

Bibliomining: Aplicación de text mining para descubrir preferencias de usuarios en el Centro de Documentación Regional “Juan Bautista Vázquez”

Wilson Pérez, Valeria Haro, Víctor Saquicela

Departamento de Ciencias de la Computación, Universidad de Cuenca, Av. 12 de Abril y Av. Loja Cuenca, Ecuador, 010150.

Autores para correspondencia: valeria.harov@ucuenca.ec, {wilson.perez, victor.saquicela}@ucuenca.edu.ec

Fecha de recepción: 17 de mayo 2017 - Fecha de aceptación: 8 de agosto 2017

ABSTRACT

Nowadays, the management of huge volumes of information in support of administrative decisions is one of the challenges libraries face. An aid thereby are text mining techniques enabling the extraction of hidden knowledge. The article describes the implementation of Bibliomining and Text Mining in the “Juan Bautista Vázquez” library of the University of Cuenca. The purpose of this study is to determine the behavior of library users by discovering patterns of similarity between the titles they seek and the titles of the bibliographic material they borrow. In this case study, we used the methodology proposed by Nicholson and the algorithms and libraries of the tool R, to create word clouds generated from search terms and book titles; demonstrating that the used visualization facilitates the perception and the interpretation of the obtained results.

Keywords: Data mining, text mining, bibliomining, library, words cloud.

RESUMEN

Hoy en día, la gestión de grandes volúmenes de información para dar soporte a las decisiones administrativas es uno de los retos que enfrentan las bibliotecas. Las técnicas de Text Mining pueden utilizarse para extraer conocimiento oculto de grandes volúmenes de datos. Este artículo describe la implementación de técnicas de Bibliomining y Text Mining en el Centro de Documentación Regional “Juan Bautista Vázquez” de la Universidad de Cuenca. El objetivo de este estudio es determinar el comportamiento de los usuarios de la biblioteca al descubrir patrones de similitud entre los títulos que buscan y los títulos del material bibliográfico que toman prestado. En este caso de estudio, se usa la metodología propuesta por Nicholson y, los algoritmos y librerías de la herramienta R, para crear nubes de palabras generadas a partir de términos de búsquedas y títulos de libros; demostrando que la visualización utilizada facilita la percepción e interpretación de los resultados obtenidos.

Palabras clave: Minería de datos, minería de texto, bibliomining, biblioteca, nube de palabras.

1. INTRODUCCIÓN

El Centro de Documentación Regional “Juan Bautista Vázquez” con la finalidad de ofrecer un mejor servicio a los usuarios, está en constante desarrollo tecnológico y de infraestructura física. Al pertenecer a un Instituto de Educación Superior, esta debe actualizar sus colecciones frecuentemente al ritmo con que evolucionan las diferentes áreas del conocimiento, permitiendo que los usuarios estén siempre a la vanguardia en las diferentes temáticas de los perfiles profesionales. Para cumplir con este objetivo, es indispensable conocer el comportamiento de los usuarios de las bibliotecas al aplicar distintas técnicas de Text Mining y Data Mining, generando reportes que permitan una gestión más eficiente de los

recursos bibliográficos según las necesidades de los usuarios y una mejora continua en la calidad de los servicios ofrecidos por la biblioteca.

Text Mining (TM) es la extracción no trivial de información oculta, previamente desconocida y potencialmente útil del texto del lenguaje natural (Waegel, 2006). TM es similar al Data Mining (DM), excepto que las herramientas del DM están diseñadas para manejar datos estructurados, pero el TM también hace uso de conjuntos de datos no estructurados o semi-estructurados (Ramez & Navathe, 2000). En el dominio de las bibliotecas, existe el concepto de Bibliomining, que es una combinación de la bibliometría y el DM (Nicholson & Stanton, 2003 & 2005), definido como la aplicación de técnicas de DM en el análisis de la información generada por los centros de documentación para la extracción del conocimiento, ayudando a tomar decisiones del dominio bibliotecario. Las técnicas usadas en Bibliomining pueden ser categorizadas en asociación, clasificación, agrupación y regresión (Siguenza-Guzman, Saquicela, Avila-Ordóñez, Vandewalle & Cattrysse, 2015).

