7ª Jornada sobre la Biblioteca Digital Universitaria
JBDU2009
"La biblioteca universitaria en la web"

"Procedimientos de la explotación de información aplicados al ámbito bibliotecológico"

Kuna, Horacio; Miranda, Mirta J.; Caballero, Sergio; Jaroszczuk, Susana.
Departamento de Bibliotecología. Facultad de Humanidades y Ciencias Sociales. Universidad Nacional de Misiones

Resumen

Se denomina Explotación de Información al proceso que permite transformar la información en conocimiento utilizando para ello técnicas de Minería de Datos. Se define a la Minería de Datos como el conjunto de pasos que permite extraer conocimiento previamente desconocido que sea comprensible y útil en grandes bases de datos. Este proceso posibilita la construcción de modelos, la predicción, el descubrimiento de grupos, la identificación de factores, detección de perfiles entre otras aplicaciones. Para esto se utilizan distintas técnicas como las redes neuronales, las redes Bayesianas, los algoritmos de inducción, etc. Se han expandido con mucha fuerza las aplicaciones relacionadas con la Minería de Datos a la Word Wide Web, estas se relacionan con analizar tanto los datos de los servidores Web como de los clientes y permite descubrir conocimientos relacionados el uso, la estructura, los accesos a la Web.

El término Bibliomining se utiliza para definir la utilización de la Minería de Datos en el ámbito bibliotecológico, sus principales aplicaciones son la personalización de los servicios, el apoyo a la toma de decisiones, análisis de las colecciones, análisis del comportamiento de los usuarios.

En este trabajo se presenta una experimentación realizada con información de una Biblioteca de la Universidad Nacional de Misiones donde se aplicó un algoritmo TDIDT (Top Down Induction of Decision Trees) con el objetivo obtener patrones de comportamiento relacionados con el cumplimiento de las fechas de devolución de los libros prestados.

Se pudo demostrar la utilidad que tiene la aplicación de Procedimientos de Explotación de Información y la enorme potencialidad que tienen los mismos en el ámbito Bibliotecológico.

Palabras clave: procesos de explotación de información, minería de datos, webmining, bibliomining

1 Explotación de Información

La explotación de Información es la sub-disciplina Informática que aporta a la Inteligencia de Negocio las herramientas (procesos y tecnologías) para la transformación de información en conocimiento, para lograr este objetivo se utiliza a la Minería de Datos.

Se define la Minería de Datos (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, de manera automática. Es decir, la Minería de Datos 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 minería de datos 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.

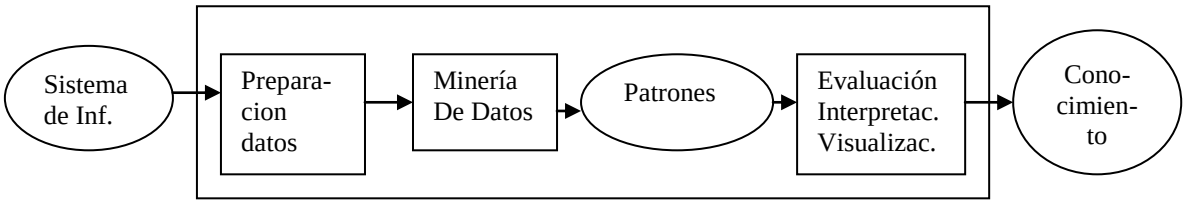

Figura 1. Proceso de KDD

Ante la necesidad existente de brindar al incipiente mercado una aproximación sistemática para la implementación de proyectos de Minería de Datos, 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 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, 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.

1.1 Procesos de Explotación de Información

En [Britos, 2005] se identificaron cinco procesos de minería de datos y el contexto en el cual deben ser aplicados: predicción, construcción de modelos, descubrimiento de grupos, identificación de factores y detección de perfiles. Complementariamente puede explorarse la optimización de los resultados obtenidos mediante el uso de algoritmos genéticos.

La abstracción que describe cada proceso y las técnicas asociadas se resumen a continuación.

