Búsqueda de patrones para mejorar productos y servicios en las bibliotecas

Esther Marina Ruiz Lobaina*
Pedro Lázaro Romero Suárez**

Artículo recibido:
14 de agosto de 2015.
Artículo aceptado:
9 de septiembre de 2016.

Resumen

Este trabajo muestra los resultados alcanzados durante la búsqueda de patrones ocultos, aplicando algoritmos estadísticos, a una base de datos bibliográfica. Para esta investigación se seleccionó el software *WinIDAMS v.1.3*, que utiliza para el manejo de los datos la construcción de un *dataset* IDAMS (BUILD) y la agrupación de datos (AGGREG), para el análisis estadístico los algoritmos Análisis de conglomerados (CLUSFIND) y los diagramas de dispersión (SCAT). Para las salidas de los resultados este software ofrece las tablas multidimensionales, capaces de crear por cada grupo de variables seleccionadas una tabla interna con resultados como la frecuencia y la

* Ministerio de Ciencias, Tecnologías y Medio Ambiente (CITMA). Instituto de Información Científica y Tecnológica (IDICT), Cuba. marinajfr@yahoo.com
** Ministerio de Educación Superior (MES). Instituto Superior de Tecnología y Ciencias Aplicadas (INSTEC), Cuba. lromerocu@instec.cu

media aritmética, que fueron las seleccionadas para estas pruebas, mientras que para la representación gráfica de los resultados se decidió utilizar los histogramas porque son gráficas de barras que permiten interpretar de forma muy fácil y rápida el comportamiento de las variables seleccionadas para el análisis. Este estudio encontró patrones a través de la clusterización con los cuales fue posible potenciar los servicios de difusión selectiva de la información y proponer nuevos servicios para que formen parte de los productos que brinda la biblioteca.

Palabras clave: Ciencias de la Información; Gestión de la Información; Tablas Multidimensionales; Minería de Datos; Estadística; Patrones; Histogramas.

Abstract

Finding patterns to improve products and services in libraries

Esther Marina Ruiz-Lobaina and *Pedro Lázaro Romero-Suárez*

This work shows the results obtained during the search for hidden patterns using statistical algorithms, in a bibliographic database. For this research the WinIDAMS v.1.3 software, used for data management, building an IDAMS dataset (BUILD) and Data Group (AGGREG) for statistical analysis algorithms, Cluster analysis was selected (CLUSFIND) and scatterplots (SCAT). In addition to the outputs of the results this software provides multidimensional tables, able to create for each group of selected variables, an internal table with results such as frequency and average arithmetic were selected for these tests, while for graphical representation of the results, it was decided to use histograms, because they are bar graphs that allow us to interpret very easily and quickly, the behavior of the variables selected for analysis. This study found patterns through clustering, with which services could enhance Selective Dissemination of Information and propose new services, to become part of the products offered by the library.

Keywords: Information Science; Information Management; Multidimensional Tables; Data Mining; Statistics; Patterns; Histograms.

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57830

Introducción

El estudio de grandes bases de datos con el empleo de la minería de datos o la estadística se ha aplicado desde hace varias décadas sólo en instituciones y empresas que manejan grandes volúmenes de información, con lo que, al no aplicarse de forma generalizada, se pierden las ventajas que aportan.

En este caso se encuentra la Red de Bibliotecas de Ciencias y Técnicas[1] y por esta razón se comenzaron investigaciones, con el asesoramiento de la Universidad de Ciencias Aplicadas,[2] para estudiar las bases de datos que guardan la información bibliográfica de esta red.

La primera razón que motivó el estudio de la información bibliográfica de esta base de datos es la necesidad de encontrar patrones que permitieran potenciar los servicios que se ofrecen en estas bibliotecas y lograr, con el uso eficiente de las ciencias de la información, una vía que permita un aporte mayor a la gestión del conocimiento; la segunda razón de este estudio se refiere a preparar la información de esta base de datos, seleccionada para entrar en el marco de esta investigación por su tamaño y por la cantidad de consultas que recibe a diario.

La reingeniería aplicada a la información de esta base de datos estuvo dirigida a dos aspectos fundamentales, el primero a la estructura de la base de datos donde se encuentra la información que se quiere analizar y el segundo a la información guardada en los campos. Este trabajo tomó más del 50 % del tiempo de la investigación porque la base de datos se encontraba con una cantidad considerable de campos vacíos e información sin estandarizar, problema que es imprescindible resolver antes de someterla a cualquier estudio, y también porque fue necesario darle una nueva estructura y formato que le permitiera ser aceptada por estos algoritmos estadísticos.

Sobre el estado del arte de las herramientas digitales dedicadas a la extracción de patrones, existen muchas que han sido creadas para trabajar el concepto de la minería de datos con las cuales se puede extraer perfectamente los patrones que se quieren lograr, pero para estas pruebas se decidió escoger la herramienta *WinIDAMS v.1.3* porque esta herramienta está considerada como un paquete estadístico que sirve tanto para analizar la información numérica como a la información textual. Es una herramienta digital capaz de procesar grandes bases de datos y la distribución del software se hace de

1 Esta red tiene alcance a nivel nacional y está dirigida por el Instituto de Información Científica y Tecnológica (IDICT), perteneciente al Ministerio de Ciencia, Tecnología y Medio Ambiente de Cuba.

