

Construcción automática de ontologías para la Web Semántica

Antonio Moreno, Aïda Valls, David Sánchez, David Isern

Grupo de Sistemas Multi-Agente (GruSMA)
Departamento de Ingeniería Informática y Matemáticas
Universidad Rovira i Virgili (URV)
ETSE. Av. Països Catalans, 26, 43007-Tarragona
Antonio.Moreno@urv.net
<http://grusma.etse.urv.es>

Resumen En este documento se hace una breve presentación del Grupo de Sistemas Multi-Agente de la URV, y se describe su principal línea de trabajo relacionada con los temas de las ontologías y la Web Semántica: la construcción automática de ontologías a partir de la información disponible en las páginas web accesibles por Internet.

1. Presentación de GruSMA

GruSMA (Grupo de Sistemas Multi-Agente) es un grupo de trabajo que forma parte del Grupo de Investigación en Inteligencia Artificial de la Universidad Rovira i Virgili. Está coordinado por los doctores Antonio Moreno y Aïda Valls, y actualmente cuenta con dos estudiantes de doctorando avanzados (David Isern y David Sánchez) y una decena de estudiantes de Ingeniería Informática. La motivación principal del grupo, fundado en el curso 1999-2000, es promover un entorno propicio para la realización de Proyectos Fin de Carrera (PFC) de primer y segundo ciclo de Ingeniería Informática en el ámbito de los agentes inteligentes y los sistemas multi-agente (SMA). Desde su inicio ya se han presentado más de 30 PFCs, que han dado lugar a numerosas publicaciones científicas.

Las líneas principales de trabajo de GruSMA son las siguientes:

- La provisión personalizada, inteligente y proactiva de servicios a los ciudadanos o turistas que visitan una ciudad. Este trabajo se enmarcó dentro de la red europea *AgentCities.NET*, IST 2000-28384, ya finalizada. El principal resultado de GruSMA fue el diseño e implementación de *HeCaSe*, un SMA que proporcionaba servicios en el ámbito de la salud ([1]).
- La aplicación de los sistemas multi-agente a problemas en el dominio de la Medicina. Algunas áreas concretas de trabajo han sido la gestión del transplante de órganos, el seguimiento automatizado de guías de práctica clínica (trabajo actual de tesis de David Isern), o la gestión, análisis y monitorización de los datos de pacientes paliativos (en el proyecto español *PalliaSys*, TIC2003-07936, coordinado por la URV, que acaba a finales de 2005, [2]).

- Implementación de agentes inteligentes en dispositivos móviles.
- El uso de ontologías para la mejora de la búsqueda de información en Internet. Este trabajo se llevó a cabo en el proyecto europeo *hTechSight* (IST 2001-33174), ya finalizado. La URV contribuyó al proyecto diseñando e implementando *MASH*, un sistema multi-agente que utilizaba la información semántica sobre un dominio, proporcionada por una ontología del mismo, para analizar los resultados de las búsquedas de páginas web asociadas a los conceptos básicos del dominio y valorar cuáles eran los que realmente tenían información relevante y de interés para el usuario ([3]).
- La construcción automática de ontologías a partir de la información accesible en páginas web.

Esta última línea de trabajo se está desarrollando principalmente en la tesis doctoral de David Sánchez, y se describe con más detalle en la siguiente sección.

2. Construcción automática de ontologías

El objetivo básico de esta línea de trabajo es la estructuración semi-automática del conocimiento almacenado en la web sobre un dominio determinado. La intención es desarrollar un sistema que funcione de forma autónoma, no supervisada, sin conocimiento previo, de forma independiente del dominio, y con la suficiente escalabilidad como para obtener resultados representativos para cualquier dominio con un rendimiento aceptable.

El proceso de construcción automática de ontologías que hemos diseñado e implementado hasta el momento tiene los siguientes pasos:

- Construcción de una taxonomía de clases del dominio.
- Detección de instancias de las clases.
- Detección de relaciones no taxonómicas.

2.1. Jerarquía de clases e instancias

Para obtener la jerarquía de clases y sus instancias, aplicamos el algoritmo mostrado en la figura 1. El funcionamiento básico del algoritmo es el siguiente (se pueden obtener más detalles en [5], [7] o [8]):

- El usuario proporciona como parámetros de entrada un *keyword* que represente el dominio de interés (p.e. *cancer*) y algunos parámetros de control del proceso (p.e. el número de páginas web a analizar por cada clase).
- Utilizamos un buscador Web estándar para recuperar un conjunto de páginas web donde aparezca el concepto dado por el usuario. Se hace un ligero análisis sintáctico para detectar candidatos a subclases del concepto. Los candidatos se detectan estudiando los sustantivos que aparecen justo delante del concepto dado (p.e. *lung cancer* sería un candidato a subclase de *cancer*). Para cada candidato se almacenan algunos datos, como el número de páginas donde ha aparecido.

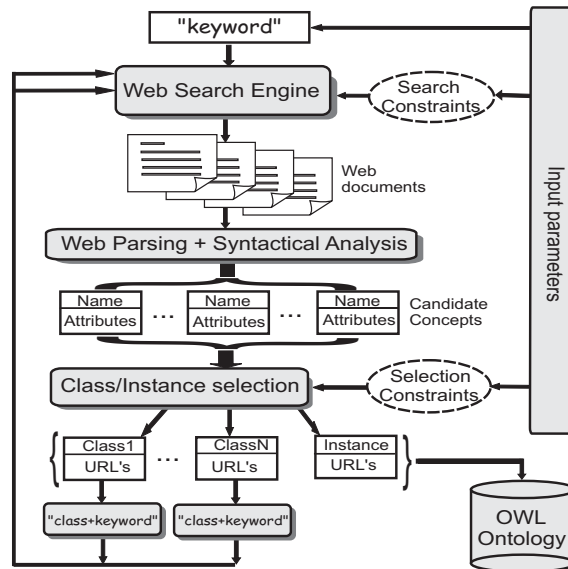


Figura 1. Algoritmo de construcción de taxonomías.