- ✓ *Proceso de Predicción*: Cuando se quiere saber el valor que tomarán algunas variables del negocio variables dependientes en función del valor que tomarán otras variables independientes. Técnica a Utilizar: Redes Neuronales.
- ✓ *Proceso de Construcción de Modelos:* Cuando se quiere saber como la variación de una o más variables del negocio incidirá sobre la variación de las otras variables. Técnica a Utilizar: Redes Bayesianas.
- ✓ *Proceso de Descubrimiento de Grupos*: Cuando se requiere identificar clases en el conjunto de registros de información que se tienen del negocio. Técnica a Utilizar: Mapas Auto Organizados (Kohonen).
- ✓ *Proceso de Identificación de Factores:* Cuando se requiere identificar cuales son los factores que inciden sobre determinado resultado del negocio. Técnica a Utilizar: Arboles de Inducción (TDIDT).

- ✓ *Proceso de Detección de Perfiles:* Cuando se requiere identificar los factores de clases en el conjunto de registros de información que se tienen del negocio. Técnica a Utilizar: Mapas Auto organizados combinado con Árboles de inducción.

La tabla 1 muestra la correspondencia entre los procesos de explotación de información, las tecnologías y sus aplicaciones.

PROCESOS EI	**TECNOLOGIAS**	**Aplicaciones**
Predicción	Redes Neuronales de Back Propagation (Perceptron multicapa)	Predicción de valores de atributos
Agrupamiento	Redes Neuronales SOM (mapas autoorganizados de Kohonen)	Descubrimiento de grupos
Inducción	Algoritmos TDIDT (Top Down Induction of Decision Trees)	Descubrimiento de reglas de comportamiento
Ponderación	Redes Bayesianas	Ponderación de interdependencia de Atributos
Agrupamiento + Inducción	SOM + TDIDT	Descubrimiento de reglas de pertenencia a grupos
Inducción + Ponderación	TDIDT + Redes Bayesianas	Ponderación de atributos relevantes en reglas de comportamiento
Agrupamiento + Ponderación	SOM + Redes Bayesianas	Ponderación de atributos relevantes en cada grupo descubierto

Tabla 1. Procesos, tecnologías aplicaciones de DM

1.1.1 Redes neuronales [Ferrero et al., 2006]

Las Redes Neuronales Artificiales (RNA) son sistemas de procesamiento de la información que se basan en el comportamiento de las redes neuronales biológicas. Se trata de un conjunto de elementos elementales de procesamiento llamados neuronas que se conectan entre sí y dichas conexiones tienen un valor numérico modificable llamado peso sináptico.

El proceso que una neurona artificial tiene es simple, se trata de sumar los valores de las entradas (inputs) que recibe de otras unidades conectadas a ella, comparar esta cantidad con el valor umbral y, si lo iguala o supera, enviar una señal de activación o salida (output) a las unidades a las que esté conectada.

Las redes de Neuronas artificiales son modelos computacionales paralelos que constan de unidades de procesos adaptativas y masivamente interconectadas entre sí.

La figura 1 muestra la estructura de una neurona artificial elemental.

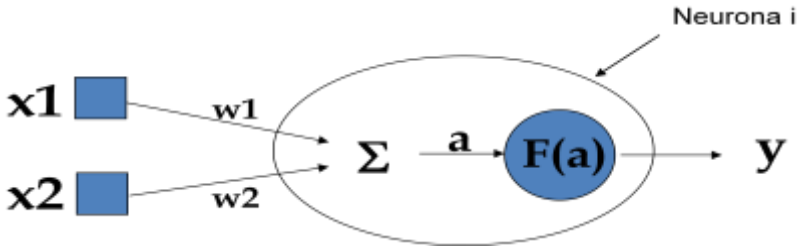

Figura 1. Neuronal artificial elemental

El objetivo es lograr una maquina formada por la interconexión de muchos elementos simples de calculo cuyo objetivo es aprender de la experiencia y generalizar. El Modelo de Neurona Artificial de McCulloch y Pitts, calcula la suma ponderada de las entradas que recibe de otras neuronas y produce un 0 o 1 dependiendo si la suma supera un determinado umbral.

En resumen si la suma de las entradas Xj que recibe de la neurona i, ponderadas por los pesos wij supera un umbral, la neurona se activa.

Algunas de sus características fundamentales son:

- ✓ Aprendizaje adaptativo.
- ✓ Autoorganización
- ✓ Tolerancia a fallos.
- ✓ Operación en tiempo real.
- ✓ No linealidad

- ✓ Implementación VLSI (Very Large Scale Integrated)

Las principales aplicaciones de las redes neuronales son:

- ✓ Asociación y clasificación
- ✓ Predicción.
- ✓ Regeneración de patrones
- ✓ Optimización.

