Análisis de Incidentes Informáticos usando Modelos de Asociación y Métodos del Análisis de Datos Multivariante.

García, Alejandro (1), Corso, Cynthia Lorena (2), Gibellini, Fabián (3), Rapallini, Marcos (4)
Laboratorio de Sistemas de Información. / Dpto. Ingeniería en Sistemas de Información / Facultad Regional Córdoba / Universidad Tecnológica Nacional
Maestro Marcelo López s/n. Ciudad Universitaria.
Teléfono: 5986011
malejandrogarcia@hotmail.com / cynthia@bbs.frc.utn.edu.ar / speaker@bbs.frc.utn.edu.ar / marco@bbs.frc.utn.edu.ar

Resumen

El objetivo de este proyecto es la caracterización y búsqueda de relaciones y asociaciones significativas entre variables relacionadas con la ocurrencia de incidentes en equipos informáticos, en el contexto de un laboratorio académico y de Investigación y Desarrollo. Para ello se implementa el uso de técnicas pertenecientes a la rama de Minería de Datos y métodos incluidos en el Análisis de Datos Multivariante.
Mediante la aplicación de técnicas y métodos apropiados, se pretende la elaboración de un modelo de conocimiento de carácter predictivo.
El mismo nos debe aportar conocimiento de las variables o factores de mayor incidencia en la presentación de incidentes, como así también el establecimiento de relaciones y modelos subyacentes en las mismas.

Palabras Clave: Minería de Datos, Análisis de Datos Multivariante, Técnicas de Asociación, Gestión de Incidentes Informáticos.

Contexto

Este proyecto surge de necesidades concretas en el contexto de un Laboratorio Informático y está relacionado con la gestión de mantenimientos y prevención de incidentes en equipos informáticos.

El mismo forma parte de un conjunto de proyectos pertenecientes al Laboratorio de Sistemas, cuyo propósito es el estudio y análisis de temáticas relacionadas con problemáticas y acciones de mejoras a implementar en este contexto.

Introducción

La gestión de las actividades relacionadas con el área de mantenimiento es clave, para que cualquier unidad de negocio pueda desarrollar sus actividades de manera eficiente. Por ello es fundamental el establecimiento de un sistema, que permita la detección de acciones de carácter preventivo con el objetivo de minimizar la presencia de fallos o incidentes.

La presentación frecuentes de fallos o incidentes, en cualquier unidad de negocio u organizacional, tienen un impacto negativo desde el punto de vista operacional, como así también económico. Las estadísticas reflejan que un alto porcentaje de las horas hombres son dedicadas a la ejecución de tareas relacionadas a la solución de fallos o incidentes y que no han sido detectados por el área de mantenimiento.

Por todo lo expuesto, se considera necesario y significativo el estudio que permita el desarrollo de una metodología que permita el descubrimiento factores influyentes que propician la presentación de incidentes o fallos en equipos, como así también detectar si existen relaciones entre los mismos. Para ello es fundamental el análisis y evaluación de datos históricos de los informes de mantenimientos correctivos reportados por el personal del Área Técnica, perteneciente al laboratorio.

Las aplicaciones de minería de datos posibilitan la identificación de tendencias y comportamientos en los datos que no son evidentes.
Para esta investigación se ha seleccionado una de las técnicas de Minería de Datos, que en función a

la problemática que se desea resolver, la más adecuada es la técnica de Asociación. Esta nos permitirá obtener un modelo de conocimiento en forma de reglas de asociación que permiten develar hechos que ocurren dentro de un conjunto de datos determinado. La selección y futura aplicación de este tipo de técnica puede resultar interesante para el descubrimiento de relaciones entre variables o atributos de un conjunto de datos.

El importante avance que ha tenido el campo de la tecnología y el abaratamiento de costos ha traído como consecuencia un aumento significativo en la cantidad de datos que son almacenados en muchas ocasiones en diferentes formatos.
La Minería de Datos es un mecanismo que nos permite facilitar precisamente, la búsqueda de información valiosa en grandes volúmenes de datos.
La implementación y construcción de un modelo de conocimiento que permita conocer el comportamiento de los incidentes en el periodo considerado, facilita la elaboración de un plan de prevención que permita la disminución de los reportes de incidentes.
Esto permite lograr mayor disponibilidad de los equipos informáticos para las diversas actividades academicas que se desarrollan en el Laboratorio de Sistemas. Desde el punto de vista económico se logra disminuir los costos relacionados con la adquisición de determinados insumos o componentes/piezas que son utilizados en las tareas de mantenimiento.

Teniendo en cuenta los resultados obtenidos, esto podría impactar principalmente en el ámbito de mantenimiento y fiabilidad de equipos informáticos en el contexto de laboratorios informáticos perteneciente a entidades públicas municipales, provinciales y nacionales.

