Los Hilos De Ariadna En El Laberinto Temático: Visualización Y Minado De Datos Para Bibliotecas

Silvia Eunice Gutiérrez De la Torre
segutierrez@colmex.mx
El Colegio de México A.C., México

Julián Alberto Equihua Benítez
julian.equihua@gmail.com
CONABIO, México

Micaela Chávez Villa
mch@colmex.mx
El Colegio de México A.C., México

Introducción

Encontrar relaciones entre los encabezamientos que se asignan a una obra monográfica es un problema histórico en el ámbito de búsqueda y recuperación de información. Por un lado, los documentos rara vez pueden ser representados con un solo tema; por otro, el número de temas que se puede asignar a una obra es virtualmente infinito (Green, 2001). En la intersección de las Humanidades Digitales y la Bibliotecología han existido diversos esfuerzos por mejorar la calidad de las ontologías de estos temas (Nurmikko-Fuller et al, 2016), su evaluación (Harper, 2016) y visualización (Duguid, 2015). Sin embargo, a nuestro conocimiento, no se han hecho estudios que aprovechen métodos innovadores para indagar relaciones entre los encabezamientos de materia. En esta comunicación breve, presentamos los resultados preliminares de un primer acercamiento al tema, que aprovecha el área de especialidad de cada participante del equipo --humanidades digitales, ciencia de datos y bibliotecas-- para analizar 249,899 registros de una de las colecciones más importantes de Ciencias Sociales y Humanidades de América Latina: la del catálogo de la Biblioteca Daniel Cosío Villegas de El Colegio de México.

Metodología

A través del portal de analíticas del Grupo Ex Libris, se extrajeron los encabezamientos de materia de todos los 249,899 registros de libros de la colección de la Biblioteca Daniel Cosío Villegas. Los encabezamientos de materia fueron subdivididos a su vez en tres niveles a partir de los subencabezamientos, sin distinguir entre sus tipos –geográficos, cronológicos y de forma (ver Salta et al., 2015)– sino sólo tomando en cuenta su posición (primer subencabezamiento, segundo, etcétera). Por ejemplo, México--Historia--1821-1861 fue dividido en: México, Historia, 1821-1861.

Se estudió la relación entre temas utilizando técnicas de minería de reglas de asociación. Estas procuran descubrir implicaciones de la forma $I \rightarrow i$ donde I es un conjunto de objetos y i es un objeto en particular, ambos tomados de un universo de objetos, en este caso temas. El soporte de I se define como el número de registros para los cuales I es subconjunto. La confianza se define como el soporte de $I \cup i$ entre el soporte de I (Leskovec, 2010).

Se debe notar que la frecuencia de los temas asociados a los registros es sumamente baja como se puede observar en la Tabla 1, lo cual puede deberse a que, tratándose de una biblioteca especializada en ciencias sociales y humanidades los temas que se asignan son muy específicos, a fin de que el usuario especializado pueda encontrar lo que realmente le sirve.

	Percentiles					
Tema	25%	50%	75%	85%	95%	99%
1	1	1	3	5	22	129
2	1	1	3	6	27	219
3	1	1	3	6	28	170

Tabla 1

Asimismo, es de notar que 231,052 (92.45%) de los registros tienen un encabezamiento de materia; 152,414 (treinta por ciento menos) llega a tener dos encabezamientos de materia y sólo 29.89% tuvo tres. Por este motivo, los encabezamientos se concatenaron verticalmente para observar indistintamente las relaciones entre éstos. Se utilizó el algoritmo *a priori* y la elección de los umbrales se llevó a cabo de manera manual; se generaron 13 conjuntos de reglas de asociación con variaciones en los umbrales de confianza y soporte. Cada uno de estos conjuntos de reglas de asociación induce un grafo que se puede visualizar y explorar como se muestra más adelante. Umbrales demasiado permisivos inducen redes que tienen demasiadas relaciones como para poderse explorar manualmente y umbrales demasiado restrictivos inducen redes que no tienen suficientes relaciones como para poder decir

algo interesante sobre la estructura de los datos en su totalidad. Finalmente se eligió una red que presenta un balance entre cantidad de información e interpretabilidad. El 'soporte' mínimo fue de 0.0001 (ver Gráfico 1) y la 'confianza' mínima de 0.4 (ver Gráfico 2) y la matriz de incidencia derivada de las reglas encontradas se utilizó para generar un grafo para la exploración visual del conjunto de asociaciones descubiertas. Para crear esta versión gráfica utilizamos la exportación de R a Gephi (Yon and Yon, 2015), la 'confianza' como un peso para los vértices y Fruchterman Reingold (1991) como algoritmo para el diseño. Dimos color a los nodos de acuerdo con su modularidad, es decir, de acuerdo a las "comunidades" de nodos que se crean por la fuerza de sus relaciones (Blondel et al, 2008). La alta modularidad de la red prueba lo conectados que están los nodos en sus grupos y lo desconectados que están de nodos fuera de su red.

