Tratamiento de incidentes informáticos mediante la utilización de redes bayesianas

Carlos A. Talay, **José L. Saenz**, **Dora S. Maglione**, **Griselda Rojas**, **Carlos Amarilla**

Unidad Académica Río Gallegos

Universidad Nacional de la Patagonia Austral

Santa Cruz (9400), Argentina

carlostalay@yahoo.com.ar, jsaenz_lacaze@hotmail.com, dmaglione@disytel.com, patagoniaustral@gmail.com, carlos.amarilla@gmail.com

Luis Marrone

L.I.N.T.I. – Universidad Nacional de La Plata

Calle 50 y 115 – 1er. Piso – Edificio Bosque Oeste

lmarrone@info.unlp.edu.ar

Resumen. La evaluación de los incidentes informáticos es un primer paso para realizar un correcto diagnóstico de las causales que producen una merma en la eficiencia en una organización. A posteriori, con este diagnóstico, podemos delinear una estrategia adecuada para intentar reducir a su mínima expresión la ocurrencia de estos incidentes. Este trabajo propone una metodología de análisis de incidentes basada en redes bayesianas, mediante la cual se pueda identificar los factores de riesgo que pueden tener una mayor incidencia en el normal desenvolvimiento de las tereas desarrolladas en un centro de cómputo.

Palabras clave: Incidentes informáticos, redes bayesiana, prevención de incidentes

1. Introducción

El relevamiento y evaluación de incidentes informáticos dentro de un centro de cómputos, conjuntamente con la optimización de los procedimientos de manejo seguro de información, es el camino que posibilita la mejora de funcionamiento de una organización ya que permite evitar la pérdida de tiempo e información ante un incidente. La meta de toda organización es realizar el trabajo de la manera más simple y eficiente posible, con el menor riesgo. Si bien este criterio está aceptado, la implementación de una política rigurosa en donde exista un sistema organizado de recopilación de incidentes, su posterior análisis y la toma de acciones proactivas que eviten que estos incidentes, no es una práctica de uso muy común entre organizaciones en donde los centros de cómputos tienen un rol importante en la gestión de la información. En general se tiene el caso de organizaciones que realizan la recopilación de incidentes, pero estos sólo son consultados ante la ocurrencia de un problema, utilizando la información recopilada para implementar las tareas correctivas ante hechos consumados. Por ello, si queremos contar con una herramienta o procedimiento que nos permita actuar en forma proactiva, debemos procesar esa información de manera que no sólo nos permita determinar la frecuencia de los incidentes sino que también proporcione una medida de su potencial peligrosidad.

Cuando iniciamos el análisis del problema, nos encontramos en forma temprana con el primer inconveniente, ya que al momento de ordenar los incidentes observamos que no existe un consenso generalizado en la manera de clasificarlos. Si bien este problema ha sido abordado desde hace tiempo [1] [2] y desde varios enfoques [3] [4] [5], no hay un criterio

unificado sobre una taxonomía de incidentes, dependiendo la determinación de su clasificación al tratamiento que se les vaya dar. En consecuencia, en este trabajo se propondrá una clasificación de incidentes, luego en referencia a ella trataremos dos casos de estudio sobre los que se han recopilado y clasificado los datos.

A posteriori se delineará un modelo de red bayesiana que permitirá realizar un análisis multivariado de los datos que refleje la interrelación existente entre la distribución de los incidentes, de esta manera podremos analizar el efecto que estos causan sobre el desempeño de la organización.

2. Las Redes Bayesianas

Las redes bayesianas, también conocidas como redes de creencias, constituyen una herramienta estadística que codifica relaciones probabilísticas entre un grupo de variables aleatorias de interés [6] [7]. En forma muy simplificada podemos explicarlo como:

Sea $\{A_1, A_2,\ldots, A_i,\ldots, A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, tales que la probabilidad de cada uno de ellos es distinta de cero (0). Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} \tag{1}$$

Donde:
$P(A_i)$ son las probabilidades a priori.
$P(B|A_i)$ es la probabilidad de B en la hipótesis A_i.
$P(A_i|B)$ son las probabilidades a posteriori.