El descubrimiento de reglas, factores y relaciones entre variables que intervienen en la presentación de incidentes en los equipos informáticos, permitirá planificar diferentes aspectos como:

- Identificación del él/los posible/s origen/es de incidentes, como por ejemplo conocer si la presentación de fallos está relacionado mayormente por inconvenientes de un determinado tipo de componente o pieza.
- Elaboración de procedimientos técnicos y tareas de mantenimiento preventivo a efectuar periódicamente.
- Capacitación al personal de Área Técnica.
- Disminución de los tiempos muertos o de parada de los equipos.
- Aprovechamiento y uniformidad en la carga de trabajo del personal de Área Técnica, debido a una planificación de actividades.
- Diseño de un proceso de compras de componentes o piezas de los equipos, que facilite la disponibilidad de un stock aceptable.

Las herramientas seleccionadas, para el procesamiento de datos y obtención de los modelos en esta investigación son Weka (Software Libre) e InfoStat (Software Comercial).

[1] Weka: esta herramienta fue concebida en la universidad de Waikato (Nueva Zelanda) implementado en lenguaje Java y que dispone de un conjunto de librerías que facilitan la integración con otras herramientas.
Además Weka contiene las herramientas necesarias para realizar transformaciones con los datos, tareas de clasificación, regresión, clustering y asociación.
La licencia de Weka es GPL, lo que significa que este programa es de libre distribución y difusión.

[2] InfoStat: es un software para análisis estadístico de aplicación general. Dispone de una amplia gama de herramientas para el tratamiento de estadísticas descriptivas y gráficos para el análisis exploratorio, como así también métodos avanzados de modelación estadística y análisis de datos multivariado.

El objetivo general de esta investigación es la implementación de herramientas relacionadas con la Minería de Datos y la Estadística; facilitando la generación de un modelo de conocimiento que permita describir y caracterizar el comportamiento de los incidentes reportados de los equipos informáticos.

Líneas de Investigación y Desarrollo

- Herramientas de Inteligencia de Negocios
- Minería de Datos
- Algoritmos de Asociación
- Mantenimiento y prevención de incidentes.

- Técnicas de asociación, aplicada al ámbito de laboratorio informático.
- Software de Minería de Datos.
- Métodos de Pre-procesamiento de datos.
- Elaboración de metodología relacionada con la implementación de proyecto de Minería de Datos, aplicada al ámbito ingenieril.
- Parámetros o métricas de calidad para el modelo de conocimiento obtenido.

Resultados y Objetivos

Entre los resultados del avance de este proyecto se pueden mencionar los siguientes:

- Se ha investigado sobre diversas herramientas estadísticas y de aprendizaje automático, con el propósito de facilitar la implementación del proyecto de minería de datos.
- Se ha investigado sobre métodos que facilitan la selección de atributos relevantes
- Se ha investigado sobre diversos algoritmos dentro de la rama de las técnicas predictivas, con el objetivo de seleccionar la más adecuada, en relación a la problemática a resolver.
- Integrante del grupo han participado en experiencias en el rol de coordinación en Panel: "Aplicación de técnicas de Minería de Datos usando software Weka", en el Congreso Argentino de Estudiantes de Ingeniería Industrial y carreras afines (CAEII 2009).
- Se ha participado en Congresos Nacionales (JAIOO 2012/CAIM y WICC (2011))

La diagramación de actividades planificadas para este proyecto son las que se detallan a continuación:

- Búsqueda de alternativas respecto a herramientas estadísticas y de aprendizaje automático para facilitar la implementación del proyecto de minería de datos.
- Identificación de criterios para la selección de la herramienta que se adapte mejor a la problemática que se desea resolver.
- Análisis de los datos o reportes existentes relacionado con la gestión de incidentes.
- Selección de muestras de datos considerados para este estudio.
- Evaluación de los atributos o variables que serán consideradas para el estudio de gestión de incidentes de los equipos informáticos.
- Carga y migración de los datos en un formato adecuado para ser interpretada por la herramienta de minería de datos seleccionada.
- Selección y Evaluación de algoritmos de aprendizaje supervisado para la implementación de técnica de asociación.
- Implementación y validación.

Dentro de los objetivos específicos.

a) Determinar dentro de la Técnica de Asociación, la variedad de algoritmos existentes y evaluar cuál de ellos se adapta mejor a la situación problemática planteada. Considerando criterios o indicadores de confiabilidad que determinen el nivel de calidad de las reglas de asociación resultantes.

b) Selección y Evaluación de herramientas para la implementación de proyectos de Minería de Datos que dispongan una variedad significativa de algoritmos pertenecientes a las Técnicas de Asociación.

c) Difundir el uso herramientas, como una alternativa de instrumento de inteligencia de negocios para la ejecución de proyectos de Minería de Datos, a través de charlas, conferencias y publicaciones en el ámbito universitario.

c) Generación de un modelo de conocimiento (Reglas de asociación) que nos permita la detección de los factores que tienen alta incidencia en el reporte de incidentes de los equipos informáticos pertenecientes al laboratorio.

d) Transferir al seno de la cátedra Inteligencia Artificial (5to. Nivel de la carrera Ingeniería en Sistemas de la U.T.N. - F.R.C.) los resultados obtenidos, con la finalidad de mejorar el diseño curricular de las asignaturas y enriquecer las mismas con el intercambio interdiscilplinario entre investigación, aplicación y análisis de la evidencias resultantes.

f) Evaluar la posibilidad de extender este estudio a otros laboratorios informáticos con esta problemática, principalmente en el contexto académico para los distintos niveles educativos

Los resultados obtenidos:

- Definición de la arquitectura tecnológica en torno al proyecto de Minería de Datos, que generalmente suele tener una arquitectura cliente-servidor.

