



CONSULTORA DE CIENCIAS DE LA INFORMACIÓN
BUENOS AIRES
ARGENTINA

Serie

DOCUMENTOS DE TRABAJO

Área: Tecnología

Bibliominería: bibliometría y minería de datos

Marcelo de la Puente

Marzo 2010

N°014

ISSN 1852 - 6411

Copyright Consultora de Ciencias de la Información

Editor: Patricia Allendez Sullivan. Asistente Editorial: Mariana Sabugueiro

Puente, Marcelo de la

Bibliominería: bibliometría y minería de datos. Buenos Aires:
Consultora de Ciencias de la Información, 2010.

ISSN 1852 - 6411

1. Bibliominería. 2. Bibliometría. 3. Minería de datos. 4. Minería
de datos web. I. Título

Resumen:

La Bibliominería puede describirse como una disciplina que combina técnicas de la Estadística, la Bibliometría y de la Minería de Datos para la elaboración de informes que permitan extraer nueva información o conocimiento sobre los patrones de uso de los sistemas bibliotecarios. Una de las diferencias entre la Bibliometría y la Bibliominería es que la primera estudia patrones de citación entre autores, mientras que la segunda extrae patrones de uso por parte de los usuarios de una colección. Implica una serie de pasos como determinación de áreas a trabajar, identificación de fuentes, almacenamiento de datos y extracción de patrones

La minería de datos o Data Mining es un conjunto de técnicas que consisten en la extracción no trivial de información que existe de manera implícita en grandes volúmenes de datos. Esta información, hasta el momento, era desconocida y puede resultar útil para algún proceso. Implica una preparación y una exploración de los datos para descubrir patrones de información, nuevo conocimiento, etc.

Es un conjunto de técnicas destinadas a la extracción de conocimiento procesable implícito en las bases de datos. Sus fundamentos se encuentran en la Inteligencia Artificial y en la Estadística, en el que las Ciencias de la Información confluyen en el ámbito de la Gestión del Conocimiento en las organizaciones. Dentro de lo que se denomina la cadena de valor de la Administración de Conocimiento en las organizaciones, puede ubicarse como una actividad destinada a la adquisición de nuevo conocimiento, junto con técnicas como las redes neuronales, los algoritmos genéticos, etc., son en conjunto las herramientas que sirven para descubrir patrones y aplicar conocimiento a la toma de decisiones concretas y a diferentes dominios del conocimiento.

La minería de datos se aplica normalmente a organizaciones que tienen en sus archivos grandes volúmenes de datos, para mejorar los procesos de negocios que requieren estos volúmenes de información, generalmente almacenada en forma estructurada en bases de datos. También es una técnica fundamental en el ámbito de la investigación científica como herramienta de análisis y descubrimiento de conocimiento a partir del análisis de datos experimentales. Se utiliza frecuentemente en el ámbito de los sistemas de información geográfica y en el campo de la Bioinformática, en el análisis de las estructuras de grandes moléculas biológicas, como las proteínas o el ADN.

Un proceso normal de minería de datos, implica una selección del conjunto de datos, un análisis de los mismos, una selección previa de las técnicas a aplicar, el proceso de extracción de conocimiento en si mismo y la interpretación de los resultados obtenidos.

La minería de datos aplicada a las bibliotecas, se denomina **Bibliominería**, término que deriva del inglés, *bibliomining*, como una derivación de los términos bibliometría (*bibliometrics*) y minería de datos (*data mining*). Se define como la combinación de minería de datos, Bibliometría, Estadística y herramientas de elaboración de informes y extracción de patrones de comportamiento, basados en sistemas bibliotecarios. Es un término reciente, pero se viene utilizando desde la década de 1990.

Se asocian ambas disciplinas, porque ambas se ocupan del análisis estadístico de datos con el propósito de descubrir patrones y tendencias en los datos, pero en el caso de la Bibliometría, la parte de la Cienciometría que aplica modelos estadísticos al estudio de la información científica, su objeto de estudio es la comunicación entre académicos en forma cuantitativa, a través de indicadores bibliométricos. Aquí se trata de un uso pretérito de la información, productividad de los autores en distintos campos disciplinarios a través de estudios de citas, etc.,

mientras que en el caso de la Bibliominería se trata de datos con un uso potencial, previamente a la acción a desarrollar para determinar los patrones de uso de la colección por parte de los usuarios, la extracción de patrones de comportamientos de los usuarios en el uso de los servicios bibliotecarios, con utilidad para la toma de decisiones para la selección de recursos, la organización de la colección y la planificación de los servicios por parte de los directores de unidades de información.

En el primer caso, en enfoque esta puesto en los autores y en las redes de citas que se elaboran en un campo dado, mientras que en el segundo, es el uso de la colección por parte de los usuarios

El proceso de la Bibliominería, según los diversos especialistas del área se compone de seis pasos o fases:

- Determinación de los campos temáticos de interés
- Identificación de fuentes de información internas y externas
- Recolección, depuración y proceso de ocultamiento de la identidad de usuarios en el almacén de datos del sistema o *data warehouse*
- Selección de las herramientas de análisis
- Descubrimiento de patrones, tendencias y elaboración de informes
- Análisis e implementación de los resultados

Las fuentes de información externas a la biblioteca, generalmente consisten en datos de tipo demográfico y sirven para contextualizar la información obtenida.