Resultados

Como hemos mencionado antes, los encabezamientos fueron divididos en los subencabezamientos que los anidan. Retomando el ejemplo anterior: "México--Historia--1821-1861" fue codificado como:

- Subject 1.1 - México
- Subject 1.2 - Historia
- Subject 1.3 - 1821-186

Este modelado de los datos, fue pensado para permitir una cierta exploración "gramática" de la asignación temática. Es decir, que permitiera ver qué niveles "sintácticos" se relacionan en qué orden con otros niveles. En números, la red tiene 394 nodos (subencabezamientos) y 339 vértices (asociaciones). De los nodos, 203 son del primer nivel, 109 del segundo, 33 de la combinación de un encabezamiento del primer nivel con el tercero, y cuatro de la combinación del primer nivel con el cuarto. El total asociaciones o reglas de implicación (si encabezamiento *I* aparece también *i*) fue de 339. De éstas la mayoría ocurre sólo en 25 registros, es decir, tuvieron un soporte bajo (ver Gráfico 1). Sin embargo, esto no es tan poco considerando lo que hemos dicho antes de la naturaleza especializada de esta biblioteca. Por otro lado, las confianzas observadas presentan una distribución menos concentrada que la de los soportes (ver Gráfico 2).

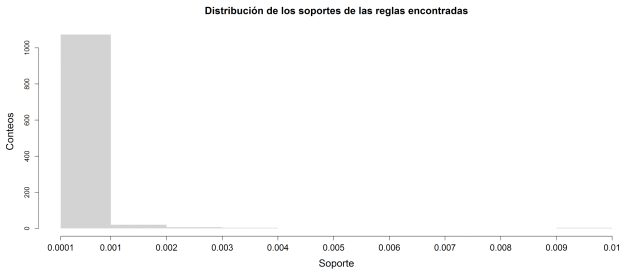

Gráfico 1

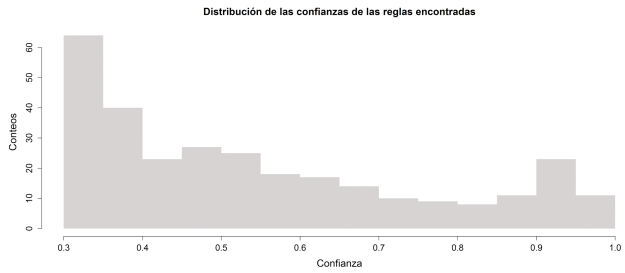

Gráfico 2

De la red de grafo interactiva que obtuvimos con el uso de Gephi y el *plug-in* de Sigma.js, pudimos identificar que el nodo con mayores asociaciones o reglas es 'Historia' en su posición como "Subject 1.2" y que entre sus asociaciones existen dos nodos de distinta modularidad y nivel (ambos "Subject 1.1"): 'México' (ver Imagen 1) y 'España' (ver Imagen 2).

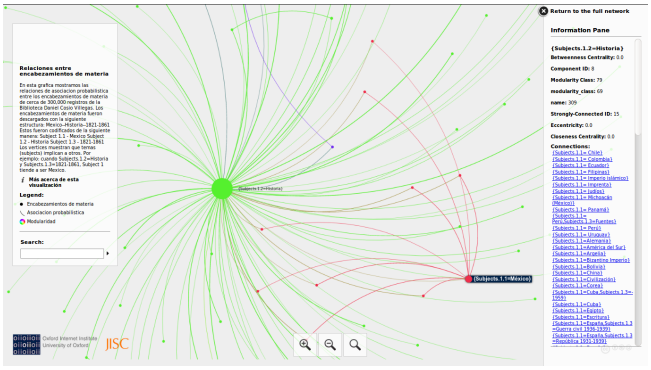

Imagen 1

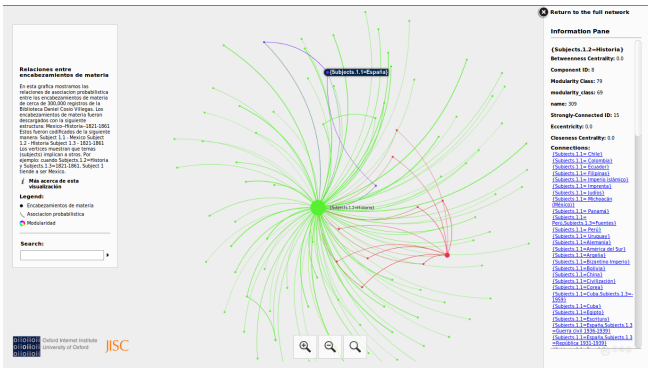

Imagen 2

A su vez, la plataforma permite explorar más a fondo el encabezamiento 'España' y darse cuenta, por ejemplo, de que este tema en primera posición tiene fuertes relaciones con subencabezamientos de la tercera dimensión que corresponden a los periodos históricos relevantes en la historia de ese país:

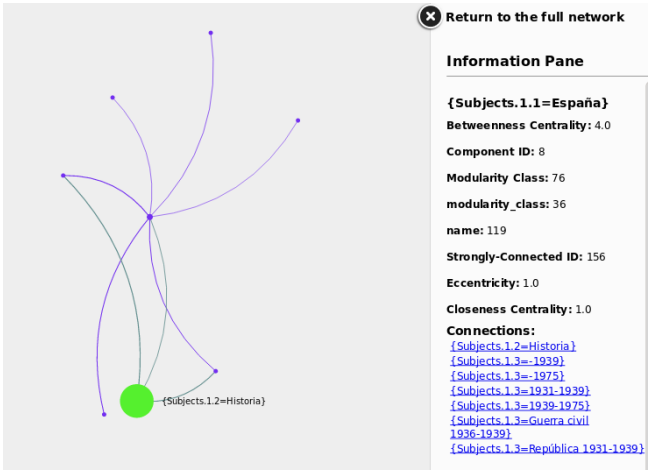

Imagen 3

En resumen, este tipo de exploración permite al usuario familiarizarse con las reglas "gramaticales" de la asignación temática pues puede "ver" tanto los niveles "sintácticos" de los temas como las formas en que se relaciona con otros, además de que incluye un botón de búsqueda de encabezamientos que permite interactuar de manera directa con el grafo (disponible en linea).

Reflexión final

Nosotros, como lo sugieren Nurmikko-Fuller et al., estamos conscientes de que si las bibliotecas quieren dar acceso a recursos de información relevantes para nuevas áreas de investigación, deben evolucionar a métodos más sofisticados y semánticos de asignación temática para proporcionar nuevos puntos de acceso que correspondan más al lenguaje natural y que permitan identificar las relaciones temáticas con mayor claridad.

Sin embargo, en lo que este paso puede ser dado en México y Latinoamérica, creemos que el uso de herramientas y métodos de las humanidades digitales pueden ayudar a analizar los datos generados en la organización de la información e incluso útil para la formación del catalogador, que aprende a asignar-elaborar los temas y con esta herramienta podría tener un acceso visual a la "sintaxis temática" de ciertos términos. En este mismo sentido, un acercamiento así, podría ser usado como elemento pedagógico de los cursos de investigación documental en el que los estudiantes deben aprender a familiarizarse con los lenguajes controlados. Otra aplicación de este trabajo, podría ser en la evaluación de colecciones para determinar las fortalezas y carencias temáticas, de acuerdo con la especialidad que la biblioteca declara. Análisis más detenidos pueden ayudarnos a determinar la representación cronológica, autoral, lingüística o geográfica de un acervo. En fin, consideramos que al continuar el análisis y desarrollo de este proyecto podremos aportar otro tipo de metodología no sólo para evaluar las colecciones sino para acercarse a ellas.

Bibliografía

Blondel, V., et al. (2008). "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, P1008.

Duguid, T. (2015), "BigDIVA: Big Data, Big Visuals, Big Searches, and Big Results." *Texas Digital Humanities Conference 2015.* University of Texas Arlington, Texas.

Fruchterman, T. M., & Reingold, E. M. (1991). *Graph drawing by force-directed placement. Software: Practice and experience,* 21(11), pp. 1129-64.

Green, R. *(2001). "Relationships in the o*rganization of knowledge: an overview." *Relationships in the organization of knowledge*. Springer Netherlands, pp. 3-18.

Nurmikko-Fuller, T., Jett, J., Cole, T., Maden, C., Page, K., Downie, J. (2016). "A Comparative Analysis of Bibliographic Ontologies: Implications for Digital Humanities". *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 639-42.

Leskovec, J., Rajaraman, A., Jeffrey, U. (2010). *Mining of Massive Datasets.* Cambridge University Press, U.K., pp. 205-14.

Salta, G., Cravero C., Saloj, G. (2005) "Lista de encabezamientos de materia de la Biblioteca del Congreso de los Estados Unidos: características generales". *Información, Cultura y Sociedad*, 12. pp. 85-97

Yon, G. V., & Yon, M. G. V. (2015). Package 'rgexf'.