EL MANEJO DE DATOS

Aproximación desde los estudios de la información

Georgina Araceli Torres Vargas

Z666.73
L56M3

El manejo de datos. Aproximación desde los estudios de la información / Coordinadora Georgina Araceli Torres Vargas. - México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2019.

viii, 116 pp. - Colección: TECNOLOGÍAS DE LA INFORMACIÓN.

ISBN: 978-607-30-2690-1

1. Datos vinculados. 2. Minería de datos. 3. Investigación bibliotecológica.
I. Torres Vargas, Georgina Araceli, coordinadora. II. Ser.

Diseño de portada: Natalia Cristel Gómez Cabral

Primera edición, 2020

Ciudad Universitaria, 04510, Ciudad de México
Impreso y hecho en México
ISBN: 978-607-30-2690-1

Publicación dictaminada
2020

Contenido

Minería de Datos, el caso de estudio de la Biblioteca Dr. Valentín Gómez Farías de la Facultad de Medicina de la UNAM.

David Flores Macías
Guadalupe Vanesa Carolina Gutiérrez Hernández
Universidad Nacional Autónoma de México

Introducción

La Minería de Datos es el proceso automatizado para la extracción de patrones de un cierto conjunto de datos. Aunque es éste un paso del Proceso de Descubrimiento de Conocimiento, normalmente se le conoce como Minería de Datos. También se puede definir como el hecho de descubrir información implícita pero útil de datos almacenados.

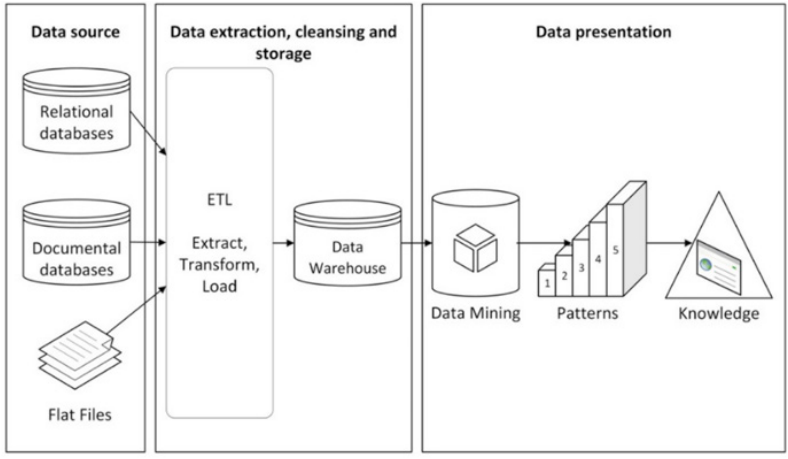
Las técnicas de minado son aplicadas en un amplio rango de dominios; por mencionar algunos ejemplos, si se genera minería de datos con datos obtenidos de la web, se conoce como *web mining,* si es usado en texto es minería de textos y si se aplica a Bibliotecas es llamado *Bibliomining* o Bibliominería. Este último término es muy interesante debido a que si uno realiza la búsqueda en inglés en la web de bibliotecas y minería de datos, normalmente los resultados proporcionan información técnica sobre las librerías utilizadas por los algoritmos de máquina. Por ello

(Nicholson 2006), se introdujo el término de Bibliomining, justamente para hacer referencia a la aplicación de la minería de datos en Bibliotecas. Siendo más específicos, en el presente trabajo la Bibliominería es usada para encontrar patrones y tendencias de los sistemas transaccionales en bibliotecas, entendiéndose como transaccionales todas aquellas operaciones que se realizan en una base de datos al realizar movimientos de circulación tales como préstamos, devoluciones y resellos (Prakash *et al.* 2004).