Las fuentes internas de información generalmente provienen de los Catálogos en línea y de los datos de circulación de materiales (en el caso de los sistemas integrales de gestión bibliotecaria), son los datos que provienen de la actividad diaria. Ambas fuentes muestran información sobre los materiales más usados, las renovaciones de los préstamos, etc.

En el caso de Bibliotecas digitales se pueden desempeñar mayor variedad de funciones, se pueden crear bitácoras que muestran la actividad de los usuarios de las mismas, mediante la identificación de la IP y con el uso de cookies y con la identificación del usuario, en el caso de las bibliotecas con acceso restringido.

El almacén de datos recopilados almacena datos actuales e históricos de potencial interés para los responsables de la toma de decisiones en una organización, toma los datos generalmente de las transacciones operativas del sistema, en el caso de las bibliotecas, en las operaciones de préstamos, reservas, devoluciones, renovaciones, consultas al catálogo, etc. La información del sistema debe ser filtrada y depurada previamente y estandarizada para facilitar su consulta. Este almacén debe permitir la elaboración de distintos informes en base a criterios específicos. Esta es la etapa que toma más tiempo de todos los pasos mencionados.

Es muy importante en el diseño del almacén de datos el objetivo de la protección de la privacidad de los usuarios, es decir no debe guardarse la información de forma que se identifique a los usuarios y se viole su privacidad.

Según Nicholson (2006), se puede considerar que existen en el mismo almacén tres tipos de datos principalmente: datos sobre la *obra* de la colección, datos sobre el *usuario* y datos sobre el *servicio*. El almacén debe contener y conectar los tres tipos de datos. En el primer caso, tenemos los datos bibliográficos propiamente dichos sobre la obra: el autor, el título de la obra, descriptores temáticos, formato, ubicación física (URL en el caso de bibliotecas digitales), etc. Esta información puede estar codificada en distintos formatos de entrada o intercambio de datos, como el MARC, Dublin Core (en el caso de los metadatos), etc., o en el sistema de gestión bibliotecaria. Esta área puede conectar la información bibliométrica, como citas o links, con otras obras. Esto requeriría, en el caso de las bibliotecas

digitales, la extracción desde la fuente original o el enlace a la base de datos referencial.

En el segundo caso, se encuentran los datos sobre el usuario en el que se almacenará lo que se denomina el *sustituto demográfico* (se verá luego): se pueden almacenar datos adicionales como IP de la computadora de acceso, que podría dar una idea sobre la localización, en el caso de las bibliotecas digitales, en bibliotecas académicas, en el caso de las bibliotecas públicas, datos del perfil del usuario, áreas de interés, etc. Todo esto podrá llegar a brindar una aproximación demográfica al usuario, pero nunca una coincidencia exacta.

En el tercer caso, el servicio bibliotecario, en dónde se encuentra la razón primaria de ser de la biblioteca, sería la parte más difícil de conceptualizar debido a la variedad de servicios que la misma provee :búsquedas, circulación, referencia, préstamo interbibliotecario y otros servicios. Deben añadirse al almacén un conjunto de campos apropiados para cada tipo de servicio. El almacén de datos debe ser capaz de manejar ambos tipos de datos: tanto los que permiten la evaluación de un servicio específico como los que brindan la posibilidad de comprender el uso que se hace de los distintos servicios de la biblioteca por parte de los usuarios.

A los datos recopilados de diversas fuentes, se le aplica el **OLAP (On Line Analytic Processing)** o procesamiento analítico de los datos en línea, que es el procesamiento de los datos en múltiples dimensiones, lo que permite visualizar los datos desde diversos puntos de vista, a través de la elaboración de informes. Se pueden efectuar consultas específicas a la base de datos y una realizar un análisis no dirigido de diversos parámetros. Se utilizan los datos provenientes del sistema de gestión integral, si la biblioteca posee un software de gestión integral.

Con los datos recopilados en el almacén de datos, se pueden efectuar distintas operaciones: En primer lugar se lleva a cabo un proceso de limpieza y de filtrado

de los datos, para descartar los datos irrelevantes y asegurar la consistencia de los datos, a continuación, se realizan diferentes tipos de operaciones que dan como resultado diferentes tipos de información: *asociaciones*, *secuencias*, *clasificaciones*, *agrupaciones* y *pronósticos*.

Las *asociaciones*, son ocurrencias relacionadas a un único evento, muestran los elementos relacionados de diferentes formas (causalidad, similitud, derivación, etc.)

En las *secuencias*, los eventos se enlazan con el paso del tiempo.

La *clasificación*, reconoce patrones que describen al grupo, al cual pertenece un elemento dado por medio del análisis de elementos existentes que se han agrupado en el mismo y por medio de un conjunto de reglas de inferencia que definen la pertenencia o no de un determinado elemento al grupo en cuestión.

El *agrupamiento* o *clustering*, funciona de manera similar a la clasificación, pero cuando aún no se han definido los grupos de pertenencia, se van formando grupos de datos con características similares.

