Ontologías en el Proceso de Descubrimiento de Conocimiento en Bases de Datos

Héctor Oscar Nigro, Sandra González Císaro, Daniel Xodo

INTIA- Departamento de Computación y Sistemas
Facultad de Ciencias Exactas - UNICEN
Campus Universitario - Paraje Arroyo Seco s/n
B7001BBO Tandil, Buenos Aires, ARGENTINA
TEL: +54-2293-439680 – FAX: +54-2293-439681
e-mail: {onigro, sagonci , dxodo}@exa.unicen.edu.ar

Resumen

En este proyecto daremos las bases para la investigación y el desarrollo del proceso de Descubrimiento de Conocimiento en Bases de Datos con Ontologías. La principal motivación para la inclusión de ontologías en dicho proceso es la necesidad de incluir el conocimiento previo en las sesiones de minería. Dicho conocimiento puede ser provenir del proceso mismo o del dominio de aplicación involucrado.
Nuestro objetivo es el mejoramiento integral del proceso, a partir de un mejor entendimiento del dominio de aplicación, de los resultados obtenidos en sesiones previas y de la aplicación de la o las técnicas más convenientes de acuerdo a problema a resolver.

1) Introducción

Descubrimiento de Conocimiento en Bases de Datos (KDD) se define como la extracción, no trivial, de información previamente desconocida y potencialmente útil, en grandes colecciones de datos (Fayyad et al, 1996). Puede ser considerado como una búsqueda de reglas interesantes, patrones o excepciones en grandes colecciones de datos. Es un área interdisciplinaria sustentada por diversos campos, tales como: Estadística, Bases de Datos, Aprendizaje Automático, Inteligencia Artificial, Teoría de la Información, Computación Paralela y Distribuida y Visualización entre otros.

De acuerdo a la bibliografía (Fayyad et al., 1996; Han et al., 2001; Hernández Orallo et al, 2004) las técnicas más frecuentes pueden ser catalogadas en:

- *Descriptivas*: El objetivo de estos procedimientos es la búsqueda de la caracterización o discriminación de un conjunto de datos. Las técnicas más conocidas *son*: Agrupamiento o Clustering, Reglas de Asociación, Análisis de Patrones Secuenciales, Análisis de Componentes Principales, Detección de Desviación.
- *Predictivas*: El propósito de estos métodos es aprender una hipótesis la cual pueda clasificar a nuevos individuos. Los algoritmos principales son: Regresión y Clasificación (Árboles de Decisión, Clasificación Bayesiana, Redes Neuronales, Algoritmos Genéticos, Conjuntos y Lógica Difusa).

Cuando realizamos un proceso de Minería de datos, necesitamos tener en cuenta el conocimiento previo; este puede derivar del proceso mismo(elección de variables, técnicas, algoritmos, interpretación de resultados) o del dominio de aplicación.

Actualmente, en la mayor parte de proyectos de KDD, el conocimiento previo está sólo presente implícitamente (en la mente del analista humano) o en la forma de documentación textual. Inclusive en acercamientos intensivos al conocimiento como ILP (Induction Language Programming), los conocimientos previos, a menudo, no son organizados alrededor de un modelo conceptual gramaticalmente correcto. Esta práctica parece no hacer caso del último desarrollo en la Ingeniería de Conocimiento, donde el conocimiento del dominio es típicamente definido por ontologías formales.

En ambientes distribuidos, las ontologías son usadas para construir el servicio semánticamente rico en descripciones. Técnicas para planificación, composición, edición, razonando y el análisis sobre estas descripciones está siendo investigado y desplegado para resolver la interoperabilidad semántica entre servicios(Canataro, et al, 2003).

Por lo expuesto, podemos observar que uno de los problemas más importantes y desafiantes a ser investigado en Minería de Datos es, la definición del conocimiento previo. Nuestra investigación se centrará en la utilización de Ontologías para la representación del conocimiento previo, ya sea durante el proceso o para la representación del dominio de aplicación.

2) Desarrollo

El Descubrimiento de Conocimiento en Bases de Datos es un proceso exploratorio que involucra la aplicación de varios procedimientos algorítmicos para la manipulación de datos, construcción de modelos desde los datos y la manipulación de los mismos. El proceso de Descubrimiento de Conocimiento (KD) (Fayyad, et al., 1996) es una de las nociones centrales del campo de Descubrimiento de Conocimiento y Data Mining (KDD).

