"Modelo de integración y preprocesamiento de información para el descubrimiento de conocimiento en bases de datos federadas utilizando ontologías"

Geovanna Evelyn Espinoza Taype
Universidad Católica de Santa María – Arequipa
Av. Arequipa 1516. Departamento F (Jesús María – Lima)

RESUMEN
El aumento de la información en las empresas ha provocado un incremento de datos en sus sistemas de información, sin embargo no ha implicado un aumento proporcional del conocimiento disponible. Para obtener este conocimiento es necesario avanzar en el proceso de descubrimiento de conocimiento en bases de datos. El presente modelo propone el uso de ontologías en las fases previas a la minería de datos dentro de un proceso de descubrimiento de conocimiento. Una de las ventajas más importantes de este enfoque es poder acceder de forma homogénea a información de varias fuentes heterogéneas unificadas y semánticamente coherentes.

ABSTRACT:
The increase of information in companies has caused an increment of data in her information system, however these hasn't involved an increase proportional of knowledge available, for get these knowledge is necessary improve in the process of knowledge discovery in databases. The present model propose the use of ontology's in the first phases in the data mining into to process of knowledge discovery. The advantage more important is the Access to databases heterogeneous through a Homogeneous interface with databases unified and coherent semantically.

1. INTRODUCCIÓN

Durante los últimos años se ha producido un aumento exponencial de la información que producen las organizaciones, lo que ha impulsado el desarrollo de soluciones cada vez más eficientes para su almacenamiento. Sin embargo, el conocimiento utilizado en los procesos de toma de decisiones no ha experimentado un incremento proporcional al de la información disponible. El descubrimiento de nuevo conocimiento impulsa, por tanto, la investigación de este modelo, con la aplicación en métodos avanzados de búsqueda, integración y procesamiento de información. Esta investigación es patrocinada por Microsoft Student Partners y Bitcomm Research Center, el cual propone un modelo para la integración y preprocesamiento de información en bases de datos federadas proponiendo la utilización de Ontologías en las fases previas a la minería de datos dentro de un proceso de descubrimiento de conocimiento.

MODELO DE INTEGRACIÓN Y PROCESAMIENTO DE LA INFORMACIÓN

El modelo propuesto para la integración y preprocesamiento de la información se basa en la utilización de ontologías para alcanzar la mayor expresividad posible. Las ontologías se utilizan como mecanismos para resolver diversos tipos de heterogeneidades presentes, incluyendo las semánticas en fuentes estructuradas, para manejar las relaciones entre los datos. El modelo propuesto aborda la integración en dos niveles:

(i) Integración de esquemas, en la que se centran la mayoría de los proyectos de

investigación en el área, y donde se suelen construir nuevas vistas virtuales de las bases de datos físicas.

(ii) Integración de instancias y preprocesamiento, donde se realizan las transformaciones necesarias de los valores contenidos en las bases de datos para homogeneizar los resultados. En este último caso se incluyen también cuestiones del preprocesamiento de datos, ya que aunque son fases que no tienen por qué abordarse al mismo tiempo, comparten objetivos desde un punto de vista funcional.

El modelo propuesto se guiara en base al siguiente esquema, y tomando como punto de partida, véase la figura 1 donde se muestra el modelo general propuesto para la integración y preprocesamiento de bases de datos federadas.

Figura 1. Modelo propuesto de KDD basado en Ontologías

Este modelo incluye la integración y preprocesamiento de la información y tiene como entradas:

(i) Fuentes de datos,
(ii) Consultas de los usuarios y
(iii) Ontologías que dirigen las transformaciones.

Como salida devuelve una base de datos general con datos integrados y preprocesados de forma transparente al usuario.

UNIFICACIÓN DE REPOSITORIOS VIRTUALES

Una vez detectadas las inconsistencias de las distintas fuentes de datos, y almacenadas en sus correspondientes ontologías, el proceso de su unificación es automático. Consiste en la fusión de distintos esquemas virtuales correspondientes a varias fuentes de datos en un esquema virtual unificado. Estos esquemas virtuales unificados son ontologías que reflejan la estructura conceptual de la información almacenada en varias bases de datos de forma transparente al usuario. La Figura 3 ilustra el funcionamiento general de los procesos de "mapping" y unificación, así como los elementos que intervienen.

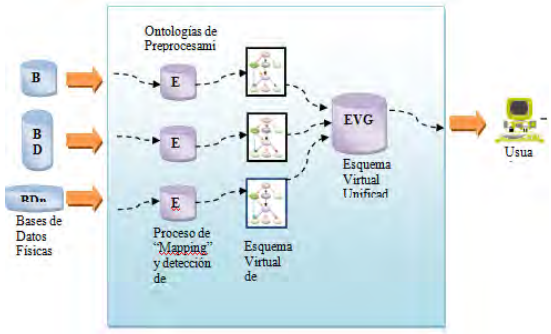

Figura 3. Procesos de integración y preprocesamiento.

2. RESULTADOS Y DISCUSIÓN

Se ha observado que las distintas baterías de consultas ejecutadas sobre la unificación virtual de fuentes heterogéneas devuelven datos correctamente integrados y preprocesados. Lo que demuestra la viabilidad del modelo propuesto basado en ontologías. Aparte del hecho de que las instancias de datos recuperadas sean correctas con respecto a las fuentes. Este proceso prueba la fiabilidad del modelo propuesto de "mapping" y unificación. Sin embargo, de las pruebas también se ha concluido que no es viable integrar fuentes

de datos de dominio muy distintos, ya que las unificaciones obtenidas serían demasiado genéricas, y se podría perder la correspondencia entre los esquemas virtuales que describen los datos y los datos físicos. La mayor ventaja en todo caso, es poder acceder de forma homogénea a la información de varias fuentes heterogéneas unificadas.

3. CONCLUSIONES

- El planteamiento de un modelo de integración y preprocesamiento de Información ha permitido integrar esquemas e instancias de datos que ha permitido obtener datos de calidad.
- El método para homogenizar los esquemas heterogéneos ha permitido la obtención de un esquema general global el cual funge como meta modelo.
- Las Ontologías juegan un papel muy importante en la integración de bases de datos federadas, a esta afirmación se ha llegado tras la creación, implementación y evaluación de un nuevo modelo de KDD basado en ontologías, centrado en el preprocesamiento e integración de esquemas e instancias de fuentes de datos heterogéneas.
- Debemos darle mayor importancia a las primeras etapas de KDD debido a que en esta fase se obtiene la esencia de los datos y de esto dependerá las futuras decisiones que se lleguen a tomar en un proceso de Data Mining.

AGRADECIMIENTOS

Microsoft Student Partners y Bitcomm Research Center, por su apoyo para el desarrollo del modelo.

REFERENCIAS Y BIBLIOGRAFÍA

[BAL02] **Baluarte, Araya César**. *Tecnología de Información (texto universitario).* 2002.

[GHJ97] **Gary W. Hansen, James V. Hansen.** *Diseño y Administración de bases de datos***.** Segunda Edición. ISBN: 0-13-308800-6. Madrid **-** 1997.

[OGX05] **Oscar Nigro Héctor, González Císaro Sandra, Xodo Daniel.** *Ontologías en el Proceso de Descubrimiento de Conocimiento en Bases de Datos.* INTIA-UNICEN. [Publicación] Argentina - 2005.

[PDD07] **Pérez del Rey David.** *Sobre Un modelo de integración y preprocesamiento de información distribuida basado en ontología*. [Doctorado**]**, Madrid - 2007.