Existe también otra herramienta denominada *generador de pronósticos*, que puede llegar a predecir tendencias futuras, en base a datos históricos y actuales

Bibliometría y Mapas SOM

El área en la que más comúnmente se han aplicado las técnicas de minería de datos al campo de las bibliotecas es el de la *Bibliometría*, que es la disciplina que estudia la comunicación entre académicos, los patrones de citación desde un

enfoque cuantitativo. Las obras de diferentes autores y las colecciones asociadas con ellos (revistas, editoriales, bibliotecas), se conectan a través de citas, términos comunes y otros aspectos del proceso de publicación.

La Bibliometría tradicional analiza información sobre la creación de una obra determinada, como la autoría y las obras citadas.. Utiliza modelos estadísticos que aportan información sobre la actividad científica, sobre la producción científica en diferentes campos y que trabajos son más citados por los autores de dicho campo y a través de dicho análisis que trabajos son considerados más relevantes en determinado campo.

Estos datos le permiten al investigador entender el contexto en el que la obra fue creada, el impacto a largo plazo de la citación de la obra y la diferencia entre los distintos campos académicos en los que emerge el patrón de citación, que muestra que trabajos están conectados con otros, cuáles son más significados en el desarrollo de un campo temático dado, etc. La Bibliometría usa generalmente indicadores basados en la frecuencia de citación de autores por otros autores, pero algunas de las nuevas técnicas utilizadas de Data Mining utilizan la visualización para explorar patrones en la elaboración de dichos estudios. La integración de las citas entre diferentes obras permite una exploración de relaciones entre académicos y temas, con mayor profundidad y los enlaces entre las distintas obras pueden ser utilizados para la recuperación automática de la información y para la visualización de las relaciones entre los diferentes autores de un campo temático determinado.

Un ejemplo de esto último son los mapas autoorganizados de Kohonen, basados en redes neuronales artificiales, métodos de procesamiento numérico en paralelo conectados entre sí en forma de grafo dirigido, basados en la Inteligencia Artificial, que modelan el comportamiento de las neuronas en el sistema nervioso humano y está relacionado estrechamente con las técnicas de Data Mining.

Los Mapas Autoorganizados se basan en el Modelo de Kohonen, informático que interesado en comprender la clasificación natural que realiza el cerebro humano, ideó el algoritmo SOM (Self Organized Map). Estos mapas organizan la información de entrada proveniente de diversas fuentes de información, de tal forma que permiten visualizar relaciones importantes entre los datos, lo que puede llevar a descubrir nuevas relaciones entre los mismos, desconocidas previamente.

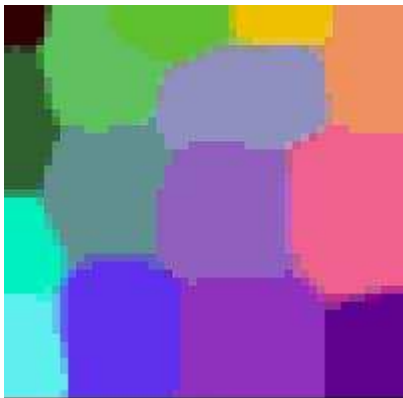
A través de los mapas, se realiza la operación de agrupamiento o *clustering*, entre los datos, es decir la clasificación de la información y su representación en mapas conceptuales de dos o más dimensiones, lo cual puede llevar al descubrimiento de información. En el *clustering* un algoritmo de la red neuronal, divide los datos de entrada en una serie de grupos con características comunes.

Los *clusters* se colocan en una red bidimensional. El concepto clave es que los representantes de cada grupo están correlacionados espacialmente, de forma tal que los puntos más próximos en la red son más similares entre sí que los que están separados. El modelo busca replicar las capacidades del cerebro humano de formar mapas topológicos a partir de los estímulos sensoriales recibidos del exterior. El modelo está compuesto básicamente de dos capas de neuronas o *nodos*: la capa de entrada y la capa de salida. La capa de entrada se encarga de recibir y transmitir a la capa de salida la información proveniente del exterior. La capa de salida es la encargada de procesar la información y de formar el mapa, estas neuronas se organizan comúnmente en una red bidimensional. La conexión entre las dos capas siempre es en una sola dirección: se propaga de la capa de entrada a la capa de salida. Cada neurona de la capa de entrada está conectada a la capa de salida mediante un “peso” específico. De esta manera, cada neurona de salida tiene asociado un vector de peso particular. El vector puede ser una descripción de las características físicas del objeto descrito o el estímulo. Cada unidad de la red se enlaza con el vector de entrada o estímulo de medio de x sinapsis de peso w . Entre estas se producen interacciones en un proceso de competencia entre las neuronas que produce la topología propia o la estructura del

mapa. Las configuraciones más comúnmente usadas son la rectangular y la hexagonal.