Este estudio toma las huellas digitales que deja el usuario tanto en búsquedas de recursos como en los diferentes préstamos de libros físicos que solicita, para analizar las preferencias en los recursos bibliográficos y el comportamiento del usuario en esos procesos. El artículo está organizado de la siguiente manera: en la sección 2 se realiza una revisión bibliográfica de trabajos relacionados; en la sección 3 se describe la metodología aplicada que empieza con la extracción y preparación de datos para posteriormente aplicar técnicas de Text Mining y descubrir conocimiento; en la sección 4 se puede observar los resultados obtenidos. Finalmente, en la sección 5 se sintetiza las conclusiones y los trabajos futuros.

2. LITERATURA RELACIONADA

En el trabajo presentado por Siguenza-Guzman, Saquicela & Cattrysse (2014), los autores utilizan una metodología para la evaluación integral de bibliotecas, donde evalúan la colección (fuentes externas e internas) y los servicios para la gestión de bibliotecas, tales como: el desempeño de los servicios, el control de calidad, el uso de la colección y la interacción con el usuario. En base a dicho análisis holístico, Haro, Pérez, Siguenza-Guzman, Cattrysse & Saquicela (2014), diseñan e implementan un sistema de soporte de decisiones basado en técnicas de Data Warehouse (DW) para la biblioteca de la Universidad de Cuenca y propone una arquitectura de DW que integra, procesa y almacena los datos.

Con el objetivo de determinar preferencias de los usuarios, los trabajos presentados en (Azam, Sohrawardi, Das, Alam, Alvy *et al.*, 2013; Uppal & Chindwani, 2013; Riveros & Klenzi, 2015; Haro, Pérez & Saquicela, 2016) utilizan técnicas de Bibliomining sobre la información de préstamos que poseen sus respectivos centros de documentación universitaria, definiendo reglas de asociación de los recursos bibliográficos y la relación entre categorías de clasificación.

Ligada al TM, los autores Funk (2013), Ahearn (2014) y Baeg (2016) identifican los principales temas a lo largo del tiempo de diversas revistas en el campo de Ciencias y la Biblioteconomía; el análisis se enfoca en representar el cuerpo literario y los principales conceptos en nubes de palabras, con el objetivo de determinar cómo ellas reflejan los temas de los artículos e identificar las características de la revista. Por otro lado, basado en los registros de préstamos de libros, pedidos en el año 2015 en la Universidad de Jinan, Cong (2017) explora la preferencia de los lectores para guiar el proceso de adquisición de libros en la biblioteca; en el estudio, guiados de técnicas de TM se analiza la frecuencia de los títulos de los libros y los resultados son presentados en nubes de palabras agrupados en la Clasificación China de Bibliotecas; para ampliar este trabajo hemos considerado no solo los datos de los títulos prestados, sino también los términos de las búsquedas realizadas por el usuario para hacer una comparativa entre los datos de estas fuentes y además, un análisis de correlación entre los términos más populares de las búsquedas.

3. METODOLOGÍA DE INVESTIGACIÓN

El caso de aplicación se desarrolla en el Centro de documentación “Juan Bautista Vázquez” (CDRJBV) perteneciente a la Universidad de Cuenca, Ecuador. El CDRJBV está conformado por las Bibliotecas de los Campus Central, El Paraíso (Áreas de la Salud) y Yanuncay (Áreas Agropecuarias y de Artes). Su conjunto documental está constituido por publicaciones convencionales en todas sus formas, así como por diversos soportes digitalizados, audiovisuales y bases de datos en línea. El CDRJBV ofrece servicios de reservas de libros, préstamos a usuarios, préstamos inter-bibliotecarios, evaluación de bibliotecarios y encuestas de satisfacción. Los usuarios finales del CDRJBV son estudiantes, docentes y personal administrativo de la Universidad de Cuenca, a quienes se ofrece el servicio de préstamos y reservas de libros (Haro *et al.*, 2014).

La Universidad de Cuenca cuenta con un total de 12 Facultades, distribuidas en los diferentes campus de estudio. La Tabla 1 presenta las facultades con mayor interacción en el CDRJBV para el proceso de préstamos (Haro *et al.*, 2016).

Tabla 1. Las 5 Facultades con mayor interacción en el CDRJBV.