1.1.2 Redes Bayesianas [Felgaer et al, 2006]

El uso del método para aprendizaje estructural de redes bayesianas que se basa en el algoritmo desarrollado por [Chow y Liu, 1968] para aproximar una distribución de probabilidad por un producto de probabilidades de segundo orden, lo que corresponde a un árbol. La probabilidad conjunta de n variables se puede representar como:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i) P(X_i | X_{j(i)})$$

donde $X_{j(i)}$ es la causa o padre de. X_i

Se plantea el problema como de optimización y lo que se desea es obtener la estructura en forma de árbol que más se aproxime a la distribución “real”. Para ello se utiliza una medida de la diferencia de información entre la distribución real (P) y la aproximada (P*):

$$I(P, P^*) = \sum_{x} P(X) \log(P(X)/P^*(X))$$

Entonces, el objetivo es minimizar I. Para ello se define una diferencia en función de la información mutua entre pares de variables, que se define como:

$$I(X_i, X_j) = \sum_x P(X_i, X_j) \log(P(X_i, X_j)/P(X_i)P(X_j))$$

Chow y Liu demuestran que la diferencia de información es una función del negativo de la suma de las informaciones mutuas (pesos) de todos los pares de variables que constituyen el árbol. Por lo que encontrar el árbol más próximo equivale a encontrar el árbol con mayor peso. Basado en lo anterior, el algoritmo para determinar la red bayesiana óptima a partir de datos es el siguiente:

- Calcular la información mutua entre todos los pares de variables (hay exactamente n (n - 1) / 2 pares).
- Ordenar las informaciones mutuas de mayor a menor.
- Seleccionar la rama de mayor valor como árbol inicial.
- Agregar la siguiente rama mientras no forme ciclo, si es así, desechar.
- Repetir (4) hasta que se cubran todas las variables (n - 1 ramas).
- [Rebane y Peral, 1988] extendieron el algoritmo de Chow y Liu para poliárboles.

En el caso de un poliárbol, la probabilidad conjunta es:

$$P(X) = \prod_{i=1}^{n} P(X_i \left| X_{j1(i)}, X_{j2(i)}, ..., X_{jm(i)} \right.)$$

donde $\{X_{j1(i)}, X_{j2(i)}, ..., X_{jm(i)}\}$ es el conjunto de padres de la variable X_i.

1.1.3 Árboles de inducción [Quinlan, 1986, 1996, 1999]

La aplicación de los distintos métodos para construirse reglas de decisión directamente a partir de los datos, analizaremos el método conocido como “separa y reinarás”. El enfoque básico de dicho método consiste en tomar cada una de las clases y buscar la manera de cubrir todas las instancias que pertenecen a ella, excluyendo las instancias que no pertenecen. La idea es tomar una clase y construir una regla que cubra todas las instancias que pertenecen a ella. La regla se va construyendo condición por condición. La condición que se agrega a la regla es

aquella que maximice p/t, donde t es la cantidad de instancias cubiertas por la regla, p es la cantidad de instancias cubiertas por la regla de clase i, con lo que t-p será la cantidad de instancias cubiertas por la regla pertenecientes a una clase distinta de i.

1.1.4 Algoritmos Genéticos [Goldberg, 1989]

Se utilizarán los mecanismos de reproducción, mutación y selección de redes individuales. Para aplicar los conceptos genéticos de evolución es necesario primero encontrar la manera de codificar la información contenida en una NN. Estos códigos reciben el nombre de cromosomas. Para una dada topología de red, la información total está contenida en los ω_{ij}, y el los umbrales (thresholds) V_i. El vector (ω_{ij}, V_i) puede ser considerado entonces un cromosoma, aunque es posible encontrar codificaciones más sofisticadas de esta información. Para aplicar el concepto de selección, se necesita además calificar a cada cromosoma, de acuerdo con su performance, de acuerdo con un problema dado. A diferencia de la Redes Neuronales, los Algoritmos Genéticos requieren de todo una ensamble de elecciones iniciales de sus parámetros, que se denomina población de cromosomas.