Los mapas SOM se emplean en la categorización automática de documentos, que se basan en la premisa de que documentos similares contienen términos similares

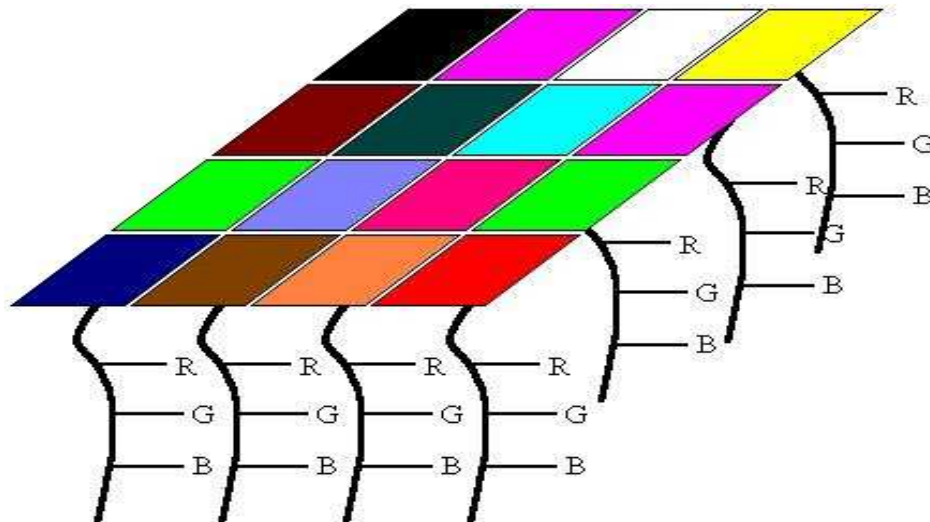
Los Mapas SOM se utilizan en la clasificación de documentos en un área temática particular, en minería de textos, por el uso del estudio de asociaciones entre términos. En dicho mapa cada documento ocupa un lugar en el espacio, según su los temas que incluye. Los temas varían a lo largo del mapa y se utiliza una escala de colores para representar la densidad de documentos según las palabras claves o términos de indización, a mayor densidad, mayor intensidad del color.



Este tipo de estudios, permiten mostrar relaciones entre áreas temáticas y publicaciones, como revistas, patentes, etc.. También, pueden observarse los desarrollos tecnológicos que ocurren en un período dado en una o más disciplinas e identificar campos emergentes. Permite visualizar además, las organizaciones que trabajan por ejemplo en un campo particular, en el desarrollo de un producto y como se relacionan entre sí en función de la estrategia de citación.

Aunque el mapa es creado por una red neuronal y no tiene en cuenta la etiqueta de clase o el tipo de cada uno de los elementos de entrada, una vez que termina el entrenamiento de la red se pueden asignar etiquetas a cada elemento del mapa, denominados *nodos*, y la red puede clasificar datos previamente desconocidos.

Pueden utilizarse también los colores para representar los vectores de peso de las neuronas: En esta representación cada color sería un dato y tenemos su localización en los ejes cartesianos xy:



Sobre los clusters desarrollados por los algoritmos en el mapa, los expertos en la temática, pueden efectuar un control terminológico sobre los términos de indización y elaborar una descripción conceptual y comentarios sobre los mapas, para descubrir áreas relacionadas, nuevas asociaciones entre distintos conocimientos, etc.

Otra de las posibles aplicaciones de las mapas SOM es la categorización automática de documentos

Un ejemplo para la creación de mapa a partir de una clasificación automática de documentos se basa en una clasificación de artículos de una revista de Astronomía: *Astronomy and Astrophysics*, en el período de 1994 a 1996, los descriptores se basan en las palabras claves de la bibliografía. Luego de una primera fase inicial de normalización, se utilizaron sólo las palabras claves que aparecen en al menos cinco artículos diferentes, para escoger las que aparecen

más frecuentemente. Se utilizó para el mapa una escala de color basada en el número de documentos x nodo, a mayor color, mayor densidad.

Para consultar el mapa el usuario parte de un mapa principal y puede ir seleccionando nodos y accediendo de esta manera a porciones más específicas a una menor escala, a medida que va seleccionando nodos. Se listan en una ventana adyacente las palabras claves más frecuentemente usadas en los documentos junto con el número correspondiente de documentos. Si la palabra clave es de interés para el usuario puede acceder a un mapa secundario más detallado en la escala cuando el nodo del mapa primario está en una región de mucha densidad de documentos. Si ese no es el caso, la lista de documentos es inmediatamente accesible. La interfase de acceso es tanto por la visualización cartográfica como por las palabras claves.

Por ejemplo si se busca en el mapa el nodo *nova*, se encuentran los siguientes descriptores: *nova*, *estrellas de neutrones*, *sistemas binarios cerrados*, *enanas blancas*, etc. Puede observarse que son términos semánticamente relacionados, ya que una nova es una estrella variable de tipo cataclísmico que explota y que se originan comúnmente en sistemas binarios cerrados, en los que uno de sus componentes es una estrella enana blanca que absorbe por su intensa gravedad material de la estrella compañera, hasta que alcanza una masa crítica y explota. Las estrellas de neutrones son formadas en explosiones de supernovas, otro tipo de estrella explosiva, por lo que puede observarse una correspondencia entre la visualización en el mapa y la lista de palabras claves correspondientes al nodo *nova*:

Node Novae: main map
stars:novae,cataclysmic variables
accretion,accretion disks
X-rays:stars

stars:binaries:eclipsing
stars:magnetic _elds
stars:binaries:close
stars:white dwarf

Node Novae: secondary map
stars:novae,cataclysmic variables
accretion,accretion disks
X-rays:stars
stars:atmospheres
stars:binaries:general

En el mapa principal se recuperan 91 documentos, demasiados para leer el abstract de cada artículo, por eso se accede al nodo para obtener una visión más detallada. Los nodos relacionados son: *discos de acreción, estrellas de rayos X, estrellas binarias eclipsantes, enanas blancas*, entre otros.

