técnicas de FCA con un potente recurso léxico (WordNet).

Seminario IULA - Grupo IULATERM
[23/02/2007]

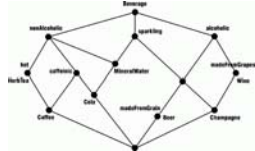
18

FCA vs. Técnicas de agrupamiento

- La principal ventaja de las técnicas de FCA [Davey, 2002] frente a las técnicas de agrupamiento radica en su capacidad para extraer relaciones de abstracción entre los agrupamientos que se generan (y no meras particiones como hacen las técnicas de agrupamiento clásicas), no sólo de tipo jerárquico, sino también de herencia múltiple.



Agrupamiento jerárquico



Retículo de conceptos FCA

Seminario IULA - Grupo IULATERM
[23/02/2007]

19

Definición del problema (I)

- Aproximación lingüística:
 - Postulado 1: Los documentos tienen contenido, que es el conjunto de las categorías que manifiestan, y viceversa, es decir una categoría tiene como forma perceptible el conjunto de documentos que la expresan.



Seminario IULA - Grupo IULATERM
[23/02/2007]

20

Definición del problema (II)

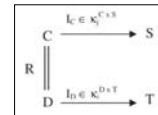
- Postulado 2: Los documentos más generales son aquellos que muestran un menor número de categorías, mientras que los más específicos serán aquellos que muestren un número mayor de categorías.

Seminario IULA - Grupo IULATERM
[23/02/2007]

21

Definición del problema (III)

- Postulado 3: Un signo lingüístico puede ser analizado a partir de sus signos componentes, en particular:
 - Los documentos D pueden ser analizados a través de sus términos componentes T.
 - Las categorías C pueden ser analizadas a partir de sus significados atómicos S.

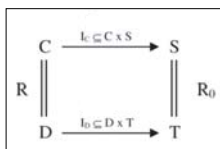


Seminario IULA - Grupo IULATERM
[23/02/2007]

22

Definición del problema (IV)

- Postulado 4: los términos tienen contenido, que son sus significados.



Seminario IULA - Grupo IULATERM
[23/02/2007]

23

Tareas...

- Estimar la relación de descomposición de documentos en términos I_D ($D \times T$) (parametrización de los documentos).
- Estimar la relación entre los términos y sus significados R_0 ($T \times S$).
- Estimar la relación de descomposición entre categorías y significados I_C ($C \times S$).
- Estimar la relación R ($C \times D$) para cualquier conjunto de documentos de un dominio.

Seminario IULA - Grupo IULATERM
[23/02/2007]

24

Diseño experimental (I)

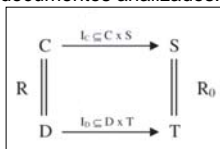
- Utilizamos una colección ya categorizada (Reuters-21578), cuyos documentos están divididos en sendos conjuntos de entrenamiento y prueba.
- Partimos de la hipótesis de que las categorías/clases en ambas colecciones son las mismas.

Diseño experimental (II)

- Descomposición los documentos en sus términos componentes (Obtención de I_D). Fases:
 - Preprocesado:
 - Stripping (eliminación de etiquetas)
 - Normalización
 - Identificación de acrónimos
 - Conversión de mayúsculas a minúscula
 - Control de cantidades numéricas y fechas
 - Eliminación de signos de puntuación y palabras vacías (artículos, determinantes, verbos auxiliares, conjunciones, preposiciones, pronombres, interjecciones, contracciones y adverbios de grado)
 - Lematización (Morphy)
 - Obtención del conjunto de significados asociados a cada término lematizado mediante el uso de WordNet (Obtención de R_0).

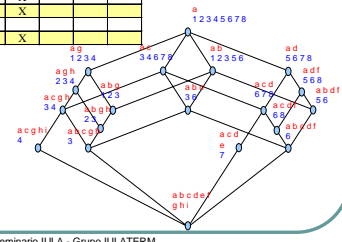
Diseño experimental (III)

- Mediante la composición de las relaciones R^E , I_D^E , y R_0^E obtenemos la relación I_C .
- Obtenida I_C en función del conjunto de entrenamiento, esperamos estimar la relación R para el conjunto de validación, que nos facilitará la taxonomía para la colección de documentos analizados.



Resultado esperado: Contexto y retículo FCA

	a	b	c	d	e	f	g	h	i
1	X	X					X		
2	X	X					X	X	
3	X	X	X				X	X	
4	X		X				X	X	X
5	X	X	X	X		X			
6	X	X	X	X	X		X		
7	X	X	X	X	X	X			
8	X		X	X			X		



Conclusiones

- La previsible aplicación de las ontologías a distintos escenarios hace imprescindible el desarrollo de métodos que permitan su fácil y rápida elaboración a un coste razonable.
- El diseño de estas herramientas de forma manual no es siempre factible debido a su elevado coste en tiempo y dinero.
- La ingeniería de ontologías, especialmente a través de las técnicas de aprendizaje de ontologías, ofrece la posibilidad de agilizar la creación de estas herramientas, a la vez que garantiza su integridad y corrección gracias a la supervisión humana durante el proceso de generación de las mismas.
- El análisis formal de conceptos representa una tecnología potente para el procesamiento de contenidos, extrayendo relaciones de abstracción entre las categorías que las técnicas de agrupamiento tradicionales no pueden.

Bibliografía

- [Davey, 2002] B. A. Davey and H. A. Priestley. *Introduction to lattices and order* (2nd edition). Cambridge University Press, 2002.
 - [Maedche, 2001] Maedche, A., and Staab, S. *Ontology learning for the semantic web*. IEEE Intelligent Systems, march-april 2001.
 - [Mizoguchi, 2004] Mizoguchi, R. *Ontology Engineering Environments*, in S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 173-189. Springer, 2004.
 - [Pedraza-Jimenez, 2006] Rafael Pedraza-Jimenez, Francisco Valverde-Albacete, and Ángel Navia-Vázquez. *A Generalisation of Fuzzy Concept Lattices for the Analysis of Web Retrieval Tasks*. Proceedings of IPMU'06, Paris, July 2006.
 - [Studer, 1998] R. Studer, R. Benjamins and D. Fensel. *Knowledge Engineering: Principles and Methods*. IEEE Trans. on Data and Knowledge Eng., vol. 25, nos. 1-2, 1998, pp. 161-197.
- Lecturas recomendadas
- Hist, G. *Ontology and the Lexicon*, in S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 173-189. Springer, 2004.
 - Hotho, A., Staab, S., Slumme, G.: *Explaining text clustering results using semantic structures*. In: Proceedings of ECAI/PKDD. Springer Verlag (2003) 217-228
 - Maedche, A., and Staab, S. *Ontology learning*, in S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 173-189. Springer, 2004.
 - Natalya F. Noy and Deborah L. McGuinness. *Ontology development 101: a guide to creating your first ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.