El proceso es como sigue. Se crea la generación 0 formada totalmente al azar. Para ello se eligen aleatoriamente N elementos de la población original que pasaran a ser los cromosomas de dicha generación. Un vez seleccionados. Una de las codificaciones más comunes es la binaria. Hasta aquí hemos obtenido la primera generación. Seguidamente se aplican reiteradamente los operadores genéticos que harán que esta generación evoluciones para encontrar el óptimo. Estos operadores son los siguientes:

a. Selección (o Reproducción): Mediante este operador se obtiene la nueva generación formada por una selección de cromosomas de la generación previa. Estos cromosomas contribuirán en una o más veces a la formación de la nueva generación,

según una probabilidad dada por la función de fitness o aptitud que es la función a maximizar por el Algoritmo Genético.

b. Cruce: Una vez creada la nueva generación, se elige un par de cromosomas en forma aleatoria. Estos cromosomas son partidos en una posición p elegida al azar y se intercambian los primeros p genes de un cromosoma con los primeros p genes del otro, con una probabilidad pc. Suponiendo que se esta usando codificación binaria y que cada cromosoma esta compuesto por 8 genes (bits) se da a continuación un ejemplo de cruce con p=3.

c. Mutación: La mutación es un operador secundario que cumple la función de evitar la pérdida de información que puede darse por los otros dos operadores descriptos.

El algoritmo genético aplicará reiteradamente los operadores 1 a 3 hasta que:

- haya transcurrido un tiempo determinado;
- se hayan cumplido una cantidad dada de iteraciones;
- se decida interrumpir el proceso.

2 Minería de datos en entornos WEB

La aplicación de técnicas de Data Mining sobre el conjunto de datos contenidos en la World Wide Web se conoce con el nombre de WebMining [Hernández 2004] [Etziony, 1996], el objetivo es aprovechar todas las ventajas de los procesos de Minería de Datos para obtener conocimiento de la información disponible en Internet.

El volumen de los datos que se disponen en la Web, las posibilidades de negocios, la necesidad de mejorar los servicios que se brindan por Internet, entre otros motivos han potenciado la investigación en esta área.

Existen dos enfoques bien diferenciados de análisis, por un lado la Minería de datos desde el lado del servidor y por el otro desde el lado del cliente.

Se utiliza la minería de datos en entornos Web para descubrir en forma automática documentos y servicios de la web y extraer información útil sobre ellos, información

que implica distintos tipos de datos: imágenes, sonido, texto, semi-estructurado, imágenes, etc.,

Se aplican técnicas de Minería de Datos para:

- ✓ Descubrir conocimiento relacionado con el contenido de la Web donde se localizan los datos de las páginas HTML, los datos multimedia, datos XML y de textos.
- ✓ Descubrir conocimientos relacionados con el uso y el acceso a la Web (Web User Mining).
- ✓ Descubrir conocimientos relacionados con la estructura de la Web y se relaciona con encontrar patrones de comportamiento en los enlaces o links que se encuentran en los documentos hipertextuales en Internet.

3 Aplicaciones de la Explotación de Información al ámbito bibliotecológico

La aplicación de técnicas de Minería de Datos en el ámbito bibliotecario se conoce con el nombre de bibliomining [Nicholson, 2003]. La llegada de las nuevas tecnologías de la Información y las comunicaciones a las Bibliotecas ha potenciado la búsqueda de patrones de comportamiento en los datos que se manejan.

Algunas de sus principales aplicaciones son:

- ✓ Apoyo a la toma de decisiones
- ✓ Análisis de los datos disponibles de la colección con el objetivo de contar con información que ayude a administrar los fondos de la biblioteca, en este caso las redes neuronales han mostrado muy interesantes resultados.
- ✓ Análisis del comportamiento de los usuarios.
- ✓ Personalización de los servicios.

Para Nicholson [Nicholson, 2003], el proceso de Minería de Datos aplicado al ámbito bibliotecológico tiene seis fases para su implementación:

- ✓ Determinación de las áreas de interés.
- ✓ Identificación de fuentes de datos internas y externas.

- ✓ Recopilar, limpiar y hacer anónimos los datos en el data warehouse.
- ✓ Selección de las herramientas de análisis apropiadas.
- ✓ Descubrimiento de patrones a través de la minería de datos y creación de informes con herramientas tradicionales de análisis.
- ✓ Análisis e implementación de los resultados.

