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

H. Kuna[1], S. Caballero[1], A. Rambo[1], E. Meinl[1], A. Steinhilber[1], G. Pautch[1], R. García-Martínez[2], F. Villatoro[3]

1. Depto. de Informática, Facultad de Ciencias Exactas Quím. y Nat... Universidad Nacional de Misiones.
2. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús
3. Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga.

hdkuna@unam.edu.ar , rgarcia@unla.edu.ar

CONTEXTO

Está línea de investigación articula el "Programa de Investigación en Computación" de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones; y el "Proyecto 33A081: Sistemas de Información e Inteligencia de Negocio" del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús.

RESUMEN

La información se ha convertido en uno de los activos más importantes para las empresas y es necesario garantizar la seguridad, calidad y legalidad de dicha información. A partir de este hecho, la auditoría de los sistemas tiene un papel central en la prevención de riesgos relacionados con el gobierno de la tecnología de la información. En general, el desarrollo y la aplicación técnicas de auditoría asistidas por computadora (CAATs) es aún incipiente, en particular la minería de datos se aplica de manera embrionaria y asistemática a tareas relacionadas con la auditoría de sistemas. En la actualidad no se encuentran procedimientos formales especialmente diseñados para aplicar técnicas de explotación de información en la auditoría de sistemas y a la búsqueda de datos con ruido, inconsistentes y faltantes Este trabajo busca establecer procesos formales de explotación de información para la detección de datos anómalos en bases de datos. Esto será muy útil para la tarea de los auditores de sistemas.

Palabras clave: *procesos de explotación de información, auditoria de sistemas, pistas de auditoría, minería de datos, cluster.*

1. INTRODUCCION

1.1 AUDITORIA DE SISTEMAS

Los sistemas de información son cada vez 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 compañías, varias son las razones que producen esta alto nivel de criticidad, 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.

La auditoría de sistemas es el conjunto de técnicas, actividades y procedimientos destinados a analizar, evaluar, supervisar y recomendar sobre cuestiones relacionadas con la planificación, el seguimiento, la eficacia, seguridad y adecuación de los sistemas de información en las empresas. 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 detección de datos anómalos en las Bases de 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. Hay algunas normas ISO relacionadas con la seguridad de la información, tales como la ISO 27001/2 e ISO 17799 para complementar las buenas prácticas desarrolladas en COBIT.

Se ha desarrollado [ISACA, 2009] la directriz G3 sobre el uso de CAATs. La norma *Statement on Auditing Standards 1009* [SAS, 2008] define a las CAATs 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 debe considerar cuando prepara la aplicación de las CAATs.

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).

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.

1.4 CLUSTERING PARA LA DETECCIÓN DE OUTLIERS

Un outlier [Hawkins, 1980], es un dato que es tan diferente a otros datos que se sospecha que han sido creados por diferentes mecanismos. Históricamente la Estadística ha tenido un papel primordial en la detección de valores atípicos, en la actualidad la minería de datos en la actualidad desempeña un papel fundamental en el proceso de outliers.

El clustering es un método de aprendizaje no supervisado en el cual los datos se agrupan de acuerdo a características similares. Es una de las principales técnicas para descubrir el conocimiento oculto, siendo muy utilizados en el descubrimiento de patrones, y en particular el descubrimiento de los valores extremos. La distancia entre los objetos es el elemento más utilizados para formar los clusters, y se considera que cuanto mayor es la distancia entre un objeto y el resto de la muestra, mayor es la posibilidad de considerar objeto, como un valor atípico. Los principales métodos para medir la distancia son la distancia euclídea, la de Manhatam y la distancia de Mahalanobis.

En la actualidad, las técnicas de la agrupación principal se pueden clasificar de la siguiente manera:

- Agrupamiento jerárquico, hay una descomposi-ción jerárquica del conjunto de datos, un gráfico conocido como dendograma ser creado, lo que representa la forma en que los grupos se están creando y de la distancia entre ellos.
- Métodos basados en particiones, divisiones sucesivas del conjunto de datos se crean, los objetos se organizan en grupos de k de modo que la desviación de cada objeto de reducir al mínimo en relación con el centro de la agrupación.
- Métodos basados en la densidad, donde cada cluster se relaciona con una medida de densidad, objetos situados en regiones con baja densidad son considerados anómalos.
- Hay otros métodos como los basados en métodos difuso, basado en redes neuronales, los métodos basados en algoritmos evolutivos, los métodos basados en la entropía, etc., que están teniendo un desarrollo interesante.

2. LINEAS DE INVESTIGACION y DESARROLLO

Existen procedimiento formales y globalmente establecidos relacionados con el uso genérico de las CATTs y procedimientos para la implementación de un proceso de descubrimiento de conocimiento en grandes bases de datos, pero no existe un procedimiento para la aplicación específica de la Explotación de Información en la obtención de outliers, 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 anómalos.

