El Proceso de Minería de Datos Asistido por Ontologías

Héctor Oscar Nigro, Sandra González Císaro

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 }@exa.unicen.edu.ar

Resumen

En este artículo mostraremos los avances obtenidos en la investigación y el desarrollo del proceso de Minería de Datos asistido por Ontologías. Además expondremos un modelo general para la aplicación de las ontologías, como así también, los tipos de ontologías planteadas.
La principal motivación para la inclusión de ontologías en dicho proceso es la necesidad de incluir el conocimiento previo en los estudios de minería. Dicho conocimiento puede ser provenir del proceso mismo o del dominio de aplicación comprendido.
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

La minería de los datos se ocupa del uso de las técnicas y de las metodologías del análisis de datos en el diseño, en el desarrollo y de la evaluación de los datos con el objetivo de hallar nuevos conocimientos. 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 (Fayyad et al., 1996; Han et al., 2001; Hernández Orallo et al, 2004).

El principal desafío que hoy enfrenta el área de Minería de Datos es la inclusión del conocimiento previo en cada sesión de minería. Este conocimiento previo puede ser contextual o del proceso mismo. Una de las formas más convenientes para la inclusión de este conocimiento está dada por las ontologías (Nigro et al., 2008). 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.

La mayoría de las propuestas o soluciones que encontramos en Minería de Datos con ontologías son parciales, es decir, se centran en algunos de los pasos del proceso de descubrimiento del conocimiento. Por ejemplo Euler y Scholz (2004) presentan un meta-modelo de las secuencias del preproceso conteniendo una ontología que describe el conocimiento conceptual del dominio. Este meta-modelo es operacional, lo suficientemente abstracto para permitir la reutilización de los usos exitosos en dominios similares. Bernstein et al. (2005) proponen una asistente inteligente basado en ontologías para guiar secuencias validas del proceso. Pan and Shen (2006) han propuesto una arquitectura para el descubrimiento del conocimiento en ambientes evolutivos. La arquitectura

crea un mecanismo de comunicación para incorporar conocimiento previo en el proceso del descubrimiento a través de facilidad del servicio de las ontologías. La continua exploración minera es transparente al usuario final; por otra parte, la arquitectura apoya independencia de datos lógica y física. Brisson y Collar (2007) presentan el proyecto KEOPS, el cual integra todo el conocimiento experto a lo largo del proceso de minería de datos en una manera coherente y uniforme.

Cannataro et al. (2003, 2004, 2007) proponen el uso de ontologías en ambientes distribuidos de minería de datos. Las ontologías son usadas para construir los servicios semánticamente ricos en descripciones. Técnicas para planificación, composición, edición, razonamiento y análisis sobre estas descripciones están siendo investigadas y desplegadas para resolver la interoperabilidad semántica entre servicios.

En otra dirección se encuentran los trabajos que incluyen el conocimiento del dominio en la información de entrada o usan 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 geográficos, 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; Pan et al., 2006).

2) Desarrollo

La naturaleza interactiva es inherente al proceso de minería de datos puesto que la comunicación con los expertos es necesaria para entender el dominio y para interpretar resultados en forma adecuada (Brisson et al, 2007).