Facultad	Porcentaje (%)
Ciencias Médicas	21.30
Ciencias Económicas y Administrativas	17.42
Filosofía, Letras y Ciencias de la Educación	12.47
Facultad de Ingeniería	10.39
Arquitectura y Urbanismo	9.57
Otros	28.85

Para descubrir conocimiento en el CDRJBV se aplican los procesos definidos por Nicholson & Stanton (2003), presentados en la Figura 1. El proceso inicia al determinar el área de interés e identificación de las fuentes de datos hasta llegar a obtener conocimiento que permita identificar a los usuarios del CDRJBV y ayudar en la toma de decisiones.

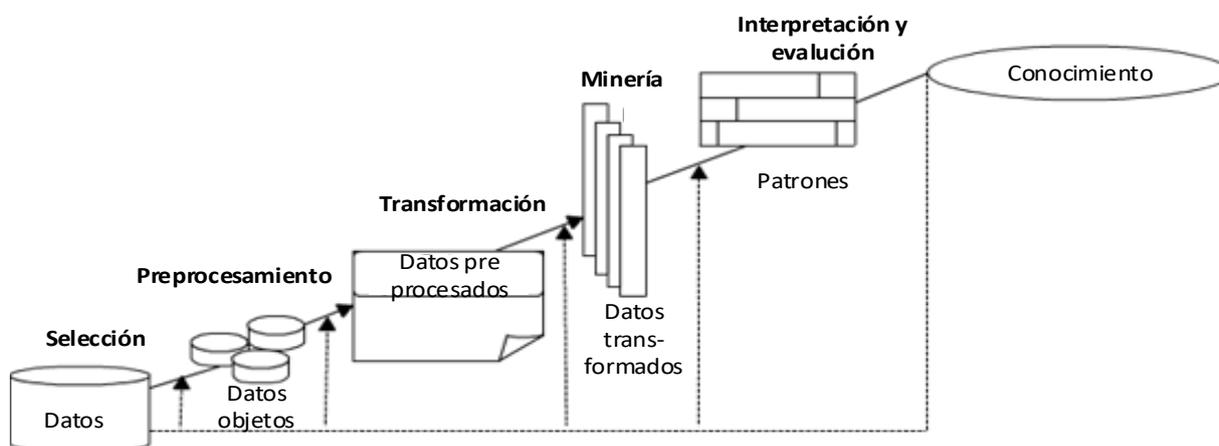


Figura 1. Metodología para descubrir conocimiento (Nicholson & Stanton, 2003).

3.1. Aplicación del proceso de bibliomining

En esta subsección se aplica el proceso de Bibliomining presentada en (Nicholson, 2003), este proceso define pasos secuenciales tales como:

- *Determinar el área de interés.* Con el propósito de analizar las tendencias del uso y disponibilidad de recursos bibliográficos en el CDRJBV, este estudio se enfoca en los procesos de préstamos y búsqueda de material bibliográfico que ofrece el CDRJBV a los usuarios finales.

- *Identificar las fuentes de datos.* En el trabajo presentado en (Haro *et al.*, 2014), se describe las fuentes de información que relacionados a los procesos de préstamos con los atributos de: fecha de préstamo, fecha de devolución, campus, categoría, autores, título, bibliotecario, usuario, facultad y carrera.
- Adicionalmente, los datos utilizados en el proceso de búsqueda se extraen del sistema *Koha*¹; dicho sistema almacena la información en la tabla de log's de búsquedas que realizan los usuarios en el CDRJBV.
- *Creación de un Data Warehouse.* El CDRJBV dispone de un DW diseñado bajo las necesidades del centro de documentación (Haro *et al.*, 2014). En dicho trabajo, los autores realizan diferentes procesos de limpieza, la integración de las diferentes fuentes y posteriormente la información es almacenada en un DW.
- *Seleccionar herramientas de análisis apropiadas.* En el estudio se utilizó el lenguaje de programación R² para el análisis de las tendencias de uso de recursos bibliográficos, debido a que R ofrece un completo banco de librerías que ayudan en el análisis estadístico y la visualización de la información.
- *Descubrir patrones de comportamiento.* En esta fase se aplica las técnicas de TM sobre los datos con el objetivo de descubrir conocimiento, este proceso es presentado en la sección 3.2 *Aplicación de Text Mining*.
- *Analizar e implementar los resultados.* Una vez descubiertos los patrones de comportamiento, el siguiente paso es analizar y validar que dichos patrones sean correctos para demostrar que los resultados obtenidos sean los apropiados, dicha fase es presentada en la sección 4.