2 Universidad autorizada por el Ministerio de Enseñanza Superior (MES) para desarrollar programas doctorales y de investigación.

forma libre, requisitos muy oportunos para optar por esta herramienta. A esto se suma que los algoritmos estadísticos que proporciona esta herramienta son extremadamente fáciles de implementar y guardan bastante similitud con algunos de los utilizados tradicionalmente por la minería de datos, como es el caso de la obtención de los clústeres de información (Unesco, 2011).

El interés en la aplicación de técnicas de este tipo se acentúa cuando se acepta que la extracción de patrones ocultos en grandes volúmenes de información es la forma más novedosa de procesamiento de la información y la única vía hasta el momento para extraer información desconocida y útil, que a través del tipo de patrón que se logre extraer brinda la posibilidad de reorientar nuevos análisis, nuevos productos o nuevos servicios que impulsan la gestión del conocimiento en cualquier área del saber donde pertenecen esos patrones.

Haciendo una síntesis del trabajo que realizan las bibliotecas con la información que poseen en sus fondos a través del empleo de los Sistemas de Gestión Bibliotecarios (SGB) (Arriola Navarrete y Butrón Yáñez, 2008), se puede decir que la gran mayoría no utilizan las técnicas de análisis de la información por diferentes razones, bien porque aún no llegan a un acuerdo sobre el asunto de los derechos de autores, bien porque en realidad son técnicas difíciles de implementar por los bibliotecarios, o simplemente porque se encuentran adaptados al trabajo de los módulos que ofrecen los SGB, los cuales están creados para manejar la catalogación, consulta, recuperación, préstamos, circulación, etc., que son los servicios que diariamente ofrecen las bibliotecas, pero que en materia de búsqueda de patrones ocultos y útiles o resultados estadísticos novedosos no incorporan estas técnicas y no ofrecen la posibilidad de hacer estos análisis desde la misma herramienta. Además, está demostrado que los algoritmos de gestión de información funcionan de forma muy diferente a los algoritmos de análisis que se utilizan en la estadística o en la minería de datos (Hernández Orallo y Ferri Ramírez, 2006).

Por lo tanto, se puede afirmar que los módulos de recuperación de información de los SGB están preparados sólo para localizar y recuperar de las bases de datos asociados a ellos aquellas palabras o metadatos que coinciden con la búsqueda del usuario y de mostrarlo en forma de listado, mientras que un proceso de minería de datos o un proceso estadístico puede no ofrecer ningún dato de los que se encuentran físicamente en el volumen de información de la base de datos, sino que ofrece patrones que reflejan el comportamiento de esa información en un periodo de tiempo.

Se puede afirmar que, con sólo el uso de un SGB (Redacción, 2014), no se logran extraer los patrones que aportan la estadística o la minería de datos y que por tanto se desperdicia la ocasión de servirse de esos patrones para

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57830

potenciar otros procesos, como los sistemas de difusión selectiva de la información (DSI) o la vigilancia tecnológica (VT). Un ejemplo se puede exponer en el caso de patrones con mayor número de frecuencia, que infieren que existe un mayor interés por parte de los investigadores sobre esas áreas de investigación, es decir, un patrón con mayor número de frecuencia tiene un mayor número de investigaciones realizadas, por lo tanto, tomándolo como referencia se puede hacer un seguimiento en internet. Por ejemplo, en las bases de datos de patentes, es posible detectar el estado de las investigaciones publicadas sobre las mismas temáticas en otras partes del mundo y tomar decisiones sobre investigaciones futuras.

En resumen, con una base de datos que esté preparada para soportar el doble procesamiento de información, es decir, tanto el proceso de consulta con su habitual módulo de recuperación de información utilizado por los SGB (Lamarca, 2013) como un proceso de análisis más riguroso, como es el caso de los algoritmos estadísticos o de minería de datos (Molina, 2002), se puede considerar que se ha logrado el máximo de aprovechamiento porque estas son las dos vías de procesamiento más reconocidas hasta la fecha para el manejo y análisis de grandes volúmenes de información. Este es un objetivo que se debe tratar de implementar para todas las bases de datos de las bibliotecas porque no sólo pone a las bibliotecas en una situación muy ventajosa con respecto a otras, sino que convierte su servicio en vanguardia de la gestión de la información y del conocimiento al hacer posible el máximo de aprovechamiento de su información.

Materiales y métodos

Entre los materiales seleccionados para este estudio está presente una base de datos con 10 492 registros de información bibliográfica. A esta base de datos se le aplicó una reingeniería que comenzó desde la selección de los campos que serían sometidos al análisis estadístico hasta la acostumbrada revisión y estandarización de la información, procesos que son imprescindibles para evitar falsos errores en los resultados de los algoritmos estadísticos aplicados. La base de datos quedó finalmente conformada con 13 campos de la siguiente manera:

- Segmento: área de la ciencia a la cual pertenece la publicación.
- Revista: nombre de la revista que publica el documento.
- Tipo de documento (TD): Premio Academia, Mención, Tesis Doctoral, etc.

