PROCEDIMIENTOS DE LA EXPLOTACIÓN DE INFORMACIÓN PARA LA IDENTIFICACIÓN DE DATOS FALTANTES, CON RUIDO E INCONSISTENTES

Horacio Kuna, Ramón García-Martínez, Francisco Villatoro Machuca

Depto. de Informática, Facultad de Ciencias Exactas Químicas y Naturales. Universidad Nacional de Misiones.
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires.
Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga.

hdkuna@unam.edu.ar, rgarciamar@fi.uba.ar

CONTEXTO

Está línea de investigación se articula con los proyectos UBACyT 2008-2010-I012 “Aplicaciones de Explotación de Información Basada en Sistemas Inteligentes” y el “Programa de Investigación en Computación” de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones; y con las líneas del Doctorado en Ingeniería de Sistemas y Computación que funciona en el marco del convenio entre la Universidad Nacional del Nordeste y la Universidad de Málaga.

RESUMEN

La información se ha convertido, en la columna vertebral de las organizaciones, la aplicación de distintas técnicas, métodos y herramientas para garantizar mediante un proceso formal de Auditoría, la calidad y seguridad de la información es una tarea de significativa importancia.
En la actualidad no se encuentran procedimientos formales especialmente diseñados para aplicar técnicas de explotación de información en la Auditoría de Sistemas en general y a la búsqueda de datos con ruido, inconsistentes y faltantes, aplicándose en algunos casos métodologías diseñadas con otros objetivos como SEMMA o CRISP que no contemplan la especificidad de los objetivos que se persiguen, en otros casos no se aplica una metodología, esta situación provoca una disminución en la calidad del proceso de Auditoría de Sistemas.
Este proyecto busca establecer una taxonomía relacionada con la calidad de los datos, analizando los procesos de explotación de información que mejor aplican a la identificación de patrones de pistas de auditoria, se explorarán esas procesos analizando las ventajas y desventajas de cada una de ellos.

Palabras clave: *procesos de explotación de información, auditoria de sistemas, pistas de auditoria.*

1. INTRODUCCION

1.1 Auditoria de Sistemas

El actual estado de desarrollo de los sistemas de información hace que los mismos sean más complejos, integrados y relacionados. La administración efectiva de la Tecnología de la Información (TI) es un elemento crítico para la supervivencia y el éxito de las empresas, varias son las razones que producen esta criticidad, son por ejemplo, la dependencia que tienen las organizaciones de la información para su funcionamiento, el nivel de inversión que tienen en el área de TI, la potencialidad que tiene la TI para transformar las organizaciones, los riesgos y amenazas que en la actualidad tiene la información, la economía globalizada que exige un alto nivel de competitividad, entre otras. Existe una relación cada vez más fuerte entre los objetivos estratégicos de una empresa y la TI, se debe implementar un sistema adecuado de control interno que permita proteger todos los elementos relacionados con la TI, el personal, las instalaciones, la tecnología, los sistemas de aplicación y los datos.
Esto hace que sea cada vez más necesario en todas las organizaciones y no sólo en las grandes, el garantizar el cumplimiento de las normas y procedimientos establecidos para el manejo de las políticas relacionadas con la Tecnología de la Información. Existe una creciente necesidad de garantizar la seguridad y calidad de los servicios que se brindan dentro de una empresa en relación con la TI. Son muchos los riesgos que amenazan los recursos relacionados con la TI, por ejemplo, accesos indebidos a las bases de datos, falsificación de información para terceros, incumplimiento de leyes y regulaciones, fraudes, virus, destrucción de soportes documentales, acceso clandestinos a redes, entre otros.
La Auditoria de Sistemas es la evaluación sistemática de todos los aspectos relacionados con la Tecnología de la Información, uno de sus objetivos es proteger los activos que tienen las empresas y organizaciones, la información en este mundo globalizado es uno de los principales activos a resguardar. La detección de ruidos, inconsistencias o

incompletitud en los datos es fundamental en el proceso de auditoria ya que brindan pistas de posibles problemas necesarios de detectar y corregir, como por ejemplo, accesos no autorizados a las bases de datos, errores en los sistemas, etc. Utilizar métodos, técnicas y herramientas que asistan al auditor en la tarea de encontrar anomalías en las bases de datos es de suma importancia ya que hacen su trabajo más eficiente, eficaz y objetivo.

