Ingeniería de Aplicaciones para la Web Semántica

Clase 10

Ingeniería de ontologías

Mg. A. G. Stankevicius

Segundo Cuatrimestre

2005

Copyright

Contenidos

- Preguntas metodológicas.
- Construcción manual de ontologías.
- Reutilización de ontologías.
- Métodos semiautomáticos disponibles.

Preguntas metodológicas

- Al considerar la ingeniería de ontologías, surgen varias preguntas:
  - ¿Cómo hacer un mejor uso de las herramientas y las técnicas disponibles?
  - ¿Qué lenguajes y qué herramientas resultan más convenientes en un determinado dominio?
  - ¿Hay que contemplar los aspectos asociados al control de calidad y del manejo de recursos al construir ontologías?

Etapas en la construcción de una ontología

1) Determinar el alcance.
2) Reutilizar si es posible.
3) Enumerar los términos.
4) Definir la taxonomía.
5) Definir las propiedades.
6) Definir las características propias.
7) Definir las instancias.
8) Verificar si se observan anomalías.

Determinar el alcance

- ¡No existen las ontologías perfectas!
  - Una ontología es una abstracción de un dominio en particular, y por ende siempre existirán otras alternativas viables.
- La frontera de qué incluir y qué no depende principalmente de:
  - El uso que se le dará a la ontología.
  - Eventuales extensiones a futuro, que se deseen contemplar en el modelo actual.

Determinar el alcance

- En esta etapa se deben abordar los siguientes interrogantes:
  - ¿Cuál es el dominio a ser modelado por la ontología?
  - ¿Para qué será usada esa ontología?
  - ¿Qué tipo de preguntas enfrentará la ontología en desarrollo?
  - ¿Quién la usará?
  - ¿Quién estará a cargo de mantenerla?

Reutilizar si es posible

- La atención que está recibiendo la web semántica ha hecho que diversas organizaciones se hayan decidido a compilar, mantener y publicar un importante número de ontologías.
- Casi nadie tiene que comenzar de cero al desarrollar una nueva ontología.
  - Usualmente siempre existe alguna ontología desarrollada por un tercero que sirva como punto de partida del desarrollo de la misma.

Enumerar los términos

- Compilar una lista plana de todos los términos relevantes que se espera puedan aparece dentro de la ontología:
  - Los sustantivos usualmente brindan los nombres de las clases.
  - Los verbos suelen denotar la presencia de una propiedad.
- Las herramientas convencionales de la ingeniería de conocimiento pueden ser aplicadas en este ámbito.

Definir la taxonomía

- Todos los términos identificados deben ser estructurados en una jerarquía:
  - No existen consenso acerca de si conviene aplicar un esquema top-down o bien un esquema bottom-up.
- Se debe verificar que la jerarquía propuesta es de hecho una taxonomía:
  - Si **A** es subclase de **B**, entonces todas las instancias de **A** deben ser instancias de **B** (es decir, el modelo es compatible con la semántica de `rdfs:subClassOf`).

Definir las propiedades

- Esta etapa se hace usualmente a la par de la anterior.
- La semántica de `subClassOf` requiere que si **A** es subclase de **B**, entonces las declaraciones acerca de las propiedades satisfechas por las instancias de **B** deben también ser satisfechas por las instancias de **A**.
  - Sugerencia: efectuar las declaraciones acerca de las propiedades en las clases superiores en la jerarquía.

Definir las propiedades

- Es convienente aprovechar el momento en que se asocian propiedades a las clases para especificar tanto el dominio como el rango de esas propiedades.
- Notemos que aparece un compromiso entre generalidad y especificidad:
  - Por un lado conviene ser flexible para que las propiedades puedan ser heradadas.
  - Pero por otro lado, podemos perder la capacidad de detectar inconsistencias.

Definir las características propias

- En esta etapa se hace la transición de RDFS a OWL, especificando:
  - Las restricciones de cardinalidad.
  - Los valores que pueden tomar:
    - `owl:hasValue`
    - `owl:allValuesFrom`
    - `owl:someValuesFrom`
  - Las particularidades de las relaciones:
    - Simetría.
    - Transitividad.
    - Otros.

Definir las instancias

- La instanciación de las clases introducidas constituye un paso aparte.
  - Usualmente, el número de instancias suele ser mucho mayor que el número de clases.
- En consecuencia, es recomendable que la definición de las instancias de las clases se haga de forma automática:
  - Tomando como punto de partida bases de datos hoy en día obsoletas.
  - Aplicando técnicas de extracción de información a un cuerpo de datos.

Verificar si se observan anomalías

- Recordemos que una de las ventajas de OWL por sobre RDFS es la detección de anomalías o inconsistencias.
  - Ya sea en la ontología, o en la ontología junto con las instancias.
- Errores típicos detectados en esta etapa:
  - Dominios o rangos incompatibles en propiedades simetricas o transitivas.
  - Violación de los requisitos de cardinalidad.
  - Dominios o rangos en conflicto con restricciones sobre los valores posibles.

Ontologías disponibles para dominios específicos

- Dominio médico: disponemos de una ontología categorizando las variantes de cancer, propuesta por el National Cancer Institue de USA.
- Dominio cultural:
  - Art and Architecture Thesaurus, conteniendo **125.000** términos relacionados con la cultura.
  - Union List of Artist Names, con **220.000** entradas acerca de artistas.