Desarrollo

El proceso de Minería de Datos utilizado en este estudio se presenta a continuación (Sigüenza-Guzmán 2015):

Diapositiva 1



1. Origen de los datos. Tomando en cuenta la estructura de la base de datos de circulación bibliográfica, se identificaron aquellos campos de la misma que podrían ser útiles para el estudio, y que

también fueran candidatos para poderse categorizar y construir la vista minable. Se determinó que éstos fueran la carrera del alumno, el material bibliográfico la clasificación, la fecha de préstamo, la fecha de devolución (indicada por sistema), la fecha de retorno (fecha real en la que se realizó la devolución) y la hora del préstamo.

2. Extracción de los datos, limpieza y almacenamiento.
Creación de una vista minable (Gutiérrez, Barranco y Méndez 2008). Para obtener dichos datos, se ejecutó una consulta SQL en el Sistema Manejador de Bases de Datos Oracle. El periodo fue del 1-08-2015 al 31-10-2018, dicha consulta proporcionó un total de 133 776 registros.

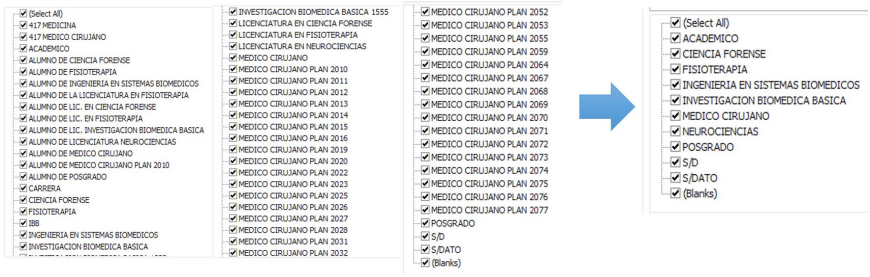
Diapositiva 2

CLASIFICACION	FECHA PR	FECHA DE	FECHA RE	HORA PRESTAMO	CARRERA
RE461 K3518 2016	20180611	20181113	0	1827	MEDICO CIRUJANO PLAN 2010
RE461 K3518 2016	20180611	20181113	0	1826	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20181107	0	1539	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20181106	0	1516	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180724	20181115	0	1725	MEDICO CIRUJANO
RM300 B3618 2016	20180813	20181107	0	906	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20181106	0	854	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180813	20180827	0	851	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180522	20180605	0	1507	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180821	20181115	0	859	MEDICO CIRUJANO PLAN 2010
RM300 B3618 2016	20180824	20181017	0	1319	MEDICO CIRUJANO PLAN 2010
RC731 C355 2017	20180724	20181115	0	1725	MEDICO CIRUJANO
RC111 M35 2016	20180829	20181108	0	1411	MEDICO CIRUJANO
RC76 M6818 2015	20180828	20181105	0	1544	MEDICO CIRUJANO
RC76 M6818 2015	20180820	20181029	0	1415	MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180629	20180803	0	1111	MEDICO CIRUJANO
RC76 M6818 2015	20180821	20181015	0	1446	MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180730	20181106	0	1433	ALUMNO DE MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180825	20181112	0	935	MEDICO CIRUJANO PLAN 2010
RC76 M6818 2015	20180823	20181026	0	1705	MEDICO CIRUJANO PLAN 2010
RA972 G65 2016	20180615	20181113	0	1522	MEDICO CIRUJANO PLAN 2010
R899 H4718 2016	20180628	20181110	0	1722	MEDICO CIRUJANO PLAN 2010
R899 H4718 2016	20180822	20181113	0	1355	MEDICO CIRUJANO
R899 H4718 2016	20180725	20181114	0	1508	ALUMNO DE MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181107	0	1258	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	851	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	1201	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	1140	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20180827	0	1020	MEDICO CIRUJANO PLAN 2010
QM601 M65418 2016	20180813	20181106	0	943	MEDICO CIRUJANO PLAN 2010

A continuación se muestra cómo se limpiaron o categorizaron cada uno de los campos seleccionados.
Campo: Carrera del alumno. Limpieza de los datos.

Como se puede ver en la figura, el campo de carrera no se encontraba normalizado, y existían diversas formas de nombrar una misma carrera. En este caso, utilizando el *software* de aplicación Excel, a través de fórmulas, se realizó la categorización, la cual quedó de la siguiente forma.