- Year: año de publicación.
- Idioma en que está redactado el documento.
- Keyword (1-6): los 6 metadatos que están relacionados a la investigación del autor.
- Autor: nombre y apellido del autor del documento.

Se debe aclarar que los nombres de estas variables, en su mayoría, son los mismos que tienen los campos de la base de datos original; en el caso de la variable *Year* se escogió su nombre en inglés para evitar la palabra año que contiene el carácter especial "ñ" y en el caso de *Keyword* para evitar nombres compuestos como "Palabras claves".

Concluida la selección de los campos, se exportó la información a un fichero de texto con extensión .txt y se utilizó el procesador de texto Notepad ++ para hacer los arreglos a la nueva estructura de la información, que requiere la importación a Microsoft Excel 2010. Esta nueva estructura mantuvo los nombres de campos antes mencionados y toda la información de los 10 492 registros, vale aclarar que se verificó que todos estos campos estuvieran separados por comas por ser requisito indispensable de identificación durante el proceso de importación a la tabla de Microsoft Excel 2010 y que se mantuvieran guardados en un fichero con formato de texto (.txt).

Este nuevo fichero fue importado en una tabla de Microsoft Excel 2010 e inmediatamente se sometieron a segunda reingeniería, que estuvo orientada a:

- la reducción y estandarización de los nombres de revistas,
- la revisión y rellenado de los campos que estaban vacíos,
- la eliminación de los errores ortográficos que se originaron en el proceso de captación de la información durante la etapa de creación de la base de datos.

Estos tres pasos fueron obligatorios porque durante la reducción y estandarización de los nombres de revistas se encontraron casos que mostraban el nombre de una misma revista escrito en diferentes formas, además de ser nombres excesivamente grandes que de someterse a los algoritmos de estadística traerían problemas.

Por otro lado, la revisión y rellenado de los campos que estaban vacíos fue necesaria porque los algoritmos que se aplicaron a la información están basados en las estadísticas que incluyen conteos y sumas de celdas con información. Por lo tanto, se analizó lo siguiente:

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57830

- Si los campos quedaban vacíos no serían contabilizados por el algoritmo.
- Si los campos vacíos eran retirados, entonces se tendría que retirar una cantidad considerable de registros que se encontraban incompletos en uno o más celdas dentro de la tabla, afectando considerablemente el volumen de la información que se quería analizar.

De esta forma, ya decidido que los campos serían rellenados porque no se debía eliminar tanta información, se analizó si el campo en su concepción debía guardar información numérica o texto. Con base en este nuevo análisis, se procedió de la siguiente manera: los campos que debían tener texto se rellenaron con el carácter especial "?", mientras que en el caso del campo *Year*, que debía tener números, las celdas se completaron con un cero "0".

Terminada la reingeniería de la información, para la cual se utilizó Microsoft Excel 2010, se obtuvo una tabla con 12 columnas, la cual se guardó con extensión .csv porque este formato es aceptado por el proceso de importación del software *WinIDAMS v.1.3*.

A continuación se procedió a la importación de la información en formato .csv al software *WinIDAMS v.1.3*, los nombres de las columnas de la tabla de Microsoft Excel 2010 fueron convertidos por el algoritmo BUILD (proceso interno de esta herramienta) en los nombres de las variables utilizados por la dataset y el diccionario que crea esta herramienta para sus análisis posteriores.

Con este proceso de preparación de la información ya terminado se puede garantizar que la aplicación de los algoritmos estadísticos de *WinIDAMS v.1.3* aportará los resultados esperados y con ello se cumpla el objetivo fundamental que plantea su tutorial: "La idea en IDAMS, es poner a disposición de los Estados Miembros de UNESCO, exento de costo, un paquete de programas para el manejo y el análisis estadístico de datos. […] entrega a los Estados Miembros de un paquete de programas integrado que permite el procesamiento de datos de texto y numéricos de una manera unificada para propósito científico y administrativo en universidades, institutos de investigación, administraciones nacionales, etc." (Unesco, 2016a),[3] por tanto es una herramienta que garantiza los buenos resultados. A estas ventajas se agrega que no necesita grandes prestaciones tecnológicas en la computadora donde

3 IDAMS. Esta herramienta proviene originalmente del paquete estadístico OSIRIS III.2 desarrollado al comienzo de la década de los años 70 en el Instituto para la Investigación Social de la Universidad de Michigan en los Estados Unidos de América. Ha sido y continúa siendo enriquecido, modificado y puesto al día por el Secretariado de la Unesco con la cooperación de expertos de diferentes países, a saber: especialistas belgas, británicos, colombianos, eslovacos, estadounidenses, franceses, húngaros, poloneses, rusos y ucranianos; de ahí el nombre *Internationally Developed Data Analysis and Management Software Package (Paquete de software para el análisis y manejo de datos desarrollado internacionalmente*).

se va a instalar y ejecutar, y brinda un entorno de trabajo amigable que permite que sea fácil de trabajar.