- Se evalúa cada uno de los candidatos, utilizando estadísticos proporcionados por el buscador Web, para decir cuáles son realmente relevantes. Éstos pasan a ser subclases del concepto dado. En el proceso de selección también se aplican técnicas para distinguir las subclases de las instancias del concepto (básicamente aprovechando el hecho de que las instancias suelen aparecer en mayúsculas).
- Para cada una de las subclases encontradas, se vuelve a aplicar el mismo proceso para seguir construyendo la jerarquía.

2.2. Detección de relaciones no taxonómicas

Cuando se construye la taxonomía de clases, se almacenan para cada una de ellas un conjunto de frases que relacionan la clase con otros conceptos (*text nuggets*). Por ejemplo, para la clase *colon cancer* podríamos haber almacenado la frase *Colon cancer starts as polyp*, tras haber notado que aparece en varias de las páginas web analizadas para ese concepto. Esta frase nos indica que existe una relación no taxonómica (*starts as*) entre un concepto conocido (*colon cancer*) y un nuevo concepto (*polyp*). Hemos desarrollado un sistema multi-agente, descrito con más detalle en [4], en el que tras analizar la frase se genera dinámicamente un agente que se encarga de analizar el nuevo concepto y construir su jerarquía de clases asociada (empleando el algoritmo de la fig.1). Este sistema multi-agente permite construir ontologías como la mostrada en la figura 2, donde se pueden apreciar relaciones no taxonómicas entre conceptos de diversas jerarquías.

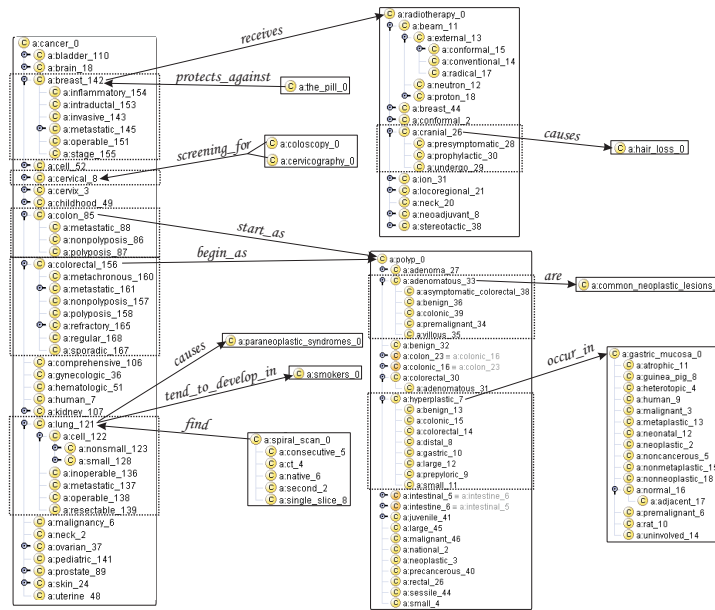


Figura 2. Ontología con relaciones no taxonómicas.

Referencias

1. Moreno, A., Isern, D., Sánchez, D.: Provision of agent-based health care services. *AI Communications*, Vol. 16, No. 3, 167–178, 2003.
2. Moreno, A., Valls, A., Riaño, D.: Agent-based alarm management in a Palliative Care Unit. III Workshop on Agents Applied in Health Care, en *IJCAI-2005*, Edinburgh. *Actas del workshop*, 60–66, 2005.
3. Riaño, D., Moreno, A., Isern, D., Bocio, J., Sánchez, D., Jiménez, L.: Knowledge exploitation from the Web. IV Int. Conference on Practical Aspects of Knowledge Management, *PAKM 2004*, Viena. *LNAI 3336*, 175–185, Springer Verlag, 2004.
4. Sánchez, D., Isern, D., Moreno, A.: An agent-based knowledge acquisition platform. Workshop on Cooperative Information Agents, *CIA 2005*, Koblenz. *LNAI 3550*, 118–129, Springer Verlag, 2005.
5. Sánchez, D., Moreno, A.: Web-scale taxonomy learning. *Actas del workshop Learning and extending lexical ontologies by using Machine Learning methods* en la 19th International Conference on Machine Learning, *ICML 2005*, Bonn.
6. Sánchez, D., Moreno, A.: Web mining techniques for automatic discovery of medical knowledge. Poster en 10th Conference on Artificial Intelligence in Medicine, *AIME 2005*, Aberdeen. *LNAI 3581*, 409–413, Springer Verlag, 2005.
7. Sánchez, D., Moreno, A.: Creating Ontologies from Web Documents. VII Congrès Català d'Intel·ligència Artificial, *CCIA 2004*, Barcelona. *Frontiers in Artificial Intelligence and Applications* 113, 11–18. IOS Press, 2004.
8. Sánchez, D., Moreno, A.: Automatic generation of taxonomies from the WWW. IV Int. Conference on Practical Aspects of Knowledge Management, *PAKM 2004*, Viena. *LNAI 3336*, 208–219, Springer Verlag, 2004.