Así mismo, si queremos hacer extensible la relación de B con todos los sucesos A_k, obtenemos la ecuación 2, también conocida como la Regla de Bayes:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k)P(A_k)} \tag{2}$$

Esta herramienta estadística es utilizada en los más variados campos de la ciencia [8] [9] [10] [11] [12], ha mostrando ser una valiosa ayuda en el diseño de estrategias para la toma de decisiones y la elaboración de sistemas expertos. Cuando se utiliza en conjunto con técnicas estadísticas y datos provenientes de conocimientos previos [13], el modelo-gráfico resultante proporciona varias ventajas al momento de realizar un análisis de las relaciones entre las causas y los efectos de un proceso aleatorio. Una de estas ventajas es que, como el modelo caracteriza dependencias entre todas las variables relacionadas mediante el vínculo de causa-efecto en las fases secuenciales del proceso, se puede obtener en forma dinámica la manera en que se vinculan las variables que afectan un evento, lo que posibilita realizar inferencias sobre el comportamiento de todo el sistema. Otro aspecto interesante es que una red bayesiana puede ser la base de un sistema experto que aprenda por sí mismo a modelar el conocimiento de las relaciones causales, y por lo tanto, se puede utilizar para obtener una mejor comprensión sobre el dominio de un problema, prediciendo las consecuencias que generan sobre todo un sistema la evolución de las variables que lo afectan. También podemos ver que un modelo bayesiano aplicado a un proceso aleatorio permite revelar la relación entre las causales no previstas inicialmente y los efectos de los eventos que se registran. Alimentando

este modelo con los eventos que se registran en forma sucesiva, a lo largo de un período de tiempo, podemos caracterizar un sistema y de esta manera comprender su dinámica.

Bajo estas consideraciones vemos que la semántica probabilística bayesiana, es una representación ideal para combinar el conocimiento previo (que a menudo viene dado en forma causal) y los datos disponibles, pudiéndose de esta manera llegar a la modelización de una serie de eventos.

3. Metodología

El presente trabajo se desarrolla a lo largo de pasos fundamentales que son la recopilación de los incidentes que dan origen a nuestra base de datos. Proponer una clasificación de estos incidentes, a fin de encuadrarlos dentro de una taxonomía que nos ayudará a generar el grafo bayesiano, imprescindible para la modelización de la interrelación de las variables asociadas con el evento principal. Por último realizaremos la evaluación de estos datos con el objeto de inferir características del comportamiento del sistema en estudio.

El diseño de la red bayesiana debe contemplar toda la gama de incidentes que pueden identificarse en un centro de cómputos. A tal fin hemos definido una red bayesiana causal que responde a un esquema de un grafo orientado acíclico G (Figura 1), cuyos nodos están compuestos por las 10 variables que conforman el evento principal "Incidentes", todas ellas con distribución binomial probabilística $B_k \sim B(n, p_k)$. Estos parámetros binomiales y su interdependencia se detallan a continuación:

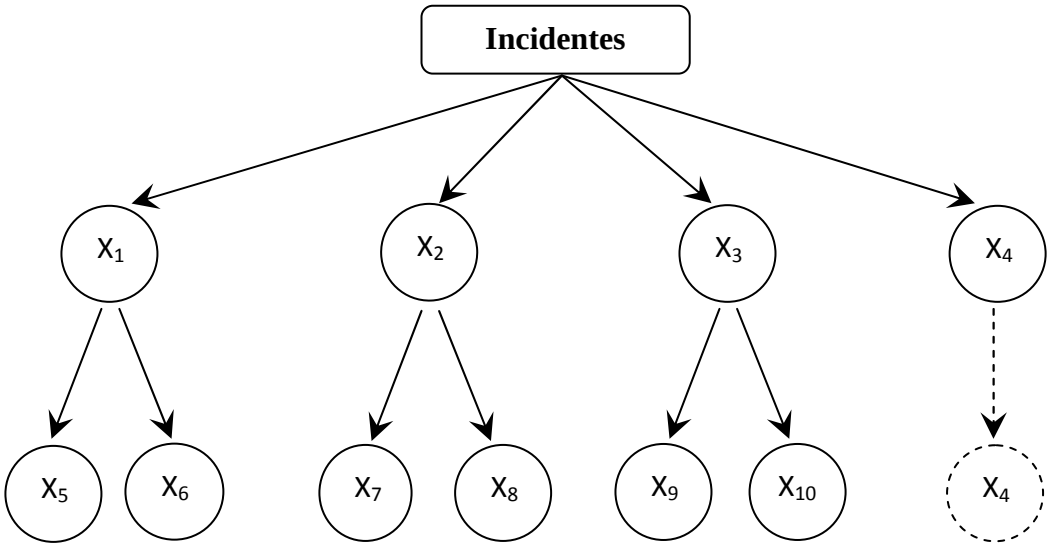

Figura 1. Gráfico orientado de la red acíclica bayesiana propuesto

Como puede verse, hay dos niveles en la relación interna de la red. En el primer nivel hay tres nodos padres X_1, X_2 y X_3 y un nodo terminal X_4. Estos nodos padres tienen a su vez dos nodos hijos cada uno, definidos por los pares (X_5, X_6), (X_7, X_8) y (X_9, X_{10}), todos ellos terminales, los que conforman un segundo nivel del grafo. Los nodos de este grafo definen los tipos de incidentes informáticos registrados. Para ello se ha adoptado una óptica operativa, es decir evaluar los incidentes que deben afrontar los encargados de soporte técnico (hardware y software) de una organización. En un primer paso plantearemos una lista de variables binomiales aleatorias, pertenecientes a la red Bayesiana [14], y considerando que cada una de estas variables puede tomar sólo valores de (0,1) como variables dicotómicas:

X_1: Hardware (1.1)
X_2: Software (1.2)
X_3: Comunicaciones (1.3)
X_4: Energia Elétrica (1.4)
X_5: CPU´s (1.1.1)
X_6: Periféricos (1.1.2)
X_7: Configuración de usuarios (1.2.1)
X_8: Configuración de/sobre aplicaciones (1.2.2)
X_9: Configuración de redes (1.3.1)
X_{10}: Conectividad física y/o de dispositivos de red (1.3.2)

El listado precedente comienza enumerando las X_i posibles causales de incidentes, que corresponden a cada uno de los nodos. Luego se indica el tipo de incidente de acuerdo al criterio de clasificación considerado y por último, entre paréntesis, la referencia a la cual se aludirá al momento de realizar las gráficas de distribución de incidentes. Téngase en cuenta que para los gráficos sólo serán considerados los nodos terminales.

Como hemos especificado, los datos tienen como referencia absoluta un pseudo nodo inicial, descripto en el grafo como el evento principal llamado “Incidente”, valorado en un primer momento con n = 0. De esta forma, cada nuevo incidente registrado aumenta a n+1 a lo largo de todos los nodos de la red. Ante la incorporación de un nuevo valor, este aporta a uno y sólo uno de los últimos nodos "hijos" de la gráfica. De esta manera, mientras el volumen de datos va creciendo, la red incorporará “conocimiento” referente a la dinámica del régimen de los incidentes que se registran y así el sistema irá adquiriendo las características de un sistema experto [15].

El otro elemento fundamental de la red bayesiana que aquí consideramos, es el conjunto de las p_k probabilidades condicionales asociadas a los nodos conectados del I-mapa, como se ve en la figura 1, que completa el modelo de dependencia que proponemos. Comenzamos asignando valores con las frecuencias empíricas a los p_k valores de los parámetros, que van a cambiar con cualquier nuevo incidente registrado, con el propósito de poner en funcionamiento la estructura de aprendizaje probabilística de la red bayesiana. Luego, con el propósito de definir la estructura probabilística de la red bayesiana, que sólo describe los átomos (0-1, como ya hemos visto) o los valores posibles adoptados por las variables aleatorias sin sus cargas, pesos o probabilidades de transición, de acuerdo a la siguiente distribución:

Primer nivel

Variable aleatoria (Xk)
X1 ; P { X1 = 0 } , P { X1 = 1 }
X2 ; P { X2 = 0 } , P { X2 = 1 }
X3 ; P { X3 = 0 } , P { X3 = 1 }
X_4 ; P { X_4 = 0 } , P { X_4 = 1 }