Al acceder al nodo *nova* y al entrar en un mapa secundario, se obtiene una distinción más precisa entre los documentos que tienen como temas *estrellas binarias en general, binarias cerradas y novas*. Las palabras claves son más limitadas pero más precisas. En esta segunda aproximación se recuperan 38 artículos.

En conclusión, en general puede observarse que las relaciones presentadas entre las palabras claves vecinas en el mapa se validan semánticamente en muchos casos. Este tipo de mapas pueden servir también, para ver los cambios que ocurren en la literatura a lo largo de un período de tiempo, que se verán reflejados en la literatura sobre el tema. A partir del mapa se pueden seleccionar documentos relevantes en función de una temática de interés particular



Astronomy & Astrophysics (1994 - 1997)

[CDS](#) · [Simbad](#) · [VizieR](#) · [Catalogues](#) · [Nomenclature](#) · [Biblio](#) ·
[StarPages](#) · [AstroWeb](#)



[Keyword query](#) · [Help](#)

Maintained by **F. POINÇOT**.
Please use the following e-mail address if you want to leave a message:
E-mail: question@simbad.u-strasbg.fr

This map gives access to 4883 articles from *Astronomy & Astrophysics* (1994 - 1997). We used the technique of "Self Organizing Maps", where documents are classified in areas on the basis of their keywords. Density of documents is smallest in blue regions, and largest in red regions.

In order to consult the *document map*, and to retrieve the abstracts of relevant papers, please follow the following steps:

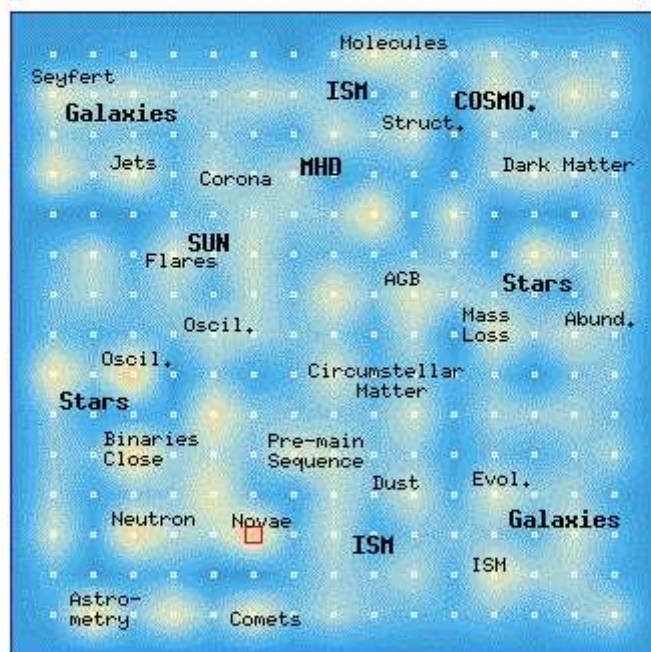
- Select, with the mouse, a region (symbolised by a white point). Keywords related to the documents of the region will appear in the adjacent frame.
- Selecting a keyword provides access to the full list of related papers.
- If there are more than 30 articles on the selected region, you can access to a more detailed map.
- To submit a keyword query, just type a few letters of an expression you are looking for. All the keywords containing these letters will appear allowing to select one or more of these keywords.

Location: <http://simbad.u-strasbg.fr/A+A/map.pl?&-prov=185>



Astronomy & Astrophysics (1994 - 1997)

[CDS](#) · [Simbad](#) · [VizieR](#) · [Catalogues](#) · [Nomenclature](#) · [Biblio](#) ·
[StarPages](#) · [AstroWeb](#)



[Keyword query](#) · [Help](#)

Maintained by **D. DOINCOT**.
Please use the following e-mail address if you want to leave a message:
E-mail: question@simbad.u-strasbg.fr

Principal map:

Node 185: 110 documents

[Get Documents](#)


Keywords:

stars:novae,cataclysmic variables 91/110
accretion,accretion disks 59/110
X-rays:stars 31/110
stars:binaries:eclipsing 13/110
stars:magnetic fields 12/110
stars:binaries:close 10/110
stars:white dwarfs 10/110

[Construct a local map](#)

Netscape: CDS Document Map

Location: http://simbad.u-strasbg.fr/A+A/map_sec.pl?&-unit=185



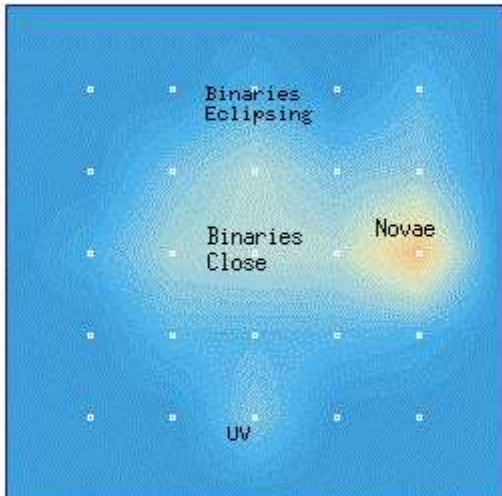
**Astronomy &
Astrophysics
(1994 - 1997)**