Diapositiva 3



Campo: Clasificación. Categorización de los datos.
A través de funciones de Excel, y de acuerdo con cada una de las clasificaciones de los registros, se recuperó el nombre de la clase o materia, de acuerdo con la clasificación LC (Library of Congress).

Diapositiva 4

CLASIFICACION	Clasificacion LC	Nombre
RE461 K3518 2016	RE	Ophthalmology
RE461 K3518 2016	RE	Ophthalmology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RM300 B3618 2016	RM	Therapeutics, Pharmacology
RC731 C355 2017	RC	Internal Medicine
RC111 M35 2016	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RC76 M6818 2015	RC	Internal Medicine
RA972 G65 2016	RA	Public Aspects of Medicine
R899 H4718 2016	R	Medicine (General)
R899 H4718 2016	R	Medicine (General)
R899 H4718 2016	R	Medicine (General)
QM601 M65418 2016	QM	Human Anatomy

Campo: Estatus de préstamo. Categorización de los datos.
Se categorizó de la siguiente forma:

PT. Libros que se encuentran prestados en tiempo
PV. Libros prestados que no han sido devueltos.
DT. Devoluciones realizadas en tiempo.
DV. Devoluciones realizadas después de la fecha de devolución indicada en el sistema.

Campo: Hora de préstamo. Categorización de los datos.
Si la hora se encuentra dentro del rango de 8:00 a 15:00, se estableció como TM (Turno matutino).
Si la hora se encuentra dentro del rango de 15:01 a 20:00, se estableció como TV (Turno vespertino).
Con todos estos campos limpios y categorizados, fue posible obtener la vista minable, de la cual se muestra a continuación un extracto.

Diapositiva 5

Clasificacion	Nombre	Estatus	Carrera	HORA
RE	Ophthalmology	PT	MEDICO CIRUJANO	TV
RE	Ophthalmology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PV	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PV	MEDICO CIRUJANO	TV
RM	Therapeutics, Pharmacology	PT	MEDICO CIRUJANO	TM
RM	Therapeutics, Pharmacology	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TV
RC	Internal Medicine	PT	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TV
RC	Internal Medicine	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PV	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TM
RC	Internal Medicine	PT	MEDICO CIRUJANO	TM
RC	Internal Medicine	PV	MEDICO CIRUJANO	TV

Esta vista minable se exportó de Excel a un archivo delimitado por comas, el cual fue el archivo de entrada para el *software* de aplicación que se encargó de realizar el minado de datos.

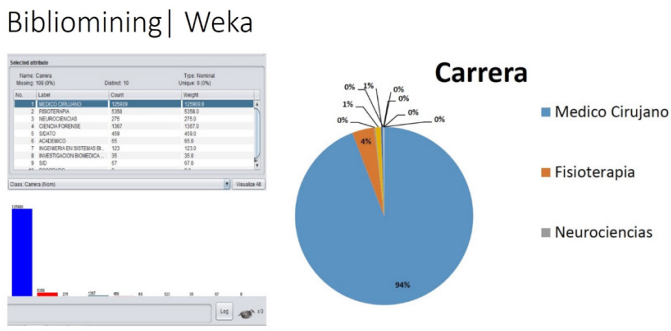
3. Minería de datos para generar conocimiento y presentación de los datos.
Con la vista minable ya generada, se decidió realizar las tareas de *Clustering* (Agrupación) y Clasificación, con el fin de encontrar patrones no triviales.

3.A *Clustering*
Objetivo: se identificaron grupos de registros que son similares entre ellos, pero diferentes del resto de los datos.

Software utilizado: Weka (Weka 3) es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación de datos, su clasificación, regresión, agrupación, extracción de reglas de asociación y visualización. Es un *software* de código abierto emitido bajo la Licencia Pública General de GNU.

Weka proporciona un primer vistazo estadístico de los datos contenidos en la vista minable.