Entre los resultados que brinda este software se seleccionaron las

> Tablas multidimensionales [porque] permite visualizar y personalizar tablas con frecuencias, porcentajes de fila, de columna y totales, estadísticas univariadas (suma, conteo, media, máximo, mínimo, variancia, desviación estándar) de variables adicionales y estadísticas bivariadas. Se pueden anidar hasta siete variables en filas y columnas. Se puede repetir la construcción de tablas para cada valor hasta tres variables de "página". También se pueden imprimir las tablas o exportarlas en formato libre (coma o carácter de tabulación como delimitador) o en formato HTML. (Unesco, 2016b)

Se logró una importación completa de los 10 492 registros de la tabla de Microsoft Excel 2010 al software *WinIDAMS v.1.3* y se sometieron al proceso de recodificación por ser datos de tipo texto en su gran mayoría. Este proceso es muy simple, se logra con sólo seleccionar la pestaña del diccionario y de la lista de campos se seleccionan todos los datos de texto en la casilla de recodificación, excepto, en este caso, la variable *Year* por ser numérica.

Terminado este último paso, quedó creada la dataset y el diccionario que utiliza *WinIDAMS v.1.3* y comienza la creación de las tablas multidimensionales, para lo cual es necesario hacer previamente una selección del juego de variables que se van a relacionar en estos resultados. Este tipo de tablas permite graficar los resultados a través de los histogramas y otros tipos de gráficos fáciles de interpretar, lo que permitió que estos resultados fueran aceptados con agrado por los bibliotecarios.

Es importante recordar que las herramientas digitales utilizadas son libres y la metodología propia, desde la selección y preparación de la información, el análisis de los patrones encontrados hasta la propuesta de las mejoras de servicios, y se rige por la lógica que impone el paso de un proceso a otro según la herramienta utilizada, logrando con esto un ahorro significativo por no incurrir en gastos por conceptos de aplicación de herramientas y metodologías ya conocidas.

Con la selección y reingeniería de la información el proceso de importación de la información a la herramienta y la búsqueda de patrones terminados se puede considerar que se ha vencido el 50 % del proceso total y que el 50 % restante continúa con el análisis de los patrones encontrados, su aplicación y la mejora de productos y servicios, que es el principal objetivo de esta investigación.

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57830

Resultados y discusión

Los patrones que muestran las tablas multidimensionales no deben crear preocupación sobre los derechos de autor y la propiedad intelectual, porque como plantea IFLA:[4]

> En 2011, la evaluación sobre Propiedad Intelectual y Crecimiento *The Hargreaves Review on Intellectual Property and Growth in the United Kingdom* presentó la recomendación de adoptar excepciones a la legislación sobre Derechos de autor y reproducción, con el objetivo de facilitar la minería de datos y conseguir los destacados beneficios significativos que esta práctica ofrece para estimular la innovación y desarrollo de nuevo conocimiento. (IFLA, 2013)

Para evitar posibles conflictos sobre derecho de autor que puedan aparecer en el futuro, el juego de variables que se escogió para realizar este estudio no comprende el campo *Autor*, como se muestra en la *Figura 1*, por lo tanto, no hay forma de relacionar a un autor o a un grupo de autores y sus datos personales con los patrones encontrados, a pesar de estar asociado a los metadatos que se estaban estudiando.

Para poder llegar a un autor o a sus datos personales es necesario ser el propietario de la base de datos, que es el único que tiene a su alcance tal información; con esto se demuestra que existen muchas formas de realizar un análisis estadístico o de minería de datos con el cual se pueda extraer patrones útiles, aun en el caso de que no existan acuerdos sobre la publicación de la propiedad intelectual en forma libre.

Para el estudio de este caso se seleccionó un juego de variables que comprende *Year* como variable de página, *Segmento* como variable de fila, *Keyword1* como variable de columna y *Revista* como variable de celda.

En la *Figura 1* se muestra que por cada variable de página (*Year*) se crea una tabla que agrupa por *Segmento* y por *Keyword1* la información de las *Revistas* por años en específico. Estas tablas hacen totales por filas y por columnas. También se observa que existe una variable de página (*Year*) con valor 0 porque estos son los años que se encontraban vacíos y fueron rellenados para poderlos contabilizar, de lo contrario el algoritmo no los hubiera tomado en cuenta y no se conocería el número de registros con esta variable vacía.

4 IFLA (*International Federation of Library Associations and Institutions*) es una organización mundial creada para proporcionar a bibliotecarios de todo el mundo un foro para intercambiar ideas, promoviendo la cooperación, la investigación y el desarrollo internacionales en todos los campos relacionados con la actividad bibliotecaria y la bibliotecología. Fue fundada en 1927 en Edimburgo, Reino Unido. La Biblioteca Real (Biblioteca Nacional de los Países Bajos), en La Haya, le proporciona desinteresadamente el espacio donde se ubica la sede.