3.2. Aplicación de text mining

El proceso de TM define un conjunto pasos para el descubrimiento de conocimiento, los mismos que son descritos a continuación:

- *Limpieza, transformación y carga.* Para el proceso de préstamos, los títulos de los recursos bibliográficos fueron exportados desde el Data Warehouse a un archivo de texto. De modo similar, para el proceso de búsquedas, las frases (texto plano) que consultan los usuarios finales fueron exportadas a un archivo de texto desde el sistema *Koha*.
- *El proceso de análisis inicia con la lectura* de los archivos de texto tanto de préstamos como el de búsquedas y posteriormente con la librería *tm_map()* de R se procedió a la limpieza del contenido, considerando cuatro casos: *i)* todos los caracteres de texto se unificaron a minúsculas, *ii)* los caracteres especiales se reemplazaron por espacio en blanco, *iii)* se removieron los caracteres numéricos y de puntuación, y *iv)* se eliminó las palabras comunes del idioma (como conectores y artículos) tanto del idioma español como del inglés, este debido a que el CDRJBV también consta con bibliografía en inglés.
- *Matriz de término del documento.* Usando la función *TermDocumentMatrix()*, se obtiene la matriz que contiene las frecuencias de cada palabra en el documento.
- *Visualización de datos.* Los resultados son visualizados en nubes de palabras; donde, las palabras son presentadas en diferentes colores y tamaños. Los términos de mayor tamaño poseen un mayor valor de frecuencia y aquellas que poseen colores similares son aquellas con frecuencias similares. Los resultados del proceso se analizan en la sección 4 de este trabajo.

4. RESULTADOS Y DISCUSIÓN

En esta sección se procede con el análisis y evaluación del proceso de Text Mining aplicado sobre las fuentes de datos de préstamos y búsqueda de material bibliográfico en el CDRJBV; en dicho análisis se

¹ *Koha* es un sistema integrado de gestión de bibliotecas, implementado sobre MySQL como gestor de base de datos. <https://koha-community.org/manual/16.11/html/>

² R es una herramienta de software libre para el análisis computacional estadístico y gráfico. <https://www.r-project.org/>

evalúa la **frecuencia** de los términos y las **asociaciones** entre ellos. Adicionalmente, se hace uso de la medida estadística **soporte** para estudiar la presencia de un término en el documento.

Las relaciones de asociación permiten descubrir ocurrencias entre elementos que pertenecen a un conjunto de datos, para lo cual se considera todas las posibles combinaciones de atributo-valor de todos los datos almacenados en un conjunto (Chen, Han & Yul., 1996). La regla $X \rightarrow Y$ tiene soporte s en el conjunto de transacciones T , $0 \leq s \leq 1$, si el porcentaje de s de las transacciones de T contienen tanto a Y como a X . El soporte puede ser considerado como la probabilidad de que las transacciones contengan un conjunto de ítems $soporte(X) = P(X)$ (Brown, 2014).

4.1. Proceso de Búsquedas

Basada en la matriz de términos, se procedió a la creación de la nube de las 50 palabras con la frecuencia más alta, presentada en la Figura 2. Los cinco términos más buscados son presentados en la Tabla 2 con su respectivo valor de frecuencia, el soporte y las asociaciones con otros términos del documento. Es así que, de cada 100 búsquedas, 1 de ellas contienen el término *biología* o *historia* o *salud* o *anatomía* o *ecuador*. Considerando el término *historia*, se puede concluir que tiene un valor de frecuencia equivalente a 650 y los cinco términos con los que más aparece junto en las búsquedas son: *consulado*, *continuación*, *imperio*, *revolución*, y *resumen*, con un soporte de 24%, 22%, 20%, y 13% respectivamente.



Figura 2. Nube de palabras. (a) Proceso de búsquedas, (b) proceso de préstamos.

Tabla 2. Términos más comunes de material bibliográfico buscados por usuarios.

Término	Frecuencia	Soporte	Asociaciones
Biología	669	1.37%	celular (0.25), robertis (0.21), molecular (0.2), práctica (0.1), universidad (0.08).
Historia	650	1.33%	consulado (0.24), continuación (0.22), imperio (0.20), revolución (0.13), resumen (0.13).
Salud	553	1.13%	Sigsig (0.19), énfasis (0.18), reproductiva (0.18), pública (0.17), locales (0.16).
Anatomía	516	1.06%	orientación (0.64), clínica (0.43), wheeler (0.39), cuello (0.20), humana (0.14).
Ecuador	500	1.03%	precolombino (0.22), república (0.18), secreto (0.17), constitución (0.15), ecuatorianismos (0.13).