Segundo nivel

Variable aleatoria (X_k)
X_5 ; P { X_5 = 0 / X_1 = 0 } , P { X_5 = 0 / X_1 = 1 } , P { X_5 = 1 / X_1 = 0 } , P { X_5 = 1 / X_1 = 1 }
X_6 ; P { X_6 = 0 / X_1 = 0 } , P { X_6 = 0 / X_1 = 1 } , P { X_6 = 1 / X_1 = 0 } , P { X_6 = 1 / X_1 = 1 }
X_7 ; P { X_7 = 0 / X_2 = 0 } , P { X_7 = 0 / X_2 = 1 } , P { X_7 = 1 / X_2 = 0 } , P { X_7 = 1 / X_2 = 1 }

X_8 ; P { X_8 = 0 / X_2 = 0 } , P { X_8 = 0 / X_2 = 1 } , P { X_8 = 1 / X_2 = 0 } , P { X_8 = 1 / X_2 = 1 }
X_9 ; P { X_9 = 0 / X_3 = 0 } , P { X_9 = 0 / X_3 = 1 } , P { X_9 = 1 / X_3 = 0 } , P { X_9 = 1 / X_3 = 1 }
X_{10} ; P { X_{10} = 0 / X_3 = 0 } , P { X_{10} = 0 / X_3 = 1 } , P { X_{10} = 1 / X_3 = 0 } , P { X_{10} = 1 / X_3 = 1 }

Para el desarrollo de este trabajo se han tomado dos casos de estudio representados por clientes reales que denominaremos caso A y caso B. A continuación, con el objeto de entender qué tipo de organización representan, para cada una de ellas realizaremos una mínima reseña de sus principales características. Luego seguiremos con los resultados obtenidos. Para ambos casos se ha solicitado a las organizaciones que informen sobre los incidentes ocurridos completando una base de datos que contempla: la fecha y hora del reporte del incidente por parte de un usuario, la fecha y hora del cierre del incidente por parte del personal técnico/operativo encargado de dar respuesta. Con estos datos podemos saber el tiempo que se demoró en solucionar este incidente. Así mismo existe un tercer parámetro que es una estimación del tiempo que ese incidente proyectó a terceros usuarios imposibilitándolos de poder desarrollar tareas en su puesto de trabajo, es decir una medida de cómo ese incidente repercutió sobre el normal desenvolvimiento de otros usuarios de esa misma organización. Este último dato se ha calculado en forma estimativa.

Caso A

Se trata de un organismo estatal, que posee una red de tipo MAN. El control de la red se realiza desde un sector que aúna el área técnica y la administración operativa. Allí se reciben los pedidos de asistencia por parte de los usuarios y luego son derivados, ya sea al soporte de hardware o software, para ser atendidos. Esta organización está conformada por un total de 27 dependencias que poseen en total 445 terminales, en donde básicamente se desarrollan tareas de carga de datos y consultas con sistemas propios desarrollados en la organización.

De acuerdo a los datos proporcionados por el cliente A y alimentando el modelo bayesiano propuesto, tomando un período que abarca unos 9 meses de recopilación de información, tenemos la siguiente representación de la distribución dicotómica de la cantidad de incidentes producidos sobre las distintas categorías de nodos terminales definidas en el modelo.

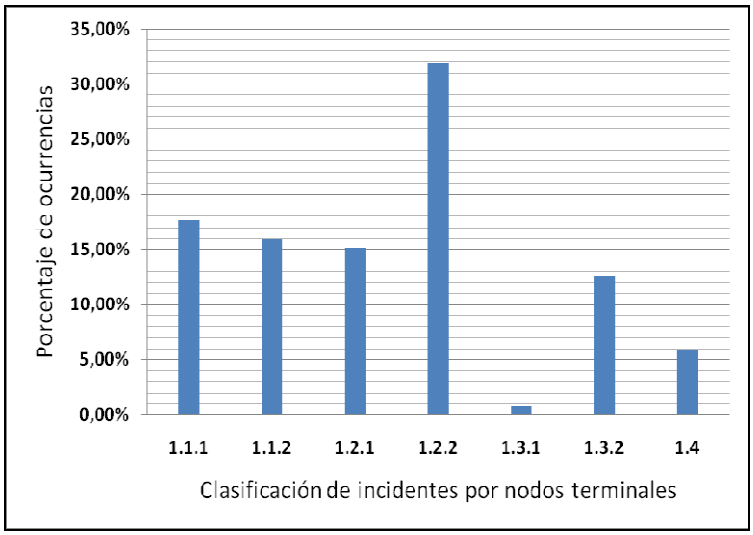

Gráfico 1. Distribución de los porcentajes de incidentes registrados (sobre el total de ocurrencias)