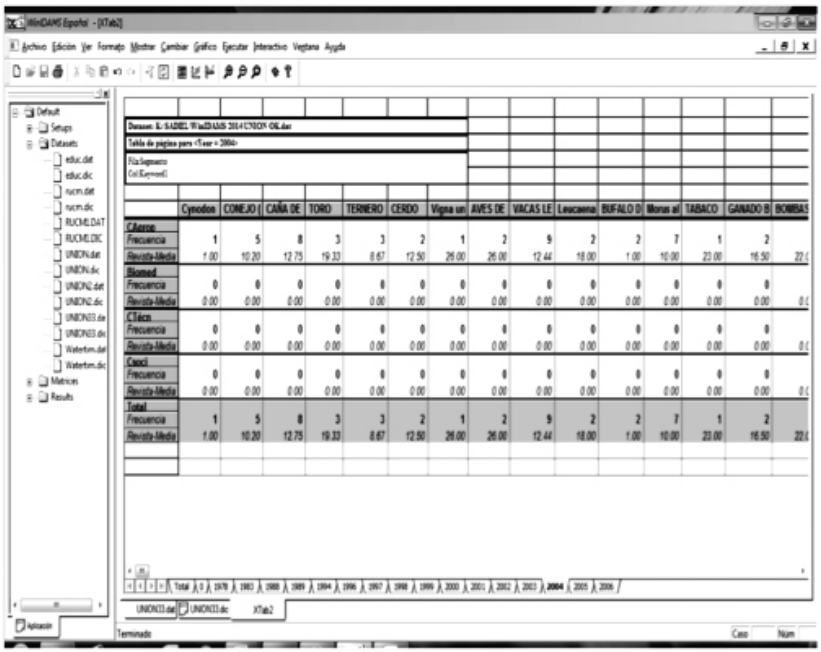

Figura 1. Tablas Multidimensionales logradas en *WinIDAMS v.1.3*

Para presentar los resultados se creó un informe con el listado de las *Keywords1* con valor de frecuencia por encima de 6 y ordenado por *Year* y *Segmento*, de igual manera se creó un segundo informe con frecuencia por debajo del valor 4. Para ambos casos se consideraron los casos extremos, basados en la relación que a mayor frecuencia, mayor número de investigaciones de esa palabra clave y por relación directa un mayor interés en ese tema y viceversa para los patrones de menor frecuencia, donde menor frecuencia, menos investigaciones sobre el tema y por lo tanto menor interés por parte de los investigadores.

Para ilustrar cómo mejorar los productos y servicios bibliotecarios de una forma rápida, se hizo un seguimiento en Internet de algunos patrones encontrados con la intención de detectar los tipos de estudios similares que han realizado otros investigadores en el mundo, este tipo de resultado permite ofrecer una nueva forma de Difusión Selectiva de Información (DSI) a los investigadores interesados.

Tomando como primer ejemplo la *Keyword1* con mayor frecuencia se obtuvo lo siguiente:

DOI: http://dx.doi.org/10.22221/ilbi.0187358xp.2017.72.57830

Vacas Lecheras con 9 investigaciones, *Year*: 2004, *Segmento*: CAgrop.

Con el patrón "Vacas Lecheras" se hizo una búsqueda en el sitio web de la OMPI,[5] que es una base de datos de patentes que permite hacer búsquedas de forma gratuita en Internet. Como resultado de esta búsqueda se encontró que existen registrados seis trabajos relacionados con "Vacas Lecheras" en diferentes años:

1. El Parmesano, rey de los quesos
   La alimentación de las *vacas lecheras* está regulada: sólo heno, ni piensos ni alimentos fermentados. Normas de producción: desde 1991, el envasado del...
   www.wipo.int/wipo_magazine/es/2011/01/article_0005.html
2. La corona del rey del queso es la P.I.
   Las granjas y las centrales *lecheras* que producen el Parmigiano Reggiano... son las habilidades y los métodos tradicionales de cría de *vacas* y de producción...
   www.wipo.int/ipadvantage/es/details.jsp?id=3664
3. La Propiedad Intelectual en Las Pequeñas y Medianas Empresas…
   Formato de archivo: PDF/Adobe Acrobat
   Vacas lecheras (cab). 17,2. 244,7. 120,9. 232,4. 615,2. Ovinos (cab). 225,8. 999, 1. 1.618,0. 824,0. 3.666,9. Caprinos (cab). 160,9. 622,7. 80,0. 175,4. 1.039,0.
   www.wipo.int/edocs/pubdocs/es/sme/795/wipo_pub_795.pdf
4. Page 1 OMPI/PIIJ U/PAN/97/1.C ORIGINAL: Espafiol FECHA: Julio…
   Formato de archivo: PDF/Adobe Acrobat
   15 Jul 1997... utilización de la imagen de una *vaca lechera* como marca de una margarina,... La *vaca lechera* da la impresión de que el producto de…
   www.wipo.int/.../OMPI.../OMPI_PI_JU_PAN_97_1%20C_S.pdf
5. REGLAMENTO (Berrocal) no 1234/2007 DEL CONSEJO de 22 de **...**
   Formato de archivo: PDF/Adobe Acrobat
   22 Oct 2007... proseguir la reestructuración de la producción *lechera* y de mejorar el medio... exclusivamente a partir de leche de *vaca* producida en la. www.wipo.int/edocs/lexdocs/laws/es/eu/eu129es.pdf

5 OMPI (Organización Mundial de la Propiedad Intelectual) es un organismo especializado del Sistema de Naciones Unidas, creado en 1967 con la firma de la Convención de Estocolmo. Presta servicios mundiales para proteger la P.I. en todo el mundo y para resolver controversias; ofrece infraestructura técnica para conectar los sistemas de P.I. y compartir los conocimientos.