[3] Para lograr una aplicación óptima de estas técnicas avanzadas, las mismas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios.
Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining.

Este warehouse puede ser implementado en una variedad de sistemas de bases relacionales y debe ser optimizado para un acceso a los datos flexible y rápido.

- Selección de variables: Aún después de haber sido preprocesados, en la mayoría de los casos se tiene una cantidad significativa de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

Los métodos para la selección de características son básicamente dos:
• Aquellos basados en la elección de los mejores atributos del problema,
• Y aquellos que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

Se ha investigado que la herramienta Weka dispone de un conjunto de algoritmos que facilitan esta tarea. En general estos algoritmos pueden ser clasificados por diversos criterios.
[4] Una categorización popular es aquella en la que los algoritmos se distinguen por su forma de evaluar atributos y se clasifican en: Filtros, donde se seleccionan y evalúan los atributos en forma independiente del algoritmo de aprendizaje y Wrappers (envoltorios), los cuales usan el desempeño de algún clasificador (algoritmo de aprendizaje) para determinar lo deseable de un subconjunto. Los algoritmos evaluadores en un subconjunto de atributos disponibles en Weka testeados son:
CfsSubsetEval, ConsistencySubsetEval (Filtros), ClassifierSubsetEval,WrapperSubsetEval (Wrappers)
Se complementará esta etapa con las pruebas de los algoritmos que permiten la evaluación de atributos individuales como:
ChiSquaredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval y OneRAttributeEval.

- Implementación de herramienta para la carga de datos.

- Algoritmos de Extracción de Conocimiento:

Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables.
También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

[5] La Minería de reglas de asociación es una técnica importante en la Minería de Datos y consiste en encontrar relaciones de implicación entre los valores de los atributos de los objetos de un conjunto de datos.

Actualmente en esta fase del proyecto se esta investigando acerca de las caracteristicas y ventajas de los algoritmos de asociación que dispone la herramienta Weka son: A Priori, Filtro Asociado, HotSpot, A priori-Predictivo y el Tertius. Una linea futura de investigación es el análisis de algoritmos de asociación implementados en otras herramientas de aprendizaje automático.

Formación de Recursos Humanos

Los integrantes de este proyecto, en su gran mayoría están conformados por docentes pertenecientes al plantel académico de la carrera de Ingeniería en Sistemas de Información.

Los mismos poseen formación académica de post-grado como especialización y magister en curso.

Este trabajo de investigación pretende integrar y articular contenidos relacionados con el campo de la Minería de Datos en un contexto del campo de la Ingeniería, con los docentes de la asignatura electiva "Inteligencia de Negocios" perteneciente al quinto nivel de la carrera de Ingeniería de Sistemas de Información.

Una de las integrantes está elaborando el plan de tesis, para la carrera de Magister en Sistemas de Información, relacionada con línea de investigación de este proyecto, complementando la investigación con técnicas avanzadas pertenecientes al análisis de datos multivariante.

Todos los integrantes docentes del PID han participado del proceso de categorizaciones en investigación dentro del Programa de Incentivos del MECyT; así como en la categorización interna que posee la U.T.N.

Bibliografía

[1] "Manual de Weka", Diego García Morate
http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf
[2] **http://www.infostat.com.ar/**
[3] "Minería de Datos", Vallejos Sofía, (2006)
http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf
[4] "Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional", Roxana Ramos, Rosa Ramos Palmero, Ricardo Avalos, María Matilde García Lorenzo,(2007)
http://bvs.sld.cu/revistas/gme/pub/vol.9.%281%29_01/p1.html
[5] "Minería de Reglas de Asociación sobre datos Mezclados", Ansel Rodríguez González, José Martínez Trinidad, Jesús Ochoa, José Ruiz Shulcloper, (2009)
http://ccc.inaoep.mx/portalfiles/file/CCC-09-001.pdf
[6] Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software Weka, Basilio Serra Araujo, Pearson Educación, Madrid 2006.
[7] "Introducción a la Minería de Datos", José Hernández Orallo, Ramírez Quintana, Mª José, Ferri Ramírez, César, Pearson Prentice Hall, 2005.
[8] "Minería de Datos. Técnicas y Herramientas", Pérez López, César, González Daniel Santin, Thompson. Madrid. (2007)
[9] "Extracción Automática de Conocimiento en Base de Datos e Ingeniería del Software"**,** Quintana Ramírez María José, Orallo José Hernández, España (2003)
[10] "Data Mining: Concepts and Techniques"**,** Jiawei Han & Micheline Kamber, Vipin Kumar, Addison-Wesley (2006)