Por otro lado, si mantenemos la misma red bayesiana pero abandonamos el modelo dicotómico y representamos, en el mismo período de tiempo (9 meses) y con la misma base de datos, una distribución que pretende mostrar el tiempo (en horas) que insumió solucionar los inconvenientes que generó ese incidente, más el tiempo acumulado (en horas) que ese

mismo incidente proyectó sobre otros usuarios, imposibilitándolos de realizar su trabajo o parte de él, tendremos una distribución como la que observamos a continuación

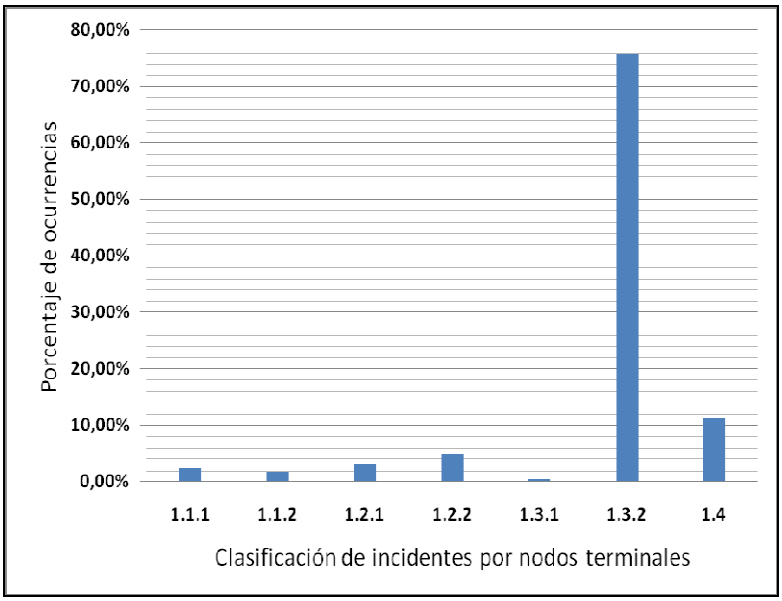

Gráfico 2. Porcentaje registrado del tiempo (en horas) que afecta/n usuario/s

Caso B

Se trata de un organismo privado que posee una red LAN en su casa central con enlaces a sucursales distribuidas a lo largo del territorio nacional, mayormente centralizadas en la provincia de Santa Cruz. Toda la organización trabaja sobre una base de datos única de clientes. En total poseen aproximadamente 300 terminales con distintos tipos de usuarios que desempeñan tareas de carga de datos, consultas y operaciones sobre servicios solicitados por los clientes. Aunque la mayor parte de los servicios brindados son de consulta y modificación de datos específicos del cliente. Estos servicios, en general, son independientes unos de otros y pueden ser brindados por un mismo usuario de la organización a un cliente dado, en el esquema de una atención personalizada. Como en el caso anterior se posee un sector de asistencia técnica que abarca el soporte técnico de hardware y software en donde se brinda la solución a los problemas y se realiza el mantenimiento de sistemas.

Nuevamente, considerando el modelo bayesiano propuesto y tomando un período que abarca unos 9 meses de recopilación de información, vemos a continuación la representación de la cantidad de incidentes producidos sobre las distintas categorías definidas en el modelo.