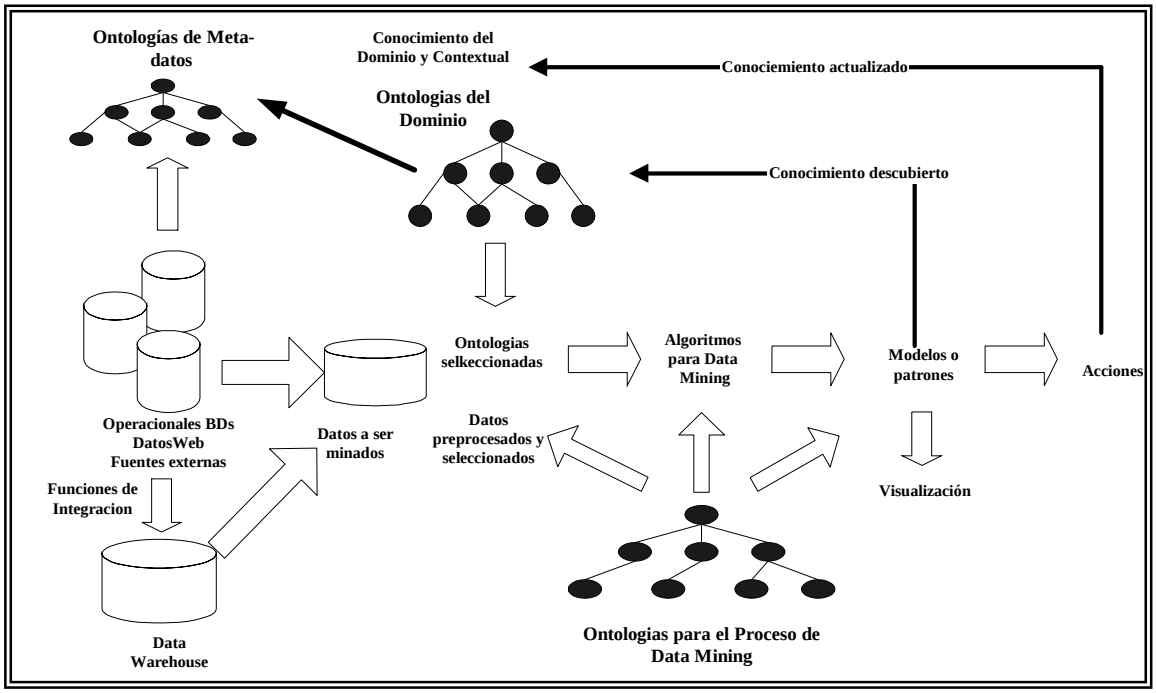

Figura 1 Modelo de Minería de Datos Asistido por Ontologías

Considerando la necesidad de incluir el conocimiento dentro del proceso de minería asistido por ontologías (definidas éstas por Gruber -2002- 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. Así, hemos elaborado un modelo que involucre todos los pasos del descubrimiento (ya sea considerando el modelo de Fayyad o el de Crisp-DM – Chapman et al., 200), el cual esta basado en tres tipos de ontologías:

- *de Dominio*
- *para Meta-datos*
- *del Proceso*

¿Por que estos tres tipos ontologías y como se relacionan?

Ontologías de Dominio: estas expresan el conocimiento del dominio de aplicación. Generalmente los científicos trabajan con diagramas causa efecto para expresar sus ideas respecto a sus hipótesis de trabajo, estos diagramas pueden ser fácilmente mapeados a mapas de conceptos y luego a ontologías.

Ontologías para Meta-datos: codifican el o los procesos que son necesarios llevar a cabo para la construcción de las variables a ser usadas en la sesión de minería.

Ontologías del Proceso: codifican el conocimiento sobre la secuencia válida de operaciones a ser realizadas. Puede ser teniendo en cuenta los datos de entrada, considerando cual es la visualización que el analista tiene del resultado esperado o ambos.

El modelo propuesto permite la interacción entre las ideas del analista tales como hipótesis del trabajo, el tipo de modelo de salida deseado y el conocimiento previo. Puesto que el analista puede visualizar el dominio y los ontologías de los meta-datos, puede entonces aprender sobre su relación y características. Con este conocimiento en mente, selecciona la/s técnica/s. Las ontologías para el proceso ayudan al usuario en la elección de las variables más convenientes, instancias de datos y algoritmos para desarrollar su modelo según los parámetros - técnica elegida, el tipo variable, precisión, exactitud, sensibilidad del costo, comprensibilidad, matriz de los datos, características de las ontologías del dominio y de los meta datos, entre otros.

Una vez que se haya seleccionado el algoritmo, los pasos de pre-procesamiento se aplican a los datos de entrada - u ontologías de dominio con la ayuda de las Ontologías para el proceso. Entonces el algoritmo se aplica al modelo preprocesado. Además las Ontologías del proceso deben brindar los pasos de post-procesamiento y las visualizaciones más apropiados para el modelo obtenido; se evalúa y se visualiza el modelo de la salida. El analista puede decidir el cambio de las características de las ontologías del dominio si un nuevo conocimiento aparece en el modelo de salida obtenido.