6. Español
   Formato de archivo: PDF/Adobe Acrobat
   24 Jun 2011... campesinos, que escasamente poseían una o dos vacas cada uno,... a) La rentabilidad económica de las explotaciones lecheras en la zona...
   www.wipo.int/edocs/pubdocs/es/geographical/798/wipo_pub_798.pdf

Se realizó una segunda búsqueda con la *Keyword1* que seguía en menor frecuencia, en este caso "Caña de Azúcar", con 8 investigaciones, *Year*: 2004, *Segmento*: CAgrop. Se obtuvo un listado en la base de patentes de la OMPI de 53 trabajos de diferentes autores. Debido a lo extenso de esta lista, sólo se muestra el primer resultado como ejemplo:

1. Estadísticas de la *caña* de azúcar de las zafras 2004/2005 a 2008...
   Formato de archivo: Microsoft Powerpoint
   20 Ago 2008 ... 92 MUNICIPIOS ABASTECEDORES DE *CAÑA DE AZÚCAR* DEL ESTADO DE VERACRUZ. ESTADÍSTICAS DE LA *CAÑA DE AZÚCAR* DE...
   www.wipo.int/edocs/.../wipo_smes_ver_09_theme01_2.ppt

Estos resultados, obtenidos en las búsquedas guiadas por los patrones encontrados a través de los algoritmos de estadística, permiten proponer nuevos productos y servicios, es decir, se puede ofrecer una nueva forma de DSI y con esto garantizar una mejor gestión de la información y del conocimiento; también se pueden fortalecer e impulsar otros proyectos, como los distintos observatorios de la ciencia donde se realiza la vigilancia tecnológica (VT).

La información que brindan las tablas multidimensionales permite conocer sobre los metadatos que tienen baja frecuencia, e inclusive de aquellos que ya no se investigan más desde hace algún periodo de tiempo. Para estos casos también se puede hacer un seguimiento en Internet y conocer el interés de otros investigadores sobre estos temas que no se están estudiando en el país.

Con estos resultados, los bibliotecarios no sólo pueden reforzar el servicio de DSI o los observatorios de las ciencias con la VT, sino también pueden crear nuevas bases de datos que sirvan de nuevos fondos a las bibliotecas.

DOI: http://dx.doi.org/10.22221/ilbi.0187358xp.2017.72.57830

Resultados gráficos

Después de calculadas las tablas multivariables existen otras opciones de exploración gráfica que el software proporciona, como muestra la *Figura 2*. Este tipo de tabla hace un ploteo de la matriz completa de los datos y estos resultados sirven para hacer estudios sobre el comportamiento general y puntual de las variables de fila, columna y celda presente en las tablas multivariables. Para el caso de los comportamientos puntuales de las variables, proporciona ventanas emergentes que presentan los valores estadísticos específicos de cada punto o registro que se encuentran ploteado en cada cuadrante de la gráfica (ver *Figura 2*).

Este tipo de exploración gráfica permite tener una valoración mucho más exacta del comportamiento que tiene toda la información de la base de datos, además de demostrar que la selección de las variables que se han escogido para el estudio es la correcta, aunque también se debe estudiar más variables de la base de datos en el caso de existir. Este procedimiento se puede repetir tantas veces como el número de variables lo permita.

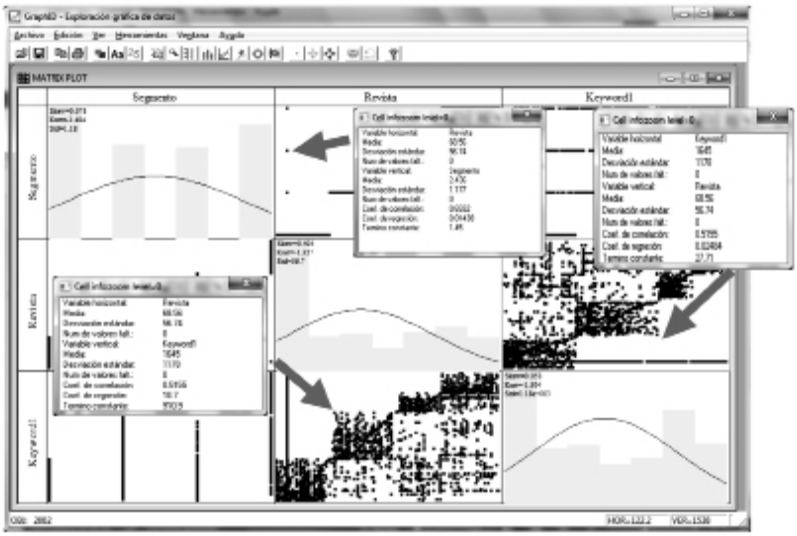

Figura 2. Matrix Plot. Exploración gráfica de datos con *WinIDAMS v.1.3*

En el caso de los valores globales esta gráfica ofrece los histogramas que se encuentran en las celdas donde coinciden una misma variable, junto a los datos estadísticos de Skew (NIST, 2013), que es un resultado estadístico utilizado para determinar el grado de asimetría que posee la distribución muestral, es decir, la distribución de toda una muestra de datos. Por ejemplo, algunos procesos de estimación y decisión son considerados más verosímiles cuanto más simétrica sea la distribución muestral (García, 2012); también presenta el dato de Kurtosis o Kurt (Statistics, 2014), que es otro resultado estadístico utilizado para determinar si las colas de la curva que representa la asimetría tienen una masa o altura diferente a la de una distribución normal (Vitutor, 2014), y la desviación estándar Std (Mora, 2010) indica la medida de dispersión de cuánto pueden alejarse los valores respecto al promedio o media (Rivera, 2011). Todos estos datos estadísticos son útiles para estudiar el comportamiento de todo el volumen de la información junto a otros resultados más específicos como el coeficiente de regresión y el coeficiente de correlación (Berrocal, 2012), que sirve para conocer el grado de relación entre las diferentes variables que se están estudiando.