A nivel internacional existen diferentes normas que intentan estandarizar el proceso de la auditoría de sistemas, una de estos estándares es COBIT [COBIT, 2008] cuya misión es investigar, desarrollar, publicar y promover objetivos de control en tecnología de la información (TI) con autoridad, actualizados, de carácter internacional y aceptados generalmente para el uso cotidiano de gerentes de empresas y auditores. La *Information Systems Audit and Control Foundation* [COBIT, 2008] y los patrocinadores de COBIT, han diseñado este producto principalmente como una fuente de buenas prácticas para los auditores de sistemas. COBIT ha sido desarrollado como estándares para mejorar las prácticas de control y seguridad de las TI que provean un marco de referencia para la Administración, Usuarios y Auditores.

Existen distintos tipos de auditorías de sistema [Piattini 2003]: auditoría física, auditoría de la ofimática, auditoría de la dirección, auditoría de la explotación, auditoría del desarrollo, auditoría del mantenimiento, y auditoría de bases de datos, entre otras.

La norma *Statement on Auditing Standards 1009* [SAS, 2008] define a las Técnicas de Auditoria Asistida por Computadora (TAACs) como el conjunto de datos y programas que utiliza el auditor durante el desarrollo de su tarea, y explicita los más importantes pasos que el auditor de sistemas debe considerar cuando prepara la aplicación de las TAACs:

- Establecer los objetivos de auditoria de las TAACs.
- Determinar accesibilidad y disponibilidad de los sistemas de información, los programas/sistemas y datos de la organización.
- Definir los procedimientos a seguir (por ejemplo: una muestra estadística, recálculo, confirmación, entre otros).
- Definir los requerimientos de salida (*output*).
- Determinar los requerimientos de recursos.
- Documentar los costos y los beneficios esperados
- Obtener acceso a las facilidades de los sistemas de información de la organización, sus programas/ sistemas y sus datos.
- Documentar las TAACs a utilizar incluyendo los objetivos, flujogramas de alto nivel y las instrucciones a ejecutar.

1.2 Explotación de Información

Se define la Explotación de Información *(Data Mining)* [Clark, 2000] como el proceso mediante el cual se extrae conocimiento comprensible y útil que previamente era desconocido desde bases de datos, en diversos formatos, en forma automática. Es decir, la Explotación de Información plantea dos desafíos, por un lado trabajar con grandes bases de datos y por el otro aplicar técnicas que conviertan en forma automática estos datos en conocimiento.

La Explotación de Información es un elemento fundamental de un proceso más amplio que tiene como objetivo el descubrimiento de conocimiento en grandes bases de datos [Fayyad *et al*. 1996; Britos *et al*., 2005], en inglés "*Knowledge Discovery in Databases*" (KDD), este proceso, como lo muestra la figura 1, tiene una primer etapa de preparación de datos, luego el proceso de minería de datos, la obtención de patrones de comportamiento, y la evaluación e interpretación de los patrones descubiertos.

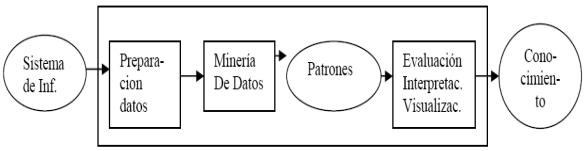

Fig.1. Proceso de KDD

1.3 Explotación de información y Auditoría de sistemas

El mayor desarrollo del uso de la Explotación de Información en actividades relacionadas con la auditoría de sistemas se relacionan con la detección de intrusos en redes de telecomunicaciones, también se encuentra en la literatura científica antecedentes relacionados con la detección de fraudes [Britos et al, 2008b], análisis de logs de auditoría, no encontrándose antecedentes de la Explotación de Información en la búsqueda de datos faltantes, con ruido e inconsistentes en bases de datos.

Ante la necesidad existente de brindar al incipiente mercado una aproximación sistemática para la implementación de proyectos de Explotación de Información, diversas empresas [Britos et al, 2008a] han especificado un proceso de modelado diseñado para guiar al usuario a través de una sucesión formal de pasos:

- SAS [2008] propone la utilización de la metodología SEMMA [SEMMA 2008] (Sample, Explore, Modify, Model, Assess).
- En el año 1999 uno grupo de empresas europeas, NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), desarrollaron una metodología de libre distribución CRISP-DM (Cross-Industry Standard Process for Data Mining) [CRISP-DM, 2008].
- La metodología P3TQ [Pyle, 2003] (Product, Place, Price, Time, Quantity), tiene dos modelos,

el Modelo de Explotación de Información y el Modelo de Negocio.