Nuestra visión del proceso con ontologías se orienta más al Descubrimiento Computacional de Conocimiento Científico (Computational Discovery of Scientific Knowledge desarrollado por Langley -2000, 2006-), que al tradicional de Fayyad; ya que:

- El conocimiento puede también asistir en la búsqueda de las características útiles (es decir, colocando restricciones en combinaciones aceptables de atributos, proporcione un sistema inicial de las variables sobre las cuales comenzar la búsqueda, predisponiendo la selección a la producción de modelos más comprensibles)
- Este acercamiento produce modelos más exactos y más fáciles de comprender que los inducidos desde la prueba y el error.
- La utilización del conocimiento para influenciar el descubrimiento pueden reducir el error de la predicción y también puede mejorar el entendimiento del modelo.
- Es un proceso intensivo de la Ingeniería del Conocimiento, con la intervención humana en la interpretación y la validación.

La próxima etapa de este proyecto, consistirá en el diseño y desarrollo de una aplicación que comprenda:

- 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.

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) Interacción Hombre-Maquina, 9) Sistemas Inteligentes, 10) Aprendizaje Automático, 11) Ingeniería Ontológica, 12) Agentes Inteligentes, 13) Visualización de datos.

4) Conclusiones

Dada la importancia actual del conocimiento, pretendemos con este proyecto asistir al usuario de Minería de Datos en los procesos de descubrimiento; brindándole una enumeración sistemática de los procesos de Minería válidos, no sólo los importantes, sino aquellos potencialmente utilizables. El orden efectivo en el que esos procesos deben realizarse según criterios diferentes.

Además, ofrecer una infraestructura y un soporte arquitectónico que permita la inclusión del conocimiento del dominio, la reusabilidad y la segmentación del mismo (esto es conocido por los economistas como redes externas).

Consideramos que toda ontología de minería de datos, estadística o de análisis de datos, en general, debe tener en cuenta las categorías del conocimiento de las técnicas de análisis de datos, las categorías del conocimiento del dominio y de los procedimientos de la investigación empírica.

Las ventajas de la utilización de las ontologías en el proceso están dadas por: la reutilización del conocimiento del dominio, modelos más compresivos, lenguaje común para la comunicación entre las aplicaciones y los expertos. Particularmente, nos centramos en cómo las ontologías pueden ayudar a construir modelos que se obtienen a través de un proceso intensivo de uso del conocimiento y no por un proceso de prueba y del error.

Referencias

1. Bay, S. D., Shapiro, D. G., & Langley, P. (2002). Revising engineering models: Combining computational discovery with knowledge. Proceedings of the Thirteenth European Conference on Machine Learning. Helsinki, Finland, pp. 10-22.

2. 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.

3. 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

4. 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.

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

6. Brisson L. & Collard M.(2007). An Ontology Driven Data Mining Process. Research report of University of Nice, France.

7. 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.

8. 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

9. 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.

10. 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

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

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. Langley, P. (2000) The computational support of scientific discovery. International Journal of Human-Computer Studies 53, pp. 393-410.

16. Langley P. (2006) Knowledge, Data, and Search in Computational Discovery. Invited talk at International Workshop on feature selection for data mining: Interfacing machine learning and statistics, (FSDM) April 22, 2006, Bethesda, Maryland in conjunction with 2006 SIAM Conference on Data Mining (SDM). 2006

17. Nigro H. O., González Císaro S. & y Xodo D. Eds (2008) "Data Mining with Ontologies: Implementations, Findings and Frameworks", Publisher: Information Science Reference. ISBN 978-1-59904-618-1

18. 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

19. Pan, D and Shen, J. Y.(2005) Ontology service-based architecture for continuous knowledge discovery. In Proceedings of International Conference on Machine Learning and Cybernetics, Volume 4, pp. 2155 – 2160. IEEE Press. 18 - 21 August 2005.

20. 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.

21. 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.