[CDS](#) · [Simbad](#) · [VizieR](#) · [Catalogues](#) · [Nomenclature](#) · [Biblio](#) ·
[StarPages](#) · [AstroWeb](#)

Local map 185:
110 documents

node 14: 38 documents

[Get Documents](#)



[Principal map](#) · [Keyword query](#)
[Help](#)

Keywords:

stars: novae, cataclysmic variables 25/25

accretion, accretion disks 16/25

X-rays: stars 8/25

stars: atmospheres 1/25

stars: binaries: general 1/25

Maintained by P. DOINCOT.
Please use the following e-mail address if you want to leave a message:
E-mail: question@simbad.u-strasbg.fr

Áreas de aplicación de la Bibliominería

La minería de datos aplicada a las bibliotecas debe ser capaz de predecir las necesidades de los usuarios de las mismas, en función de la evolución a través del tiempo de los temas consultados, de los materiales, autores, etc. de los patrones de uso de la colección, identificar los materiales poco consultados, por escaso interés de los usuarios en los mismos, errores en la catalogación, etc. ; debe ayudar al director de la biblioteca como soporte para la toma de decisiones estratégicas y brindar a los usuarios un servicio más personalizado al localizar información de potencial interés para los mismos, como en el caso de los servicios de DSI (Diseminación Selectiva de Información), en los que se les brinda a los usuarios información de alerta sobre nuevas publicaciones en base a un perfil de interés.

El estudio del comportamiento de los usuarios respecto al uso de la colección, resulta de gran utilidad para la *selección y adquisición bibliográficas*, conocer que materiales se utilizan más y cuáles cuentan con un número limitado en cantidad de ejemplares, resulta fundamental para planear las adquisiciones futuras y para la asignación del presupuesto a las mismas. Se pueden buscar correlaciones entre documentos de escaso uso, con proveedores, fechas, formatos, etc. y pueden descubrirse áreas de la colección que presentan problemas y en las que se debe modificar la política de selección. Pueden utilizarse también para estudiar problemas como la selección de proveedores, en caso de detectarse por correlaciones de precios, que los utilizados normalmente son costosos.

En el almacén de datos pueden agregarse campos que incluyan solicitudes o requerimientos de materiales, proveedores, tiempo de procesamiento del material, tiempo de entrega, costos, etc. Las técnicas de minería de datos pueden mostrar las relaciones complejas entre las diferentes variables y elaborar por *clustering*, por tipos de material clasificados por el tiempo promedio de procesamiento, etc., se pueden usar diferentes criterios para así lograr una mejora en la política de selección de la biblioteca.

En el caso de la *organización de la colección, en la catalogación y en la clasificación del material*, es importante aplicar al catálogo, teniendo en cuenta el nuevo modelo FRBR (Requerimientos Funcionales para los Registros Bibliográficos) en el que se basarán las nuevas normas de catalogación RDA (que reemplazan a las AACR), lograr el objetivo de que las técnicas de Data Mining ayuden a identificar las diferentes *manifestaciones e ítems*, procedentes de una misma *obra* que posee una biblioteca, por ejemplo en el caso de las grandes obras literarias, relacionarlas con adaptaciones posteriores, versiones cinematográficas, etc., poder recuperarlas juntas para los usuarios interesados en todo lo relacionado con esta obra en particular.

En el caso del sistema de clasificación y de lenguajes de indización utilizados por el sistema, con las técnicas de minería de datos puede verificarse si los usuarios utilizan o no el lenguaje documental del sistema para clarificar las búsquedas en el Catálogo en línea, cuál es la frecuencia de uso de los términos de indización, para luego adecuarlos mejor al comportamiento de búsqueda del usuario, puede descubrir asociaciones entre términos de indización usados por los usuarios para efectuar las búsquedas y lograr de esta forma, un mejor acceso a la colección.

También se han efectuado estudios en los que se combinaron el lenguaje natural y la minería de datos textual (aplicada a textos), para descubrir potenciales asociaciones desconocidas entre los documentos de colecciones digitales.

En el caso de los servicios bibliotecarios, en las búsquedas, puede evaluarse el contenido de la misma (si fue o no exitosa), el camino o patrón de búsqueda, la resolución de la misma y la concurrencia con el uso de los materiales. Con estos datos se podría, por ejemplo, efectuar una mejora al lenguaje documental usado por la institución, con la adición de sinónimos a los términos correspondientes, que deriven de los términos del lenguaje natural más utilizados por los usuarios en el

proceso de búsqueda según los patrones que van emergiendo, de forma similar a la forma implementada por el motor de búsqueda de Google.

Pueden establecer patrones de búsqueda en el uso del OPAC, según las clases de consultas efectuadas, se pueden recomendar términos relevantes según el tipo de búsquedas realizadas y también ítems relevantes según los registros de búsquedas anteriores y expandir las asociaciones entre términos y documentos como una forma de ayuda o guía para el usuario. También se pueden proponer correcciones si se detectan errores ortográficos sistemáticos en determinados términos o en búsquedas específicas.