En este estudio los resultados globales permitieron conocer que la distribución muestral que se tomó tiene un comportamiento ascendente en forma lineal y que se denotan los cuatro bloques de información existentes en la variable *Segmento*, además de un gran número de registros que entrecruzan su *Keyword1*, es decir, registros que están asentados en un *Segmento* pero cuya *Keyword1* muestra información que no está en relación al *Segmento* asignado, formado una nube al costado de los cuatro bloques, el gráfico refleja perfectamente el solapamiento que producen estas *Keyword1* dentro de los *Segmentos*. Los gráficos de ploteo están entre las variables *Segmento-Revista*, *Revista-Keyword1*.

En el caso de los resultados puntuales el gráfico es sensible y con sólo pasar el puntero del mouse sobre los puntos del gráfico que se quiere analizar aparece una ventana emergente con información sobre las dos variables que dieron origen a ese punto, esta ventana emergente muestra una serie de información estadística incluyendo el grado de correlación y regresión entre las variables.

Por ejemplo, en la *Figura 2*, de izquierda a derecha, se aprecia que la primera ventana emergente corresponde al punto que está situado en la celda formada por la *Revista* y *Keyword1* y que existe un grado de correlación de estas dos variables de 0.5155; este valor indica que es una correlación buena porque el valor máximo es 1 (Yat, 2008). El siguiente punto estudiado se encuentra en la celda formada entre *Revista* y *Segmento*, en este caso muestra una ventana emergente con un coeficiente de correlación de 0.6962, por lo

DOI: http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57830

tanto existe un mayor grado de correlación entre estas dos variables, lo que prueba que la selección de las variables *Revista* y *Segmento* para el estudio es buena. En el caso de la tercera ventana el coeficiente de correlación es nuevamente 0.5155 porque son las mismas variables (*Keyword1* y *Revista*).

Estos resultados sirven para comprobar que la selección de las variables que se han estudiado es una selección acertada porque existe una buena correlación entre ellas.

Este estudio se puede repetir utilizando las variables *Keyword2*, *Keyword3* y hasta *Keyword6*, y cada una de ellas aportará valores dependiendo del contenido o información que tiene la *Keyword* que se encuentre en estudio en ese momento. Es bueno recordar que existe un factor que puede variar el grado de correlación entre las variables y que depende directamente del autor cuando asigna las palabras claves a su artículo, porque es visto con frecuencia que el grado de correlación entre el metadato y el tema del artículo no es buena, resultando con esto que al estudiar las palabras clave no se refleja realmente el grado de correlación real de las variables que se encuentran en estudios.

Hasta aquí se lograron dos tipos de resultados muy útiles, uno que está relacionado con la frecuencia y otro que ofrecen las tablas multidimensionales (*Figura 1*), que en realidad son patrones logrados por tiempo (*Year*), que permiten conocer la cantidad de artículos o de investigaciones realizadas por cada uno de los metadatos, mientras que el segundo gráfico consiste en la exploración gráfica y en la exploración estadística de toda la información a través de ventanas emergentes (*Figura 2*), que permiten comprobar que la selección de las variables estudiadas es acertada para el estudio porque son variables que tienen una correlación buena, por encima del 0.5.

Se comprueba que aunque el software *WinIDAMS v.1.3* está considerado un paquete absolutamente estadístico, evidentemente emplea algoritmos estadísticos al igual que lo hacen los software dedicados a la minería de datos, por tanto son algoritmos que han permitido extraer patrones de gran utilidad para esta investigación. Entre estos valores están los de frecuencia, que se pueden tomar como valores de clústeres, y los de correlación, que han servido para comprobar que la selección de las variables que han participado en el estudio es la correcta.

Conclusiones

Dentro de los resultados alcanzados se lograron satisfacer los objetivos propuestos para la investigación:

1. Mejorar la calidad de la información en la base de datos, por lo que también mejoró el funcionamiento del sistema de recuperación del sistema gestor.
2. Los valores estadísticos encontrados como mayor y menor frecuencias de las tablas multivariables sirvieron como patrones de comportamiento para proponer una nueva forma de Difusión Selectiva de Información (DSI) independiente a la tradicional que ofrecen las bibliotecas, además de ser material para crear una nueva base de datos que puede ser utilizada en reforzar la vigilancia tecnológica y potenciar la gestión de la información en las bibliotecas.
3. Contar con los nuevos fondos creados para las bibliotecas permitió un ahorro de tiempo significativo para los investigadores interesados en esas temáticas recuperadas.
4. La implementación de algoritmos estadísticos o de minería de datos es una forma de análisis novedosa para las bibliotecas que no cuentan con estas técnicas, independientemente del motor de búsqueda y recuperación del SGB que tengan en uso, porque son sistemas de tratamiento de información totalmente diferentes.
5. La usabilidad del software libre garantiza que no existan gastos por la implementación de esta herramienta y tampoco por la contratación de personal altamente calificado.