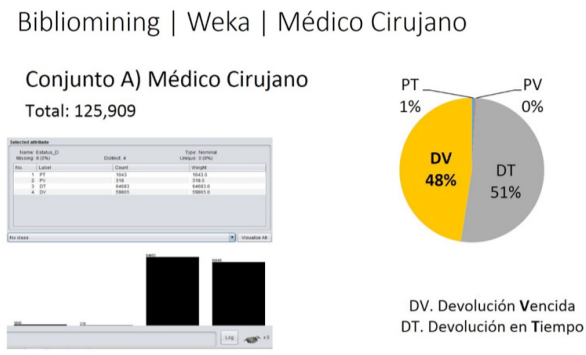
Diapositiva 6



Esta imagen nos indica que el 94% de los datos pertenecen a la carrera de Médico Cirujano, por lo que se decidió dividir el conjunto en dos apartados, lo que quedó de la siguiente forma:

Conjunto A.1) Médico cirujano (125 909 registros)
Conjunto B.1) Otras carreras y clasificaciones.
Comenzado con el Conjunto A.1) Médico cirujano.

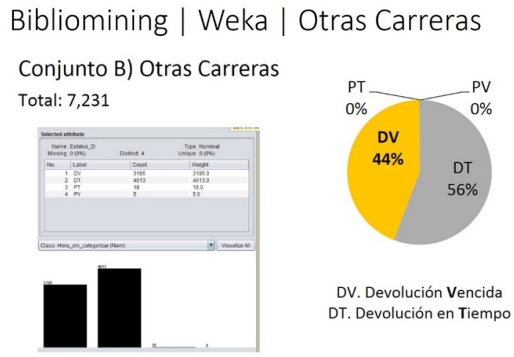
Diapositiva 7



Esta gráfica nos indica que el 99% de los libros que se prestan, son devueltos a la biblioteca, pero de ellos, el 51% se regresa de manera tardía; es decir, después de la fecha indicada en el sistema.

Un fenómeno muy parecido sucede con el conjunto B.1, de las otras carreras.

Diapositiva 8



Tomando las devoluciones vencidas, se decidió utilizar dicho campo como base para implementar la tarea de *Clustering* con el fin de identificar grupos de registros que son similares entre ellos, pero diferentes del resto de los datos.

Los resultados para el grupo A.1 (Médico Cirujano), fueron los siguientes:

Diapositiva 9

```
Final cluster centroids:
                                                Cluster#
Attribute                                Full Data          0          1          2          3
                                         (125909.0) (43046.0) (35589.0) (40945.0) (6329.0)
=============================================================================================
Materia                                  Human Anatomy Internal Medicine     Human Anatomy     Human Anatomy Internal Medicine
Estatus_D                                           DT                DT                DT                DV                DT
Carrera                                  MEDICO CIRUJANO   MEDICO CIRUJANO   MEDICO CIRUJANO   MEDICO CIRUJANO   MEDICO CIRUJANO
Hora_Prestamo                                       TM                TV                TM                TM                TM



Time taken to build model (full training data) : 0.48 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       43046 ( 34%)
1       35589 ( 28%)
2       40945 ( 33%)
3        6329 (  5%)
```

El libro que corresponde a la clasificación de Anatomía Humana, que normalmente se presta en el turno matutino, tiende a devolverse de manera tardía.

Con relación al grupo de otras carreras, los resultados proporcionados por la herramienta fueron:

Diapositiva 10

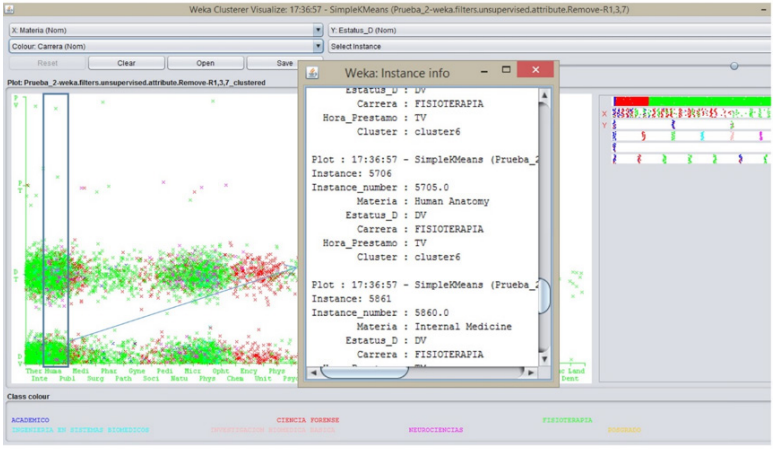
Bibliomining | Weka | Otras Carreras

```
                                5                   6                   7
                          (611.0)             (1958.0)             (137.0)
=============================================================================
Public Aspects of Medicine              Human Anatomy   Natural History, Biology
                       DV                          DV                        DV
                CIENCIA FORENSE         FISIOTERAPIA        FISIOTERAPIA
```

Lo que esto nos indica es que los alumnos de las carreras de Ciencia Forense que obtienen los libros de aspectos públicos de la medicina y los alumnos de Fisioterapia que se llevan en préstamo los libros con clasificación de Anatomía Humana, Historia Humana y Biología, representan a aquellos que devuelven los libros de manera tardía.

Adicionalmente WEKA nos muestra de manera gráfica, cómo es que se visualizan los datos; aquí el ejemplo para el conjunto B, de otras carreras.

Diapositiva 11



3.B Clasificación

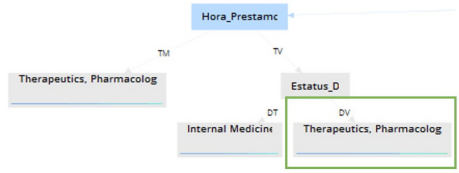
Al ser éste un aprendizaje supervisado, asigna elementos de una colección a categorías o clases de destino.

Software utilizado: RapidMiner es un programa para realizar minería de datos. No es *software* libre, cuenta con una versión educativa.

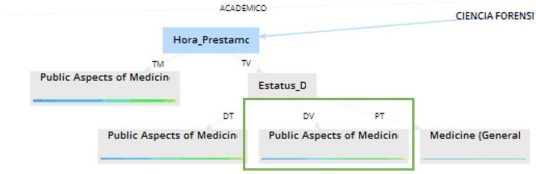
RapidMiner, con la ayuda del asistente, de manera muy rápida, permite establecer la tarea de minería de datos. Siguiendo los pasos del asistente y seleccionando la tarea de clasificación, es posible obtener arboles de decisión, que presentan información de cada una de las carreras.

Diapositiva 12

Bibliomining | rapid miner | Otras Carreras



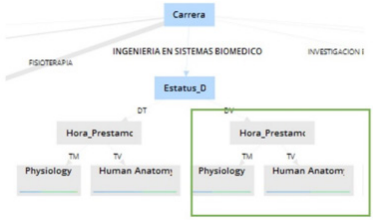
Académico



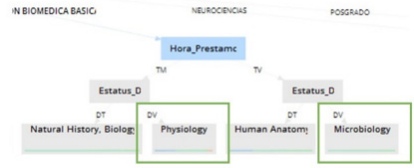
Ciencia Forense

Diapositiva 13

Bibliomining | rapid miner | Otras Carreras

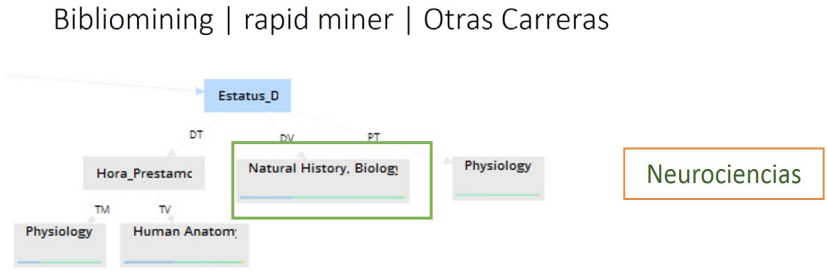


Ingeniería en Sistemas Biomédicos



Investigación Biomédica Básica

Diapositiva 14



Recopilando toda esta información, se puede resumir el conocimiento generado relacionado con el comportamiento de devoluciones tardías.