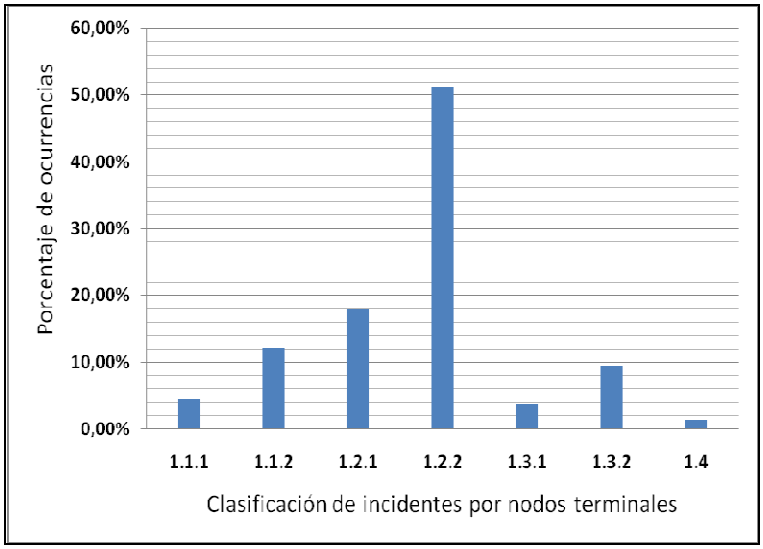

Gráfico 3. Distribución de los porcentajes de incidentes registrados (sobre el total de ocurrencias)

Al igual que en el caso anterior, abandonamos la clasificación dicotómica y representamos, en el mismo período de tiempo (9 meses), y con la misma base de datos, una distribución que pretende mostrar el tiempo (en horas) que insumió solucionar los inconvenientes generado por ese incidente, más el tiempo acumulado (en horas) que ese mismo incidente proyectó sobre otros usuarios, imposibilitándolos de realizar su trabajo o parte de él. Así obtenemos el siguiente gráfico

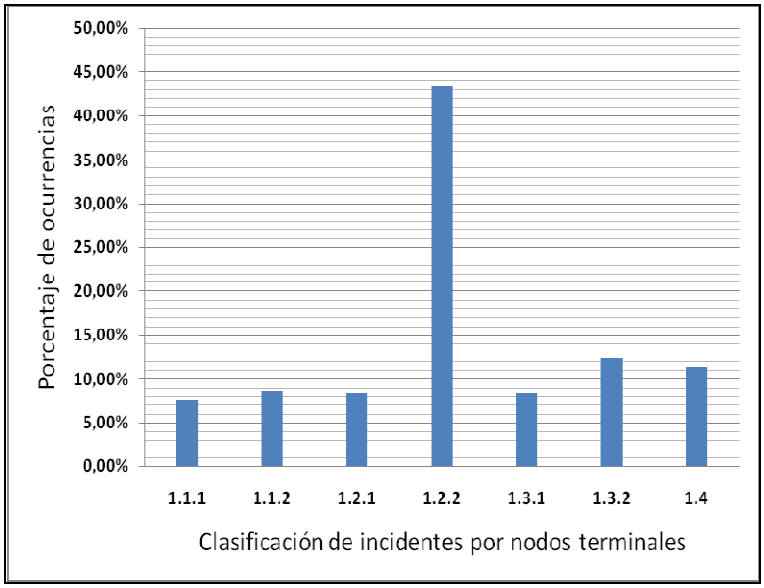

Gráfico 4. Porcentaje registrado del tiempo (en horas) que afecta/n usuario/s

4. Conclusiones

De acuerdo a los resultados del caso A, si realizamos una comparación fundamentalmente entre las columnas 1.2.2 y 1.3.2 de los gráficos 1 y 2, y menor medida entre las columnas 1.1.1, 1.1.2, 1.2.1, vemos que la distribución de las mismas se modifica en forma radical y por tanto cambia en forma notable la perspectiva del análisis. Para esta organización podemos inferir que es más perturbador a nivel operativo, tener incidentes que afecten a una cierta cantidad de usuarios, que un incidente de bajo impacto pero que se manifiestan en forma repetitiva. Si bien todo análisis es relativo, los datos parecen indicar que los incidentes que generen imposibilidad de operar en forma masiva, o a un grupo considerable de usuarios, generan un perjuicio mayor que el generado por una serie de incidentes individuales que tengan un tiempo acumulativo total de afectación distribuido a lo largo de un cierto período de tiempo. Este efecto se podría pensar de la siguiente manera, los incidentes de tipo focalizado, aunque tengan cierta periodicidad, permiten continuar con sus tareas normales al resto de los usuarios que no han sido afectados, mientras que los incidentes con afectación masiva de usuarios o sobre grupo de usuarios, generan una pérdida mayor en operatividad a la organización.