En el caso de la circulación, pueden extraerse datos como análisis de los registros o bitácoras de transacciones, datos del mostrador, etc. Con los datos de tiempo de transacciones diarias, por ejemplo, se puede intentar inferir patrones de comportamiento en el área de circulación. En base a los patrones encontrados, puede efectuarse un cambio en la asignación del personal de la biblioteca a circulación, si se encuentra que el número actual asignado es insuficiente. Combinando los datos de adquisiciones y circulación, por ejemplo, pueden examinarse títulos de alto o de bajo uso y consecuentemente, se realizarán modificaciones en la política de adquisición y selección de material, aumentando el número de pedidos en el primer caso y disminuyéndolo en el segundo.

En el servicio de referencia pueden evaluarse el método, las preguntas de referencia, las preguntas temáticas, el tiempo, las respuestas, el o camino seguido en el proceso de referencia, etc. Se ha planteado también, el uso de la minería de datos web para apoyar a los servicios de referencia electrónica. La minería web es precisamente, el uso de técnicas de minería de datos para descubrir y extraer información desde Internet, a través del descubrimiento de recursos web relevantes, de la extracción de información de una variedad de documentos web en diferentes formatos y a través del descubrimiento de patrones mediante la aplicación de la técnica de *clustering* a recursos web.

La minería web debe analizar documentos en una variedad de formatos, textual, audiovisual, etc., puede tratar de descubrir los patrones de uso o de acceso a los recursos web, tanto de una forma general como más personalizada, para ir delimitando el perfil de los usuarios de los mismos. Una conexión con la bibliometría, puede encontrarse en la llamada “minería web de estructura”, que analiza los patrones potenciales que pueden hallarse en el estudio de los hipervínculos de los documentos hipertextuales, esta técnica es similar a los estudios de citas bibliométricos y permiten estudiar la estructura de los enlaces entre documentos, permite descubrir: autoridades que proporcionan la mejor fuente sobre un tema determinado.

Pueden incluirse datos sobre programas de entrenamiento a usuarios: tipos de programa, títulos de programas, programas temáticos, etc.

En los estudios de usuarios, la Bibliominería puede ayudar a conformar la formación de grupos de usuarios con características similares, para llegar al descubrimiento de patrones de búsqueda más generales y poder adaptar así, la colección de la biblioteca a las necesidades de los usuarios que emergen de estos patrones.

En base al análisis de los registros de circulación de materiales, pueden efectuarse predicciones sobre que tipos de materiales solicitarán un grupo de usuarios, en función de sus características demográficas (profesión, en el caso de bibliotecas universitarias o académicas, carrera, materias en las que está inscripto, etc.).

También puede estudiarse a través de las técnicas de bibliominería, el comportamiento del personal de la biblioteca al analizar el uso dado por estos a las bases de datos del sistema, lo cual puede utilizarse para optimizar el desempeño del mismo en las operaciones diarias de la biblioteca.

Otro de los puntos fundamentales sobre los datos que se obtienen de la minería de datos aplicada a las bibliotecas es el dilema ético del respeto a la privacidad del usuario, es problema de preservar el anonimato o la confidencialidad de la información recabada sobre los usuarios, materiales consultados, patrones de búsqueda, etc. Nicholson, considera que una posible solución a este dilema podría ser la de desarrollar políticas de aislamiento de los datos, de encriptación de los mismos para preservar el anonimato del usuario, a esto lo denomina "*sustitutos demográficos*", es decir, se almacenarían solo los datos del perfil del usuario que permitan clasificarlo en un grupo determinado, pero sin identificarlo personalmente. Se toman un conjunto de campos para representar al usuario en las técnicas de bibliominería, de la misma forma que el registro bibliográfico es un conjunto de campos que representan una obra, un documento, etc. Ese conjunto de campos reemplazan a la información de tipo personal en el almacén de datos, es decir, desde el punto de vista de la bibliominería no importa el nombre y apellido del usuario particular al que se le prestó un título específico una x cantidad de veces durante el año, sino saber que ese título particular fue prestado x cantidad de veces durante ese año a un usuario que se identificará con un código numérico en el sistema y que responde a un perfil determinado (estudiante de una carrera específica, profesor, en el caso de bibliotecas universitarias, etc.)

Esto es importante, ya que muchas decisiones tomadas en la gestión bibliotecaria, involucra más grupos de usuarios, sus perfiles, comportamiento, necesidades, etc., antes que a usuarios individuales. Las actividades realizadas por estos grupos de usuarios emergen en patrones que las técnicas de Bibliominería pueden descubrir y analizar. Estas comunidades de usuarios pueden ser representadas por variables demográficas, por ejemplo, en el caso de las bibliotecas académicas, departamentos, localización (en el campus o fuera de él), código postal, grupos de pertenencia, áreas temáticas de interés, departamentos o cargos en bibliotecas corporativas, etc.

De esta manera el director de la biblioteca puede obtener información de cómo las comunidades usan los recursos de la biblioteca sin sacrificar la privacidad individual de los usuarios.

No obstante, todo el problema de la privacidad de los datos plantea diversos problemas de índole ética y legal que por su complejidad, excede el presente análisis, pero con una buena política de protección de los mismos y con el consentimiento informado del usuario, es factible que puedan ser de utilidad para la mejora en el servicio de la biblioteca, en la gestión bibliotecaria y en la toma de decisiones por parte de los directivos de la institución.

De todas formas, se presentan algunos problemas en el desarrollo de estos sustitutos que se deberán resolver, como por ejemplo, cuanta información puede llegar a perderse en el proceso de desidentificación de los datos, normalización de tipos de datos desidentificados a incluir en el almacén de datos por redes bibliotecarias, la opinión de los usuarios acerca del procedimiento, etc.