El promedio de frecuencias entre las 20 palabras más buscadas es equivalente a 414, con un soporte de 0.016. Basados en estos valores, se generó el grafo de relación entre los 20 términos más populares en las búsquedas y que posean el umbral mínimo de 0.016 de correlación (ver Figura 3). El gráfico

presenta la relación que tienen los términos más comunes entre sí, de este modo, el término *historia*, es más buscado con los términos *ecuador*, *cuena*, *arquitectura* y *américa*. Así como también el término *química* se busca comúnmente con el término *general* o *introducción*, lo cual indica que los usuarios mayoritariamente prefieren las colecciones de conceptos generales y básicos de química.

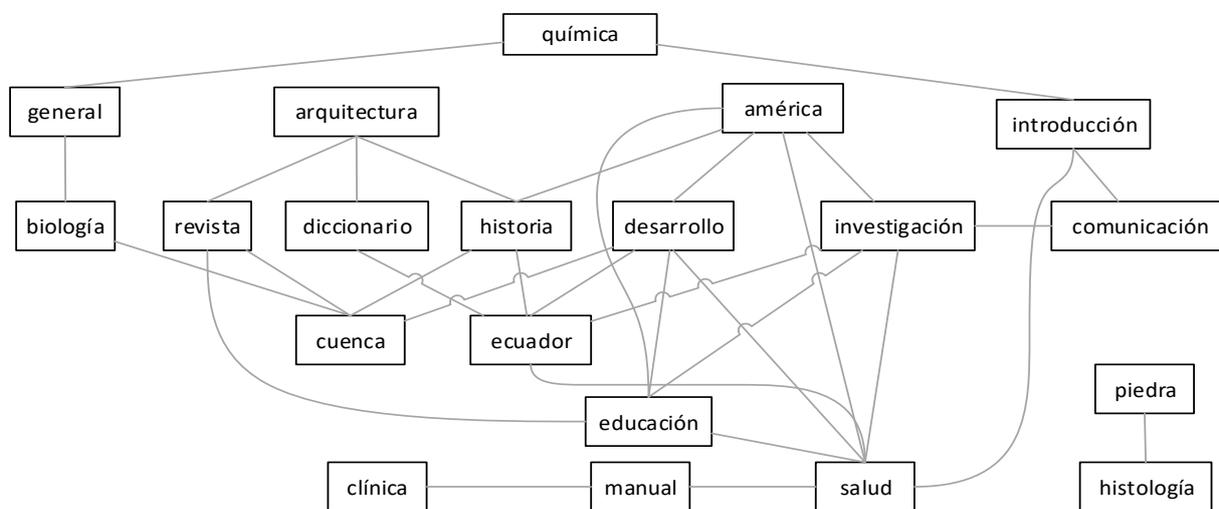


Figura 3. Grafo de correlación entre términos de búsquedas.

4.2. Proceso de préstamos

Basado en el mismo análisis del proceso de búsquedas, se procedió a la creación de la nube de las 50 palabras con la frecuencia más alta, presentada en la Figura 2b. Los cinco términos de material bibliográfico más prestados son presentados en la Tabla 3 con el respectivo valor de frecuencia, el soporte y las asociaciones entre términos. Es así que, de cada 100 libros prestados, 8 de ellos contienen el término *clínica*, 6 el término *anatomía*, 4 el término *principios* o *médica* o *manual*. Tomando el término *anatomía*, se puede concluir que tiene un valor de frecuencia equivalente a 7769 y los cinco términos con los que más aparece junto en los préstamos son: *clínica*, *humana*, *wheeler*, *descriptiva*, y *topográfica*, con un soporte de 37%, 17%, 17%, 13% y 11% respectivamente.

Tabla 3. Términos más comunes de material bibliográfico prestado a usuarios.