En el caso B, el análisis de incidentes no es tan concluyente. En la comparación de los gráficos 3 y 4 vemos que ambos registran una distribución similar en los valores de las columnas que representan los distintos tipos de incidentes. Aquí la columna 1.4 es la que presenta mayor diferencia en sus valores comparativos, aunque por su poco peso relativo no es un factor determinante en un análisis general.

Teniendo en cuenta los casos de estudio podemos concluir que el análisis de incidentes registrados en una organización debe ser abordado no sólo bajo la óptica de un análisis cuantitativo (cantidad de incidentes registrados) sino que también debe analizarse su

incidencia en su aspecto cualitativo (grado en que el incidente afecta a un usuario o grupo de usuarios en el desarrollo de su/s tareas), ya que los resultados pueden variar considerablemente según cuál de los dos enfoques tomemos y el tipo de organización que tenga la empresa. Así, el correcto análisis de los incidentes, permitirá realizar una mejor elaboración de los procedimientos y optimizar la utilización de los recursos, a la hora de intentar reducir su impacto en una organización determinada.

5. Reconocimientos

Queremos agradecer los alumnos Hernán Hernández e Iván Pezzuti, alumnos de la Carrera de Licenciatura de Sistemas de la UNPA-UARG, que colaboraron en la construcción de las bases de datos que son fundamentales para la realización de este proyecto. Así mismo también queremos agradecer a las entidades que desinteresadamente aportaron los datos necesarios e imprescindibles para alimentar los modelos matemáticos.

Referencias

[1] Aslam, Taimur. *A taxonomy of security faults in the unix operating system*. Master's thesis, Purdue University. 1995

[2] Aslam, Taimur; Krsul, Ivan; and Spafford, Eugene H. *Use of A Taxonomy of Security Faults.* Computer Science Technical Reports. Paper 1305. Purdue University. 1996

[3] Carl Landwher et al. *A taxonomy of computer program security flaws*. Technical report, Naval Research Laboratory. 1993

[4] Sandeep Kumar. *Classification and Detection of Computer Intrusions*. PhD thesis, Purdue University. 1995

[5] Ortiz Bayona, Z. y Galindo Pulido, F. *Hacia una Taxonomía de Incidentes de Seguridad en Internet*. Revista de Ingeniería - Universidad Distrital Francisco José de Caldas. 2005

[6] Jensen, F. *An Introduction to Bayesian Networks*. Springer. 1996

[7] Langley, P., W. Iba, & K. Thompson. *An analysis of Bayesian classifiers*. In Proceedings, Tenth National Conference on Artificial Intelligence (pp. 223–228). Menlo Park, CA: AAAI Press. 1992

[8] Boys, R. J., D. J. Wilkinson and T. B. L. Kirkwood. *Bayesian inference for a discretely observed stochastic kinetic model*. Statistics and Computing, 18(2), 125-135. 2008

[9] Neil, M., Fenton, N., Forey, S. & Harris, R. *Using Bayesian belief networks to predict the reliability of military vehicles*. Computing & Control Engineering Journal, Feb. 2001, pp 11-20. 2001

[10] David Heckerman. *Bayesian Networks for Data Mining*, Journal of knowledge Discovery and Data Mining 1(1), pag. 79-119 Kluwer Acedemic Publishers. 1997

[11] Ferat Sahin, John S. Bay. *Structural Bayesian network learning in a biological decision theoretic intelligent agent and its application to a herding problem in the context of distributed multi agent system*, 2001 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, pages: 1606 – 1611. 2001

[12] De la Fuente, E. I., García, J., y De la Fuente, L. *Estadística bayesiana en la investigación psicológica*. Metodología de las Ciencias del Comportamiento, 4, 185-200. 2002

[13] D. Heckerman, D. Geiger, and D.M. Chickering. *Learning Bayesian Networks: The Combination of Knowledge and Statistical Data*. Machine Learning, vol. 20, pp. 197-243. 1995

[14] Enrique Castillo Ron, José María Gutiérrez and Ali S. Hadi. *Expert systems and probabilistic network models*. Springer – Verlag. 1997

[15] Kevin B. Korb and Ann E. Nicholson. *Bayesian artificial intelligence*. Computer Science and Data Analysis. CRC / Chapman Hall, Boca Raton. 2004