Vocabularios integrados

- Es posible integrar ontologías desarrolladas de forma independientes en un único recurso.
- Por caso, el Unified Medical Language System integra 100 vocabularios propios de la medicina.
  - Cuenta con **750.000** conceptos y más de **10 millones** de enlaces entre los mismos.
- La semántica de un recurso obtenido por este medio es un tanto pobre.

Ontologías de amplio espectro

- Las ontologías suelen tener un dominio específico, con una frontera claramente delimitada.
- No obstante, se han ensayado algunas propuestas cuyo objetivo era definir una ontología lo más general posible:
  - El proyecto CYC (http://www.opencyc.org), que cuenta con **60000** declaraciones acerca de unos **6000** conceptos.
  - El Standard Upperlevel Ontology (http://suo.ieee.org).

Jerarquía de tópicos

- Algunas ontologías no merecen esa denominación.
  - Meros conjuntos de términos, con poca o nada estructuración.
- Se trata de jerarquías que no constituyen taxonomías, erigidas en torno relaciones altamente especificas (parte-de, es-un, etc.).
- No obstante, constituyen buenos puntos de partida para definición de ontologías.

Librerías de ontologías

- Actualmente se está tratando de compilar librerías de ontologías:
  - Sería muy raro que podamos usar una ontología sin tener que adaptarla.
  - Es usual que los conceptos contemplados deban ser refinados.
  - Es posible que se tengan que introducir nombres alternativos para las clases o las propiedades.
  - Resulta muy cómodo apelar a la posibilidad de refinar ontologías de forma privada.

El cuello de botella

- La construcción manual de ontologías es una tarea que consume tiempo, es cara y requiere personal calificado.
- Por qué no apelar a alguna técnica de aprendizaje automático que simplifiquen alguna de las tareas asociadas:
  - La adquisición de conocimiento.
  - El mantenimiento del conocimiento.

Tareas que admiten automatización

- Extracción de ontologías a partir de datos tomados de la web.
- Extracción de relaciones entre datos y metadatos tomados de la web.
- Combinación de ontologías producto del análisis de los conceptos definidos.
- Adecuación del comportamiento de las aplicaciones de la web semántica a sus usuarios.

Técnicas de aprendizaje automático

- Existen diversas técnicas de aprendizaje automático que pueden ayudar en la ingeniería de ontologías:
  - Clustering.
  - Mantenimiento incremental de ontologías.
  - Asistencia al ingeniero de conocimiento.
  - Mejorar las ontologías para el lenguaje natural.
  - Aprendizaje de ontologías.

Ontologías para el lenguaje natural

- Las ontologías para el lenguaje natural recopilan las relaciones existentes entre distintos conceptos.
  - Suelen tener un gran tamaño, y no requieren actualizaciones frecuentes.
- El estado del arte para este tipo de ontologías es muy prometedor:
  - Existen ontologías de amplio espectro.
  - Existen técnicas automáticas o semi-automáticas para desarrollarlas.

Aprendizaje automático para ontologías

- Las ontologías para dominios específicos suelen ser construídas a mano.
- Cuentan con mucho nivel de detalle.
- El aprendizaje automático de este tipo de ontologías es complejo:
  - Las técnicas de aprendizaje automático no juegan un rol preponderante.
  - Se limitan a detectar dependencias comprobadas estadísticamente y para luego sugerirselas al ingeniero de conocimiento.

Aprendizaje automático en la instanciación de ontologías

- Las instancias de la ontologías pueden ser generadas automáticamente y suelen ser actualizadas con frecuencia, aún cuando la ontología en si no lo sea.
- Este modelo encaja perfectamente en las premisas del aprendizaje automático.
- Las aplicaciones en funcionamiento usualmente:
  - Dependen totalmente de la ontología.
  - Apenas populan los marcados predefinidos.

Potenciales usos para el aprendizaje de ontologías

- Tareas asociadas al aprendizaje:
  - Creación de ontologías de la nada.
  - Extración de información ontológica.
  - Extración de información acerca de las instancias de una ontología.
- Tareas asociadas al mantenimiento:
  - Integración de ontologías.
  - Actualización de parte de una ontología.
  - Afinado de ontologías.

Tareas asociadas al aprendizaje

- Creación de ontologías de la nada.
  - ML asiste al ingeniero de conocimiento al sugerir las relaciones más importantes o bien verificando el conocimiento codificado.
- Extracción de información ontológica.
  - ML permite tomar los datos y los metadatos de web como entrada y generar ontologías listas para usar como salida, con la eventual colaboración del ingeniero de conocimiento.

Tareas asociadas al aprendizaje

- Extración de información acerca de las instancias de una ontología
  - Esta tarea consiste en inspeccionar documento en la red en busca de instancias de una cierta ontología para generar los correspondientes marcados.
  - Es una tarea análoga a la extracción de información convencional o la anotación de páginas y las técnicas desarrolladas para esos ámbitos pueden ser aplicadas a este contexto.

Tareas asociadas al mantenimiento

- Integración de ontologías.
  - Abarca el navegado y la integración de grandes bases de conocimiento ontológicas.
- Actualización de parte de una ontología.
- Afinado de ontologías.
  - Esta tarea no incluye la modificación de conceptos, sólo el ajuste fino de propiedades para hacer más precisa a la ontología.

Areas donde puede ser posible aplicar técnicas de ML

- Aprendizaje de reglas proposicionales.
- Aprendizaje bayesiano.
  - Generación de pares atributo-valor más probables.
- Aprendizaje de reglas de primer orden.
- Algoritmos de clustering.
  - Estos algoritmos permite agrupar instancias en base a su similitud, medida como la distancia entre los valores de sus atributos.