Término	Frecuencia	Soporte	Asociaciones
Clínica	9,537	8.04%	orientación (0.58), anatomía (0.40), exploración (0.24), propedéutica (0.21), semiología (0.15).
Anatomía	7,769	6.55%	clínica (0.37), humana (0.17), wheeler (0.17), descriptiva (0.13), topográfica (0.11).
Principios	5,165	4.36%	clásica (0.25), médica (0.18), implementación (0.13), sector (0.1), diseño (0.07).
Médica	5,084	4.29%	investigación (0.15), guía (0.12), principios (0.13), científica (0.1), clínica (0.08).
Manual	4,773	4.03%	veterinaria (0.20), ginecología (0.18), asociación (0.17), obstetricia (0.17), laboratorio (0.16).

4.3. Análisis del código de clasificación Dewey

En esta subsección se procede al análisis de la categoría Dewey³ de los recursos bibliográficos en el CDRJBV. Dewey es un código que permite una adecuada clasificación de textos, formado por diez grandes categorías de conocimiento que son: Bellas artes, Ciencias puras, Generalidades, Ciencias

³ Sistema para organizar los contenidos de una biblioteca basada en la división de conocimientos en grupos.
<https://www.britannica.com/science/Dewey-Decimal-Classification>

sociales, Filosofía y disciplinas afines, Geografía e historia, Lenguas, Literatura, Religión, Tecnología y adicionalmente se incluye la categoría tesis manejada internamente en el CDRJBV. Para el análisis se divide en los siguientes procesos:

- **Búsquedas.** Se considera el conjunto de datos buscados por el identificador del material bibliográfico (código del libro) sobre el sistema Koha, que corresponde al 16% del total de registros almacenados, esto debido a que del 84% restante solo se posee un fragmento del título, resultando inadecuado asociar el texto de búsqueda a un determinado libro. Un total de 7,916 registros conforman el conjunto de estudio.
- **Préstamos.** Para el proceso de préstamos se considera el 100% de los registros almacenados en el Data Warehouse del Centro de Documentación. Un total de 111,210 registros conforman el conjunto de estudio.

Tabla 4. Distribución del material bibliográfico buscado y prestado en base al código Dewey.

Categoría	Búsquedas		Préstamos	
	Cantidad	%	Cantidad	%
Artes y recreación	755	9.54	7,262	6.53
Ciencia	343	4.33	18,010	16.19
Ciencia de los computadores y obras generales	330	4.17	1,152	1.04
Ciencias sociales	2,185	27.60	15,916	14.31
Filosofía y psicología	523	6.61	4,584	4.12
Historia y Geografía	1,057	13.35	2,156	1.94
Lenguas	263	3.32	1,296	1.17
Literatura	1,279	16.16	2,678	2.41
Religión, teología	106	1.34	98	0.1
Tecnología	1,071	13.53	58,057	52.12
Tesis	4	0.05	1	0.001