3. Experimentación

3.1 Diseño experimental y variables

El objetivo del trabajo fue tratar de entender la causa por la cual un usuario se retrasa en la devolución de libros, de un sistema de gestión bibliotecaria que funciona en un entorno Web de una Facultad de la Universidad Nacional de Misiones, donde se realiza la reserva a través de internet.

Se trató de obtener patrones automáticos de comportamiento de la base de datos del sistema de gestión bibliotecaria Koha con información de los años 2006 al 2009 mediante el uso de procesos de explotación de información estandarizados [Britos, 2008].

Se realizó un preprocesamiento con el objetivo de mejorar la calidad de los datos y se detectaron algunos problemas relacionados con datos faltantes, se agregaron algunas variables y se completó en forma aleatoria el contenido de las mismas con el objetivo de optimizar el proceso de explotación de información.

Algunas variables fueron descartadas ya que no brindaban información sustantiva al objetivo planteado y fueron creadas nuevas variables a partir de variables ya existentes.

Las principales variables utilizadas se muestran en la Tabla 1.

Nombre de la variable	Tipo de variable	Descripción	Valores posibles
claustro	Dependiente	Claustro al que pertenece el socio de la biblioteca	1 alumno 2 docente 3 no docente
signatura de clase	Dependiente	Clasificación temática del libro	<=13 >3
Cod_carrera	Dependiente	Código de carrera	<=4

			>1
reserva_c	Dependiente	Si se realizó o no reserva del libro prestado	0 = si 1 = 2
Semestre	Dependiente	Semestre en el cual se realizó el prestamo	1 =primer semestre 2 =segundo semestre
cumplimiento	Independiente	Informa si el socio devolvió en termino o no el libro prestado	si = Verdadero no = Falso

3.2 Resultados

El principal objetivo fue encontrar características de los prestamos donde se produce un atrasa en la devolución del libro. La variable objetivo planteada fue cumplimiento, definiéndose como cumplimiento al socio de la biblioteca que devuelve el libros prestado en la fecha prevista.

Para llegar al objetivo propuesto se aplicó un algoritmo de inducción que permitió obtener un conjunto de reglas que posibilitan explicar porque los socios se retrasan en la devolución de un libro.

Se realizo la experimentación utilizando la herramienta software TANAGRA (Open Source) en su versión 1.4.25 y se utilizó el algoritmo C4.5 (figura 2)