Conclusiones

Puede decirse que los informes que resultan de la aplicación de las técnicas de Bibliominería le permiten al director de la biblioteca o centro de documentación una justificación de las actividades realizadas ante la autoridad mayor a la que sirve la biblioteca. Puede fundamentarse con datos estadísticos los programas y servicios desarrollados por la biblioteca, los resultados alcanzados y el presupuesto asignado a la misma.

En este trabajo hemos seguido la concepción de Nicholson que concibe a la Bibliomiería como una combinación de minería de datos, bibliometría y herramientas generadoras de informes para extraer modelos de comportamiento de los sistemas automatizados de bibliotecas, en el que se establece tanto la conexión entre ambas disciplinas a través de un campo en común: las obras de la

colección, impresa o digital. Al combinar los dos tipos de datos, tanto el aspecto de la creación de la obra, los autores y por otra parte, el acceso a las mismas se llega a un único almacén de datos con los que se pueden efectuar la totalidad de las operaciones descritas en el modelo. Esto se logra conectando las obras elaboradas por autores con la población de usuarios de la institución.

Aunque Nicholson reconoce que este modelo rara vez se aplica en forma completa, puede servir para el desarrollo futuro de almacenes de datos para bibliotecas que soporten un análisis completo de técnicas de Bibliominería y que permita tanto a directores de bibliotecas como a investigadores tener una mayor idea tanto de los recursos, impresos o digitales contenidos en la institución, cómo de la forma de acceso a los mismo por parte de los usuarios. No obstante debe comprenderse que los dos grupos a los que pueden servir las técnicas de bibliominería, por un lado los directores de bibliotecas y por el otro los investigadores en el área de la Bibliotecología, tienen metas diferentes: los administradores pretenden comprender el uso de su propia biblioteca mientras que los investigadores quieren generalizar sus hallazgos a una población mayor.

La Bibliometría ayuda a ambos grupos de diferentes formas: por un lado, para los directores de unidades de información les provee una comprensión más profunda y detallada del uso de los servicios de su institución, mientras que los investigadores pueden crear almacenes de datos que incluyan datos de más de una institución para obtener una visión más general y amplia.

Los administradores de biblioteca deben comprender que las técnicas de Bibliominería por sí solas no logran determinar si la colección está cumpliendo con la misión de la biblioteca o con los objetivos establecidos por la política de la institución. Deben aplicarse en conjunción con otras técnicas más tradicionales de los estudios de usuarios, como las encuestas, cuestionarios, etc. que brinden información sobre la relevancia del material hallado en búsquedas para el usuario o sobre la dificultad que encuentra al usar el sistema, ya que el almacén de datos

sobre el que se aplican las técnicas de minería de datos no muestra esta información solo registra lo que el usuario seleccionó pero no cuán relevante es para su búsqueda, si satisface o no sus necesidades de información. Se deben tener múltiples perspectivas para poder realizar una evaluación integral, tanto una interna (desde el punto de vista de los bibliotecarios y de los investigadores), como una externa (punto de vista del usuario) y debe tomarse en cuenta tanto el contenido de la sistema (medidas en base a políticas, estándares, relevancia, pertinencia, usabilidad) así como el uso del mismo (bibliominería aplicada al uso de materiales, estudios bibliométricos, estudios de citas de usuarios al material en trabajos académicos., etc.)

Bibliografía

Candás Romero, J. (2006). Minería de datos en bibliotecas: bibliominería. Textos universitaris de Biblioteconomia i documentació, 17. En: <http://www.ub.es/bid/17canda2.htm>, Acceso, 30 de marzo, 2010

Germano, Tom (1999). Self-Organizing Maps. En: <http://davis.wpi.edu/~matt/courses/soms/>, Acceso, 30 de marzo, 2010

Herrera Varela, R. (2006). Bibliomining: minería de datos y descubrimiento de conocimiento en bases de datos aplicados al ámbito bibliotecario. En: http://www.bibliotecarios.cl/Conf2006/C2006_019.pdf, Acceso, 30 de marzo, 2010

Laudon, K. (2008), J. Laudon. Sistemas de información gerencial. México: Pearson Education. 736 p.

Los mapas autoorganizados de Kohonen. En: <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema5dm.pdf>, Acceso, 30 de marzo, 2010.

Nicholson, S. (2003) The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries* 22 (4) En: <http://www.bibliomining.com/nicholson/biblioprocess.htm>, Acceso: 30 de marzo, 2010

Nicholson, S. (2006). [Preprint of The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services.](#) *Information Processing and Management* 42(3), 785-804. En: <http://bibliomining.com/nicholson/nicholsonpdfs/nicholsonbibliointro.pdf>, Acceso: 30,marzo, 2010

Poincot, P. (1998), S.Lesteven y F. Murtagh. A spatial user interface to the astronomical literature. *Astronomy & Astrophysics Supplement Series*, 130, 183-191. En: <http://aas.aanda.org/index.php?option=article&access=standard&Itemid=129&url=/articles/aas/pdf/1998/10/ds1464.pdf>, Acceso, 1 de abril, 2010

Witten, I. (2005), E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd.ed. Boston: Elsevier. 524 p.