2. LINEAS DE INVESTIGACION y DESARROLLO

Existen procedimiento formales y globalmente establecidos relacionados con el uso genérico de las TAACs y para la implementación de un proceso de descubrimiento de conocimiento en grandes bases de datos, pero no existe un método para la aplicación específica de la Explotación de Información en la obtención de pistas de auditoría, otro problema es que existen trabajos relacionados con la comparación de las distintas técnicas de Explotación de Información aplicadas en general a la auditoría de sistemas pero no se encuentran antecedentes en lo relacionado al análisis de las distintas técnicas aplicadas a la búsqueda de datos faltantes, con ruido e inconsistentes. Ante esta situación aparecen dos realidades a la hora de implementar en el proceso de auditoría de sistemas la Explotación de Información, en algunos casos no se aplica ninguna metodología formal y en otros casos surge la necesidad de adaptar las metodologías existentes para implementar la Explotación de Información a la particularidad que implica utilizar esta tecnología en la Auditoría de Sistemas, siendo este proceso de adaptación empírico e informal.
Existen antecedentes de procedimientos para la implementación de un proceso de explotación de información, pero no procedimientos específicos de explotación de información para datos faltantes, con ruido e inconsistentes, en ese contexto en este proyecto se propone establecer procedimientos que identifiquen este tipo de datos. Se espera establecer una taxonomía relacionada con la calidad de los datos, analizando las técnicas de minería que mejor aplican, se explorarán esas técnicas analizando las ventajas y desventajas de cada una de ellas.

3. RESULTADOS OBTENIDOS/ESPERADOS

El proyecto presentado comenzó a fines del año 2008 y tiene prevista como tareas para su primer año:

- Refinamiento investigación documental orientada a la identificación de trabajos previos vinculados a la explotación de información aplicados al proceso de auditoria de sistemas.
- Identificación de técnicas de explotación de información aplicadas la auditoria de sistemas.
- Identificación de casos de estudio aceptados por la comunidad internacional de aplicación de explotación de información en la auditoria de sistemas para su utilización en pruebas de concepto y validación del proyecto.
- Identificación y delimitación de problemas vinculados a la detección de datos faltantes con ruido e inconsistentes en bases de datos.
- Exploración de la aplicabilidad de técnicas de exploración de información a la resolución de los problemas planteados.

4. FORMACION DE RECURSOS HUMANOS

En el marco de este proyecto se esta desarrollando una tesis doctoral y se prevé el inicio de dos tesis de grado. Esta línea vincula al Grupo de Auditoria del Departamento de Informática de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones, al Laboratorio de Sistemas Inteligentes de la Facultad de Ingeniería de la Universidad de Buenos Aires y al Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga.

5. BIBLIOGRAFIA

Britos, P.; Hossian, A.; García Martinez, R.; Sierra, E. 2005. *Minería de Datos Basada en Sistemas Inteligentes*. Nueva Librería,

Britos, P.; Dieste, O.; García Martínez, R. 2008. *Requirements Elicitation in Data Mining for Business Intelligence Projects*. En: Advances in Information Systems Research, Education and Practice. Springer,. p. 139–150.

Britos, P.; Grosser, H.; Rodríguez, D.; Garcia Martinez, R. 2008. *Detecting Unusual Changes of Users Consumption*. In Artificial Intelligence and Practice II. Springer. p. 297-306.

Clark, P.; Boswell R. 2000. *Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann Publisher.

COBIT. 2008. *Control Objectives for Information and related Technology*. http://www.isaca. org/cobit/. Vigencia 16/04/08.

CRISP-DM. 2008. http://www.crisp-dm.org/. Vigencia 15/09/08.

Fayyad U.M.; Piatetsky Shapiro G.; Smyth P. 1996. *From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, p 1-34.

Piattini, M.; Peso, E. 2003. *Auditoría Informática, un enfoque práctico*. Alfaomega-Rama,

Pyle, D. 2003. *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers,

SAS. 2008. *Statement on Auditing Standards*. http://www.aicpa.org/. Vigencia 15/09/08.

SEMMA. 2008. http://www.sas. com/technologies/ analytics/datamining/miner/semma.html. Vigencia 15/09/08.