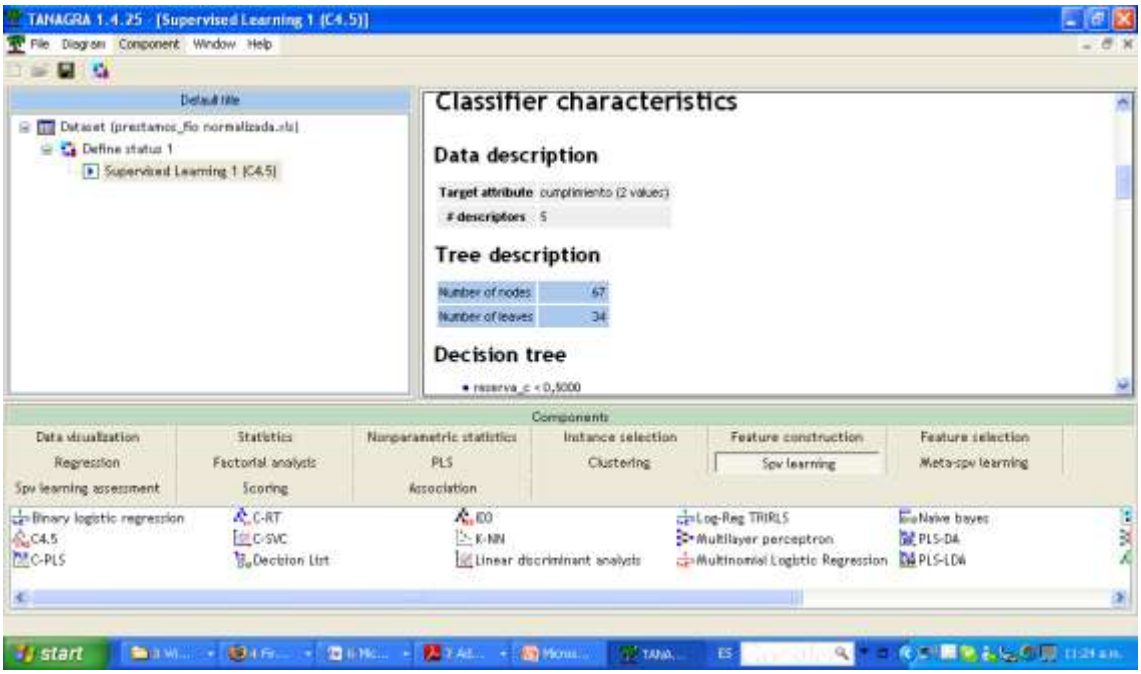

Figura 2. Experimentación con Tanagra

Se encontraron 48 reglas de comportamiento de la base de datos. A continuación se muestran algunas de las reglas encontradas.

- reserva_c < 0,5000
    - signatura de clase < 6,5000
        - claustro < 2,5000 then cumplimiento = **si** (72,78 % of 36344 examples)
        - claustro >= 2,5000
            - cod_carrera_c < 1,5000
                - signatura de clase < 5,5000 then cumplimiento = **si** (57,45 % of 188 examples)
                - signatura de clase >= 5,5000 then cumplimiento = **no** (57,81 % of 64 examples)
            - cod_carrera_c >= 1,5000
                - signatura de clase < 5,5000
                    - cod_carrera_c < 2,5000 then cumplimiento = **si** (80,88 % of 136 examples)
                    - cod_carrera_c >= 2,5000
                        - signatura de clase < 4,5000
                            - cod_carrera_c < 3,5000 then cumplimiento = **si** (81,36 % of 59 examples)
                            - cod_carrera_c >= 3,5000
                                - Semestre_c < 1,5000 then cumplimiento = **no** (66,67 % of 6 examples)
                                - Semestre_c >= 1,5000 then cumplimiento = **si** (77,78 % of 9 examples)
                        - signatura de clase >= 4,5000 then cumplimiento = **si** (65,96 % of 47 examples)
                - signatura de clase >= 5,5000 then cumplimiento = **si** (80,33 % of 122 examples)

Una de las reglas aparece por ejemplo que cuando se realizó reserva previa, la signatura de clase es 3/4/5/6, el claustro es alumno o docente, el cumplimiento en la devolución del libro es de alrededor del 73%.

- que cuando se realizó reserva previa, la signatura de clase es 3/4/5, el claustro es no docente, el código de carrera es = 1, el cumplimiento en la devolución del libro es de alrededor del 57%.
- que cuando se realizó reserva previa, la signatura de clase es 6, el claustro es no docente, el código de carrera es = 1, el no cumplimiento en la devolución del libro es de alrededor del 57%.
- que cuando se realizó reserva previa, la signatura de clase es 3/4/5, el claustro es no docente, el código de carrera es = 2, el cumplimiento en la devolución del libro es de alrededor del 81%.

3.3 Interpretación

El conocimiento que surge en la base del sistema de gestión de la biblioteca aporta un conocimiento fundamental para entender cual es la lógica de funcionamiento del sistema de reserva / préstamo y devolución de libros. Este conocimiento que no es visible sin la aplicación de procedimientos de explotación de información es de suma utilidad ya que permite entender en que casos se producen atrasos en la devolución de libros y poder de esta manera tomar las medidas preventivas que permitan corregir esta situación.

Es posible aplicar otros algoritmos de Minería de datos por ejemplo redes SOM para clusterizar y de esta manera analizar cual es el agrupamiento que surge o Redes Bayesianas para entender la interrelación entre atributos.

4. Conclusiones y futuras líneas de Investigación

Analizando los resultados obtenidos después del proceso de explotación de la información aplicando un algoritmo de inducción, es posible afirmar que estas herramientas resultan de gran importancia para determinar las causales del cumplimiento o no de las fechas de devolución de libros en un sistema de gestión de

bibliotecas que funciona en un entorno WEB, dando elementos para el análisis y la toma de decisiones como por ejemplo adoptar una política de capacitación de usuarios ante prestamos de determinada signatura topográfica, o dirigir la capacitación a los alumnos de determinada carrera, etc. Es importante destacar que la confiabilidad de los resultados del proceso de explotación de información tiene directa relación con la calidad de los datos de los sistemas de gestión.