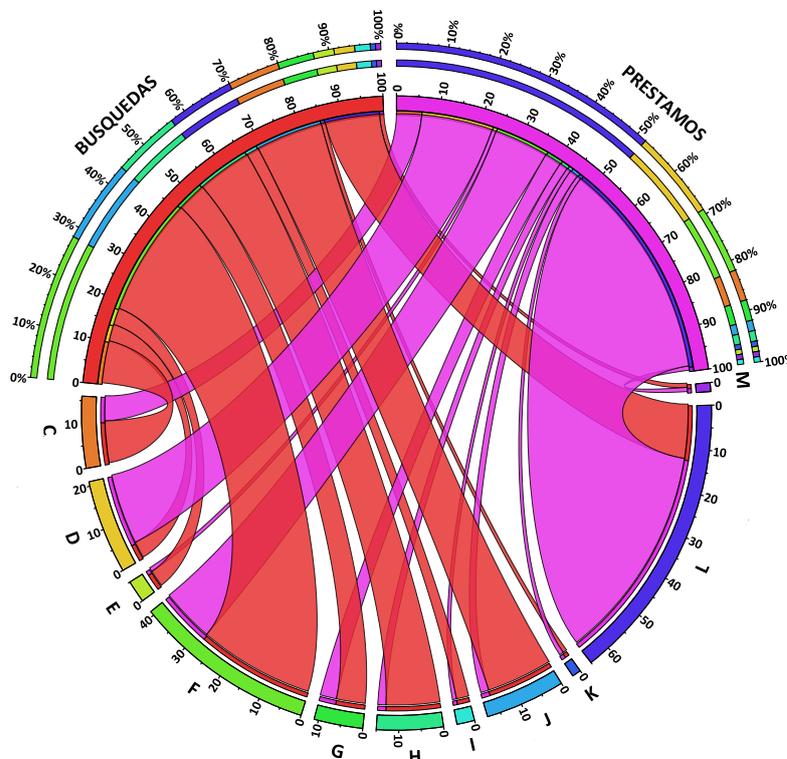


Figura 4. Relación de Dewey entre el proceso de préstamos y búsquedas. C - Artes, D - Ciencia, E - Ciencia de los computadores e información, F - Ciencias sociales, G - Filosofía y psicología, H - Historia y geografía, I - Lenguas, J - Literatura, K - Religión, L - Tecnología y M - Tesis.

La Tabla 4 presenta los materiales bibliográficos buscados y prestados en el CDRJBV, valorados por la temática de la categoría Dewey. En la Figura 4, se presenta la relación de los procesos de búsquedas y préstamos con las once clasificaciones del material bibliográfico, donde, se puede visualizar que más del 50% de préstamos corresponden a la categoría tecnología; la categoría tesis y religión son las menos solicitadas en los préstamos de recursos físicos del CDJBV con un porcentaje alrededor del 1%. Del mismo modo, con respecto al proceso de búsquedas en el sistema Koha, la gráfica indica que los términos más populares de búsquedas están dentro de la clasificación de ciencias sociales y las categorías menos buscadas también son la categoría tesis y la categoría religión.

4.4. Discusión de resultados

- Luego de presentar los resultados se pone en evidencia que:
- i) Desde la nube de palabras de búsquedas, un subconjunto de los términos está incluido en la nube de palabras de préstamos; de un total de 50 palabras que conforma la nube de préstamos, 23 de ellas están incluidas en la nube de búsquedas; siendo clínica, anatomía y manual los términos más representativos en los dos procesos (ver Figura 5). Esto permite concluir que, los usuarios del CDRJBV poseen un comportamiento similar tanto en las búsquedas como en la petición del material bibliográfico, pese a que los préstamos de recursos están dirigidos solamente a usuarios de la Universidad de Cuenca (alumnos, docentes y personal administrativo) y las consultas del material bibliográfico es abierta para el público en general.
- ii) Desde la distribución del material bibliográfico en base al código Dewey se tiene que, la mayor cantidad de recursos prestados están orientados a la categoría Tecnología para el proceso de préstamos, mientras que Ciencias Sociales es la categoría más consultada por los usuarios. De la Tabla 1 se tiene que Ciencias médicas es la Facultad con más préstamos en el CDRJBV (Haro *et al.*, 2016), lo que confirma que la categoría Tecnología posea más del 50% de recursos prestados, esto debido a que los recursos bibliográficos de medicina se encuentran en esta categoría.



Figura 5. Términos comunes entre las nubes de palabras. (a) Proceso de búsquedas, (b) proceso de préstamos.

5. CONCLUSIONES Y TRABAJOS FUTUROS

El descubrimiento de nuevo conocimiento oculto en los datos de los centros de documentación mediante técnicas de Bibliomining, permite a los administradores de los centros de documentación obtener información acerca de los intereses que tienen los usuarios y focalizar los recursos para ofrecer un mejor servicio mediante una adecuada toma de decisiones. En este artículo se presentó la implementación de técnicas de Bibliomining, sobre la información de búsquedas y préstamos del material bibliográfico en el centro de documentación de la Universidad de Cuenca, con la finalidad de descubrir patrones de comportamiento de usuarios entre los títulos que buscan y los libros que piden.

Los resultados obtenidos fueron presentados en nubes de palabras, de un total de 100 términos (50 de ellas que conforman las nubes de palabras de préstamos y 50 la nube de búsqueda), 23 términos son comunes tanto en la nube de préstamos como en las búsquedas, obteniendo que existen un comportamiento similar entre los títulos buscados y el material bibliográfico prestado con un porcentaje cercano al 50%.

A futuro se pretende extender este trabajo considerando la información académica y financiera de los usuarios, para aprovechar la gran cantidad de datos, y así descubrir nuevo conocimiento. Además, la aplicación de técnicas de predicción, clasificación y agrupación permitirá al CDRJBV proporcionar programas más apropiados, resolver las necesidades de información del usuario y observar problemas en su colección bibliográfica.