El proceso KD es digno de la mayor atención en la comunidad científica. Una razón es que los procesos comprenden múltiples componentes algorítmicos, que interactúan en recorridos no triviales. Todos los especialistas en Minería de Datos no se encuentran familiarizados con todo el rango de componentes, y asimismo con el vasto espacio de diseño, de procesos posibles (Bernstein et al., 2005).

No obstante, tanto novicios como especialistas en Minería de Datos, se encuentran aptos para una utilización más abstracta o de un nivel mayor de generalización de las instancias de un proceso KD. Se consideran herramientas que ayuden a los profesionales en Minería de Datos a navegar por el espacio de procesos KD, de una manera más sistemática, y más eficiente.

En particular el proyecto de investigación se centraliza en un subconjunto de estados de los procesos de KD (estos estados a su vez tienen múltiples componentes de algoritmos que pueden ser aplicados). A este proceso le denominamos Minería de Datos, distinguido del proceso más extenso de Descubrimiento de Conocimiento en Base de Datos. El proceso de KD se encuentra descrito por Fayyad et al. (1996) y Chapman et al., (2000). Hay que poner énfasis en tres procesos de KD: preproceso automático de datos, aplicación de algoritmos de inducción, y post-proceso automático de modelos.

Se selecciona este conjunto de pasos, porque individualmente, se encuentran relativamente bien comprendidos y pueden ser aplicados a una amplia variedad de conjunto de datos.

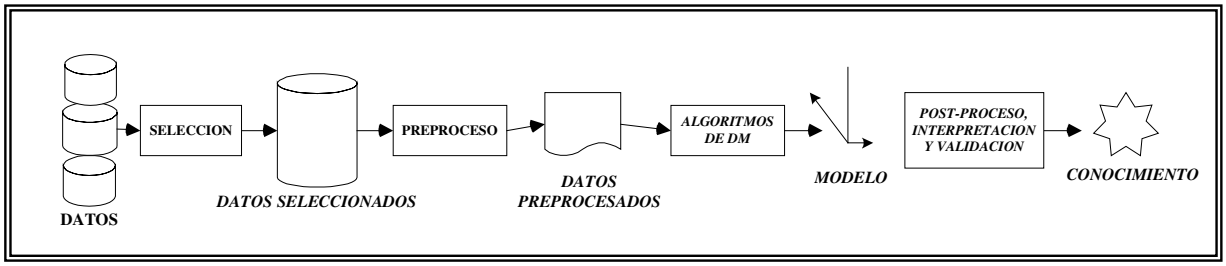

Figura 1 El Proceso de descubrimiento según Fayyad(KD)

Considerando la necesidad de incluir el conocimiento dentro del proceso de descubrimiento por medio de las Ontologías (definidas por Gruber como: "Especificación formal explícita de una conceptualización compartida"). Vemos que la base ontológica es una condición previa para el uso automatizado eficiente de ese conocimiento.

En la figura se pueden observar las áreas donde se pueden aplicar las ontologías en el

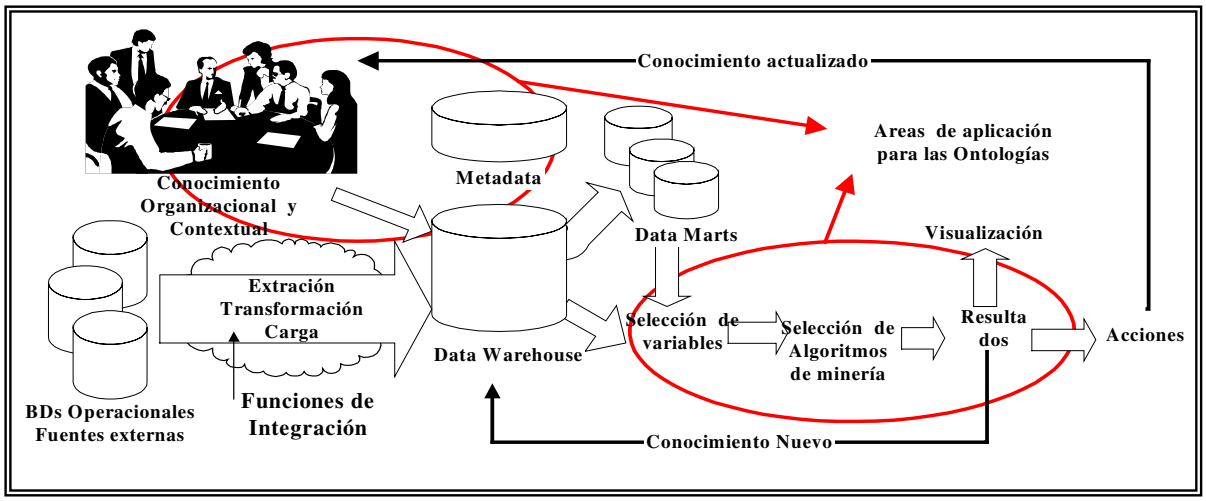

Figura 2 Áreas de aplicación para las Ontologías en el Proceso KDD

proceso de descubrimiento de conocimiento.