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, siendo el objetivo final el desarrollo de uno o varios procedimientos que permitan aplicar de manera sistemática la Minería de datos en el proceso de descubrimiento de datos anómalos.

3. RESULTADOS OBTENIDOS/ESPERADOS

3.1. GRADO DE AVANCE

El proyecto presentado comenzó a fines del año 2008 y durante el año 2009 se desarrollaron las siguientes tareas:

- Identificación de trabajos previos vinculados a la explotación de información aplicados al proceso de auditoría de sistemas.
- Identificación de técnicas de explotación de información aplicadas la auditoria de sistemas.
- Identificación de bases de datos para realizar la experimentación.
- Identificación y delimitación de problemas vinculados a la detección de datos faltantes con ruido e inconsistentes en bases de datos.
- Determinación de las técnicas de Minería de Datos que mejor aplican al proceso de detección de outliers.
- Experimentación inicial con distintas técnicas de Minería de Datos para identificar Outliers.
- Definición inicial de algunos procedimientos para detectar Outliers

En paralelo, durante el 2009 se analizaron y utilizaron herramientas de Minería de Datos basadas en la filosofía Open Source como: RapidMiner 4.4.000, Tanagra 1.4.25, Weka 3.6.

Se determinó que el proceso de Clustering permite identificar los datos anómalos, se utilizaron métodos de agrupamiento jerárquico como HAC, métodos basados en particiones como K-Means, métodos basados en la densidad como LOF y redes neuronales del tipo SOM.

Uno de los problemas que presentan los algoritmos de clustering es que identifican la tupla que consideran que contienen outliers, pero no identifican que atributo de esa tupla particular contiene el dato anómalo, en grandes bases de datos con estructuras complejas esto puede ser una complicación en la tarea del auditor de sistemas por este motivo se orientó la investigación no solo a la determinación de la tupla que contiene outliers sino específicamente que atributo dentro de esa tupla puede considerarse anómalo.

Se determinó que no existe una única técnica que brinde resultados ideales para todas las situaciones en la detección de datos anómalos, se concluyó que la mejor solución es la combinación de distintas técnicas con el objetivo de optimizar los resultados, realizándose experimentaciones iniciales combinando K-means y HAC, SOM y HAC. Se desarrolló un procedimiento utilizando el método basado en la densidad LOF que identifica específicamente que campo puede considerarse como anómalo.

La producción científica del año 2009 relacionada con el proyecto:

- "Procedimientos de explotación de información para la detección de datos con ruido, faltantes e inconsistentes." WICC 2009. REDUNCI. San Juan.
- "Pattern Discovery in University Students Desertion Based on Data Mining". IV International Meeting on Dynamics of Social and Economic Systems. Pinamar
- "Identificación de Causales de Abandono de Estudios Universitarios. Uso de Procesos de Explotación de Información". TEyET 2009. REDUNCI. La Plata
- "Procedimientos de explotación de información".VII Jornadas de Biblioteca Digital Universitaria 2009.JBDU. Rosario
- "Auditoria de sistemas asistida por computadora". VII Jornadas Tecnológicas. Facultad de Ciencias Exactas Quim.y Nat. Universidad nacional de Misiones. Posadas.
- "Aplicaciones de la minería de datos en la detección de datos con ruido en base de datos". Primer Jornada de integración, extensión y actualización de estudiantes universitarios de informática. FCEQyN. UNaM. Apóstoles.

3.2. TRABAJOS PREVISTOS EN LA PROXIMA ETAPA

Para el año 2010 se tiene previsto:

- Analizar otras herramientas open source
- Experimentación sistemática con las técnicas de Minería de Datos para detectar outliers

- Formalizar un conjunto de procedimientos que permitan la detección de outliers en Bases de Datos.
- Comparar y clasificar los distintos procedimientos desarrollados.

4. FORMACION DE RECURSOS HUMANOS

En el marco de este proyecto se conformó un equipo de investigación dentro del "Programa de Investigación en Computación", con siete integrantes (todos ellos alumnos y egresados de la carrera de Licenciatura en Sistemas de Información de la facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones) de los cuales dos finalizaron su tesis de grado durante el año 2009, dos están en proceso de finalización de sus tesis de grado y dos están comenzando a realizar su tesis de grado, uno de los recientes egresados está por comenzar una Maestría, en el marco de este proyecto también se está desarrollando una tesis doctoral.

Esta línea de investigación vincula al Grupo de Auditoria del "Programa de Investigación en Computación" del Departamento de Informática de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones, al Grupo de Ingeniería de Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús 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.

Hawkins, D. 1980. Identification of Outliers. Chapman and Hall. London.

Information Systems Audit and Control Association. http://www. isaca.org. Vigencia 10/09/2009

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.