AGRADECIMIENTOS

Se agradece al Departamento de Ciencias de la Computación por el apoyo y asesoramiento brindado, así como también al personal del Centro de Documentación Regional "Juan Bautista Vázquez" que siempre están prestos a colaborar en los proyectos planteados.

REFERENCIAS

- Ahearn, L. M. (2014). Detecting research patterns and paratextual features in AE word clouds, keywords, and titles. *American Ethnologist*, 41(1), 17-30.
- Azam, I., Sohrawardi, S., Das, H. S., Alam, M. S., Alvy, M. S., & Rahman, R. M. (2013). *Bibliomining on North South University library data*. 8th International Conference on Digital Information Management, USA.
- Baeg, J. H. (2016). *The analysis of journals topics and trend: text mining and word cloud*. Illinios, USA: IConference 2016 Proceedings, 5 pp. Disponible en <https://www.ideals.illinois.edu/bitstream/handle/2142/89376/Baeg578.pdf?sequence=1&isAllowed=y>
- Brown, M. S. (2014). *Data mining for dummies*. Hoboken, USA: John Wiley & Sons.
- Chen, M.-S., Han, J., & Yu, P. (1996). *Data mining: An Overview from Database Perspective*. IEEE Transactions on Knowledge and Data Engineering, pp. 866-883.
- Cong, D. (2017). *Application of text mining in library book procurement*. 13th Global Congress on Manufacturing and Management, MATEC Web Conf., Vol. 100, 6 pp.
- Funk, M. E. (2013). Our words, our story: a textual analysis of articles published in the Bulletin of the Medical Library Association/Journal of the Medical Library Association from 1961 to 2010. *Journal of the Medical Library Association*, 101(1), 12-20. doi: 10.3163/1536-5050.101.1.003
- Haro, V., Pérez, W., & Saquicela, V. (2016). *Bibliomining para descubrir reglas de asociación en el Centro de Documentación Regional "Juan Bautista Vázquez"*. Maskana, 7(No. especial) -

- TIC.EC: Congreso Ecuatoriano de Tecnologías de la Información y Comunicaciones, pp. 203-211.
- Haro, V., Pérez, W., Siguenza-Guzman, L., Cattrysse, D., & Saquicela, V. (2014). *Diseño e implementación de un sistema de soporte de decisiones para el Centro de Documentación Regional "Juan Bautista Vázquez"*. MASKANA, Vol. 5, No. especial - TIC.EC: Congreso Ecuatoriano de Tecnologías de la Información y Comunicaciones, pp. 245-256.
- Nicholson, S. (2003). The bibliomining process: Data warehousing and data mining for library decision-making. *Information Technology and Libraries*, 22(4), 4-9.
- Nicholson, S., & Stanton, J. (2003). *Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries*. In: Nemati, H., & Barko, C. (Eds.). *Organizational data mining: Leveraging enterprise data resources for optimal performance*. Hershey, PA: Idea Group Publishing, pp. 247-262.
- Nicholson, S., & Stanton, J. (2005). *Bibliomining for library decision-making*. In: Wang, J. (Ed.). *Encyclopedia of Information Science and Technology*, IGI Global Disseminator of Knowledge, pp. 272-277.
- Ramez, E., & Navathe, S. B. (Eds.) (2010). *Fundamentals of database systems*. (6th ed.) Boston, USA: Addison-Wesley, 1201 pp.
- Riveros Malberti, M. A., & Klenzi, R. O. (2015). Reglas de asociación en el cómputo de utilización de libros en una biblioteca universitaria. *Enfoque UTE*, 6(2), 86-101.
- Siguenza-Guzman, L., Saquicela, V., & Cattrysse, D. (2014). *Design of an integrated decision support system for library holistic evaluation*. Proceedings of IATUL Conferences, Helsinki, Finland, pp.1-12.
- Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J., & Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. *The Journal of Academic Librarianship*, 41(4), 499-510, <https://doi.org/10.1016/j.acalib.2015.06.007>
- Uppal, V., & Chindwani, G. (2013). An empirical study of application of data mining techniques in library system. *International Journal of Computer Applications*, 74(11), 42-46.
- Waegel, D. (2006). *The development of text-mining tools and algorithms*. Pennsylvania, USA: Ursinus College, 32 pp.