Como consecuencia de estas conclusiones surgen una serie de preguntas con relación a los datos que se recogen de cada préstamo: ¿son los necesarios? ¿Son pocos? ¿Son bien interpretados? ¿Son excesivos? ¿Están bien categorizados? ¿Se necesita incorporar datos nuevos? ¿Se debe realizar un control de calidad más exhaustivo de los datos que están en la base de datos?

Si bien es muy importante la potencialidad que tiene para los Bibliotecarios el uso de procedimientos de Explotación de información en general y en entornos Web en particular, su aplicación es aún aislada e insipiente, siendo un importante campo de investigación.

Bibliografía

BRITOS, P. Objetivos de Negocio y Procesos de Minería de Datos Basados en Sistemas Inteligentes. Reportes Técnicos en Ingeniería del Software. 2005. 7(1): p. 26-29.

BRITOS, P.; HOSSIAN, A.; GARCÍA MARTINEZ, R.; SIERRA, E. Minería de Datos Basada en Sistemas Inteligentes. Nueva Librería, 2005.

BRITOS, P.; DIESTE, O.; GARCÍA MARTÍNEZ, R. Requirements Elicitation in Data Mining for Business Intelligence Projects. En: Advances in Information Systems Research, Education and Practice. Boston: Springer : IFIP International Federation for Information, 2008. p. 139–150.

BRITOS, P.; GROSSER, H.; RODRÍGUEZ, D.; GARCIA MARTINEZ, R. Detecting Unusual Changes of Users Consumption. In Artificial Intelligence and Practice II. Boston: Springer : IFIP International Federation for Information Processing, 2008. p. 297-306.

CHOW, C.; LIU, C. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory, 1968. p. 462-467.

CLARK, P.; BOSWELL R. Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann Publisher, 2000.

CRISP-DM. [En línea]. Disponible en: http://www.crisp-dm.org/. [Consultado el 15/09/09].

ETZIONY, Oren. The World-Wide Web: Quagmire or Gold Mine. Communications of

the ACM. Vol.39, No.11, November 1996 .pp. 65-68

FAYYAD U.M.; PIATETSKY SHAPIRO G.; SMYTH P. From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. p 1-34.

FELGAER, P.; BRITOS, P.; GARCÍA MARTÍNEZ, R. Prediction in Health Domain Using Bayesian Network Optimization Based on Induction Learning Techniques. International Journal of Modern Physics C, 2006. 17(3): p. 447-455.

FERRERO, G.; BRITOS, P.; GARCÍA MARTÍNEZ, R. Detection of Breast Lesions in Medical Digital Imaging Using Neural Networks. Boston: Springer : IFIP International Federation for Information Processing, 2006. Vol. 218, p. 1-10.

HERNANDEZ, José et al. Introducción a la minería de datos. Madrid : Pearsons,

2004. 656 p.

GOLDBERG, D. Genetic Algorithms in Search, Optimization, and Machine Learning. Adisson Wesley Publishing Company, 1989.

NICHOLSON, Scott. and STANTON, J. [En Línea] 2003. Gaining strategic dvantage

through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. In Nemati, H. & Barko, C. (Eds.). Organizational data mining: Leveraging enterprise data resources for optimal performance. Hershey, PA: Idea Group Publishing. 247-262. Updated on 2/16/04 . Disponible en: <http://www.bibliomining.com/nicholson/odmcom.html > [Consultado: 02/09/2009]

PYLE, D. Business Modeling and Business intelligence. Morgan Kaufmann Publishers, 2003.

QUINLAN, J. Induction of decision trees. Machine Learning, 1986.1(1): p. 81-106.

QUINLAN, J. Learning Decision Tree Classifiers. ACM Computing Surveys, 1996. 28(1): p. 71-72.

QUINLAN, J.R. Simplifying decision trees. International Journal of Man-Machine Studies, 1999. 51(2): p. 497-510.

REBANE, G.; PEARL, J. The recovery of causal poly-trees from statistical data. International Journal of Approximate Reasoning.1988. 2(3): p. 341.

SEMMA. [En línea]. disponible en http://www.sas.com/technologies/ analytics/datamining/miner/semma.html. [Consultado el 15/09/09].