De esta manera, podemos ver a la relación entre Ontologías y Minería de Datos de dos modos:

- ***Desde las Ontologías a la Minería de Datos***, incorporamos el conocimiento al proceso por el uso de ontologías, es decir como los expertos entienden y realizan las tareas de análisis. Las aplicaciones representativas son ayudantes inteligentes para el proceso de descubrimiento, la interpretación y la validación del conocimiento extraído, Ontologías para recursos y descripción de servicios y Knowledge Grids(Hotho et al., 2003; Bernstein et al., 2005; Cannataro et al. 2003, 2004, 2007; Gridminer Project; Gottgtroy et al., 2005; Rennolls, 2005).
- ***Desde la Minería de Datos a Ontologías***, incluimos el conocimiento del dominio en la información de entrada o usamos las ontologías para representar los resultados. Por lo tanto el análisis es realizado sobre estas ontologías. Las aplicaciones más representativas están en Medicina, Biología y Datos Espaciales, como: la representación de Genes, Taxonomías, aplicaciones en Geociencias, aplicaciones médicas (Breaux et al., 2005; Tadepalli et al., 2004; Bogorny et al., 2005, 2006; Sidhu et al., 2006).

En primer término, nuestra investigación se centrará en la definición y clasificación de ontologías dentro del proceso de descubrimiento de conocimiento en Base de Datos con el objetivo de mejorar la performance del proceso en sí mismo, como así también, la calidad de los descubrimientos realizados.

La segunda etapa de este proyecto, consiste en el diseño y desarrollo de una herramienta que abarque:

- Base de conocimiento ontológico para el dominio de aplicación.

- Base de conocimiento ontológico para las técnicas de Minería o estadística empleadas. Esta base será empleada en la implementación del Asistente Inteligente de Descubrimiento ideado por Bernstein(2005).
- Funciones de aprendizaje sobre la utilización de la herramienta, lo que nos permitirá mejoras para distintos perfiles de usuario
- Funciones de meta aprendizaje para la evaluación de cada uno de los modelos inducidos.
- Base Conocimiento conteniendo los patrones descubiertos.

El proyecto pretende prestar a un usuario de Minería de Datos, al menos, los siguientes beneficios:

- Una enumeración sistemática de los procesos de Minería válidos, no sólo los importantes, sino aquellos potencialmente utilizables.
- Un orden efectivo de dichos procesos válidos según criterios diferentes y una ayuda para seleccionar entre las distintas opciones.
- Una infraestructura para segmentar el conocimiento de Minería, algo que los economistas denominan redes externas.
- Un soporte arquitectónico genérico para permitir la inclusión del conocimiento del dominio en las sesiones de minería.

3)Temas involucrados en el proyecto

Las áreas incluidas en el proyecto son: 1)Data Warehouse, 2)Bases de Datos, 3)Estadística, 4)Análisis de Datos, 5)Ingeniería del Conocimiento, 6)Data Mining, 7)Inteligencia Artificial, 8)Sistemas Inteligentes, 9)Aprendizaje Automático, 10) Ingeniería Ontológica, 11)Agentes Inteligentes, 12) Visualización de datos.

4) Conclusiones

En la actualidad, el conocimiento tiene una importancia relevante en las organizaciones, razón por la cual es necesario optimizar el proceso de su descubrimiento. El conocimiento empírico o heurístico, ya sea proveniente del dominio de aplicación o del conocimiento de las técnicas, mejora el resultado de los modelos inducidos en el proceso de Minería de Datos. Nuestra meta es captar ese conocimiento y representarlo mediante ontologías.

La inclusión de la Ingeniería Ontológica en el proceso de descubrimiento, nos permitirá la integración de diferentes técnicas de Minería de Datos, como así también su uso adecuado.