Conclusiones

Con el apoyo de la estadística, se detectó que un gran porcentaje de los libros que se prestan y que son devueltos, lo son de manera tardía (DV) (44%-46%).

Aplicando tareas de minería de datos, es posible conocer de dichas devoluciones vencidas, a qué clasificación pertenecen y en qué horario fueron prestadas.

Con dicha información se podría establecer que la multa no es factor importante para la devolución del material bibliográfico; se tendría que revisar la política para mejorar el regreso de libros en tiempo.

El hecho de conocer la clasificación de los libros que se devuelven de manera tardía, motiva a realizar nuevos análisis de estudio de la colección, poniendo atención en dichas clasificaciones.

Fuentes consultadas

Bin, Chen. 2013. "Study on Data Mining in Digital Libraries." In , 282–91. *Springer, Berlin, Heidelberg.* https://doi.org/10.1007/978-3-642-53703-5_30.

Candás Romero, Jorge. 2006. "Minería de datos en bibliotecas: bibliominería." 2006. http://bid.ub.edu/17canda2.htm.

Nicholson, Scott. 2006. "The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services." *Information Processing & Management* 42 (3): 785–804. https://doi.org/10.1016/j.ipm.2005.05.008.

Sarma, Pankaj Kumar Deva, and Rahul Roy. 2010. "A Data Warehouse for Mining Usage Pattern in Library Transaction Data." Assam University Journal of Science and Technology. http://www.inflibnet.ac.in/ojs/index.php/AUJSAT/article/view/194.

Zhu, Tingting, and Lili Zhang. 2011. "Application of Data Mining in the Analysis of Needs of University Library Users." 2011 6th International Conference on Computer Science & Education (ICCSE), Computer Science & Education (ICCSE), 2011 6th International Conference On. https://doi.org/10.1109/ICCSE.2011.6028662.

Juan Camilo Giraldo, Mejía, and Builes Jovani Alberto Jiménez. "Caracterización del Proceso de Obtención de Conocimiento y Algunas Metodologías para Crear Proyectos de Minería de Datos." *Revista Latinoamericana de Ingeniería de Software,* Vol 1, Iss 2, Pp 42-44 (2013) no. 2 (2013): 42. Directory of Open Access Journals, EBSCOhost (accessed September 7, 2018).

Gutiérrez Hernández, Guadalupe Vanessa Carolina, Verónica Barranco Serrano, and Carlos Francisco Méndez Cruz. Minería de datos dentro del proceso de KDD aplicado a la base de datos de circulación bibliográfica de la Biblioteca Central. n.p.: 2008. TESIUNAM, EBSCOhost (accessed September 7, 2018).

Prakash, K & Chand, Prem & Gohel, Umesh. (2004). Application of Data Mining in Library and Information Services. Presented at the 2nd Convention PLANNER, Manipur Uni., Imphal.

Weka 3: Data Mining Software in Java. https://www.cs.waikato.ac.nz/ml/weka/

RapidMiner. Lightning Fast Data Science for Teams. https://rapidminer.com/

Manejo de datos. Una aproximación desde los estudios de la información. La edición consta de 100 ejemplares. Coordinación editorial, Israel Chávez Reséndiz; revisión especializada, Francisco Xavier González y Ortiz; revisión de pruebas, Valeria Guzmán González; formación editorial, Natalia Cristel Gómez Cabral. Instituto de Investigaciones Bibliotecológicas y de la Información / UNAM. Fue impreso en papel cultural de 90 gr. en los talleres de Grupo Fogra. Año de Juárez 223. Col. Granjas San Antonio. Alcaldía Iztapalapa. Ciudad de México. Se terminó de imprimir en febrero de 2020.