Bibliografía

Arriola Navarrete, Óscar, K. Butrón Yáñez. 2008. "Sistemas integrales para la automatización de bibliotecas basados en software libre". *Biblioteca Virtual de la Salud*. http://bvs.sld.cu/revistas/aci/vol18_6_08/aci091208.htm

Berrocal S., Celia. 2012. *Distribuciones bidimensionales. Regresión y correlación*. http://recursostic.educacion.es/descartes/web/materiales_didacticos/Correlacion_regresion_recta_regresion/correlacion_y_regresion.htm

García C., María J. 2012. *Distribuciones muestrales*. http://recursostic.educacion.es/descartes/web/materiales_didacticos/inferencia_estadistica/distrib_muestrales.htm

Hernández Orallo, José, Cèsar Ferri Ramírez. 2006. *Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software. T.2 Minería de Datos y Extracción de Conocimiento de Bases de Datos*. http://users.dsic.upv.es/~jorallo/docent/doctorat/t2a.pdf

IFLA. 2013. *IFLA publicó su Declaración sobre Bibliotecas y Minería de Datos*. http://www.anabad.org/noticias-anabad/28-bibliotecas/2161-ifla-publico-su-declaracion-sobre-bibliotecas-y-mineria-de-datos

Lamarca Lapuente, M. J. 2013. SGBD y STRID. "Hipertexto: El nuevo concepto de documento en la cultura de la imagen". Tesis doctoral, Universidad Complutense

de Madrid. http://www.hipertexto.info/documentos/sgbd.htm
Microsoft, SQL Server. 2016a. *Contenido del modelo de minería de datos para los modelos de asociación* (*Analysis Services - Minería de datos*). https://msdn.microsoft.com/es-es/library/cc645767.aspx
Microsoft, SQL Server. 2016b. *Contenido del modelo de minería de datos para los modelos de regresión lineal* (*Analysis Services - Minería de datos*). https://msdn.microsoft.com/es-es/library/cc645754.aspx
Molina Félix, L. C. 2002. *Data mining: torturando a los datos hasta que confiesen*. http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html
Monografias.com. 2016. *Desviación estandar*. http://www.monografias.com/trabajos89/desviacion-estandar/desviacion-estandar.shtml
Mora, L. A. 2010. "Qué es la desviación estándar y como interpretarla #1", *Trading Center. Formación & Información para el Inversor*. http://tradingcenter.wordpress.com/2009/11/11/que-es-la-desviacion-estandar-y-como-interpretarla-1/
NIST. 2013. "1.3.5.11. Measures of Skewness and Kurtosis", en *NIST/SEMATECH e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
Redacción. 2014. "El uso del Big Data y los grandes volúmenes de información, es el gran desafío para las empresas", *Puro Marketing*. http://www.puromarketing.com/30/19583/data-grandes-volumenes-datos-gran.html
Rivera L., M. 2011. *El papel de las redes bayesianas en la toma de decisiones*. http://www.urosario.edu.co/Administracion/documentos/investigacion/laboratorio/miller_2_3.pdf
Statistics, I. 2014. *Assessing Skewness and Kurtosis in the Return Distribution*. https://www.evestment.com/resources/investment-statistics-guide/assessing-skewness-and-kurtosis-in-the-return-distribution/
Unesco. 2008. *Manual de Referencia de WinIDAMS V.1.3 - Índice General*. http://www.unesco.org/webworld/portal/idams/html/spanish/TOC.htm
Unesco. 2011. Introducción a *WinIDAMS V.1.3*. http://www.unesco.org/webworld/portal/idams/html/spanish/S1intro.htm
Unesco. 2016a. *Manual de Referencia de WinIDAMS V.1.3 - Prefacio*. http://www.unesco.org/webworld/portal/idams/html/spanish/S1pref.htm#Ktw_1
Unesco. 2016b. *Manual de Refencia de WinIDAMS V.1.3 - Tablas multidimensionales y su presentación gráfica*. http://www.unesco.org/webworld/portal/idams/html/spanish/S1mtabs.htm#Ktw_0
Vitutor. 2014. *Distribución normal*. http://www.vitutor.com/pro/5/distribuci%C3%B3n_normal.html
Yat P., O. 2008. *Regresión y Correlación*. http://oscarmanuelyatpop.blogspot.com/2008/06/re

Para citar este texto:

Ruiz-Lobaina, Esther Marina y Pedro Lázaro Romero-Sánchez. 2017. "Búsqueda de patrones para mejorar productos y servicios en las bibliotecas". *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información* 72 (31): 209-225.
http://dx.doi.org/10.22201/iibi.0187358xp.2017.72.57830

DOI: http://dx.doi.org/10.22221/iibi.0187358xp.2017.72.57830