Creemos que la inclusión de las Ontologías no solo mejorará la performance del proceso sino que también mejorará la calidad de los conocimientos descubiertos. Podemos considerar entonces, que ya no estamos realizando Minería de Datos sino Minería de Conocimientos.

Referencias

1. Bernstein A., Provost F. y Hill S. (2005). "Towards Intelligent Assistance for the Data Mining Process:An Ontology-based Approach for Cost/Sensitive Classification". En IEEE Transactions on Knowledge and Data Engineering 17(4), pag.503-518, Abril 2005.

2. Bogorny, V.; Engel, P. M.; Alvares, L.O. (2005). A reuse-based spatial data preparation framework for data mining. In J. Debenham, K. Zhang (Eds.), *Fifteenth International Conference on Software Engineering and Knowledge Engineering* (pp. 649-652). Taipei: Knowledge Systems Institute

3. Bogorny, V.; Camargo, S.; Engel, P. M.; Alvares, L.O. (2006). Towards elimination of well known geographic domain patterns in spatial association rule mining. *In Third IEEE International Conference on Intelligent Systems* (pp. 532-537). London: IEEE Computer Society.

4. Breaux T. y Reed J. (2005). "Using Ontology in Hierarchical Information Clustering". En Proceedings of the 38 Hawaii International Conference on System Sciences.

5. Cannataro M. y Comito C.(2003). "A Data Mining Ontology for Grid Programming". En I Workshop on Semantics Peer to Peer and Grid Computing. Budapest, 20/24 Mayo, 2003. http://www.isi.edu/~stefan/SemPGRID.

6. Cannataro, M.; Congiusta, A.; Pugliese, A.; Talia, D.; Trunfio, P., Distributed Data Mining on Grids: Services,Tools, and Applications, IEEE Transactions on Systems, Man and Cybernetics, Part B, 34(6): 2451- 2465, December 2004

7. Cannataro M., Guzzi P. H., Mazza T., Tradigo G. y P. Veltri(2007), Using ontologies for preprocessing and mining spectra data on the Grid. Future Generation Computer Systems, 23(1),. pp. 55-60.

8. Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T.,Shearer C., and Wirth R., CRISP-DM 1.0: Step-by-step data mining guide, SPSS White paper– technical report CRISPWP-0800, SPSS Inc., 2000

9. Fayyad U., Piatetsky-Shiapiro G., Smyth P. y Uthurusamy R. (1996). "Advances in Knowledge Discovery and Data Mining". Merlo Park, California: AAAI Press.

10. Gottgtroy P., MacDonell S., Kasabov N y Jain V. (2005). "Enhancing data analysis with Ontologies and Olap". Proceedings Data Mining 2005 Conference, 25-28 Mayo, 2005, Skialhos, Grecia.

11. Gridminer Project. http://www.gridminer.org

12. Gruber T. (2002). What is an Ontology? http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

13. Han J. y Kamber M. (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann.

14. Hernández Orallo J., Ramírez Quintana M y Ferri Ramirez C. (2004) "Introducción a la Minería de Datos". Editorial Pearson Educación SA, Madrid.

15. Hotho, A.; Staab, S. & Stumme, G. (2003). Ontologies Improve Text Document Clustering. In Proceedings of the 3rd IEEE Conference on Data Mining, Melbourne, FL, USA, pp.541-544.

16. Noy, N, McGuinness D. (2000); Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, Stanford, CA.

17. Pan, D. & Pan Y. (2006). Using Ontology Repository to Support Data Mining. In Proceedings of the Sixth World Congress on Intelligent Control and Automation, June 21-23, 2006 in Dalian, China. WCICA 2006, pp. 5947 - 5951

18. Rennolls, K. (2005). An Intelligent Framework (O-SS-E) For Data Mining, Knowledge Discovery and Business Intelligence. Keynote Paper. In Proceeding 2nd International Workshop on Philosophies and Methodologies for Knowledge Discovery, PMKD'05, in the DEXA'05 Workshops pp. 715-719. IEEE Computer Society Press. ISBN 0-7695-2424-9.

19. Sidhu, A. S., Dillon, T. S. & Chang, E. (2006) Advances in Protein Ontology Project. 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2006). Salt Lake City, Utah, IEEE CS Press.

20. Tadepalli S., Sinha, A.K., y Ramakrishnan N (2004). "Ontology Driven Data Mining for Geoscience". Annual Meeting and Exposition of the Geological Society of American, November 7–10, 2004 Denver USA. http://gsa.confex.com/gsa/2004AM/finalprogram/index.html.