Jornadas Temas Actuales en Bibliotecología

Desarrollo de una herramienta para el análisis y representación semántica de colecciones documentales a través del factor TF-IDF

Liberatore, Gustavo
gliberat@mdp.edu.ar
Vuotto, Andrés
avuotto@gmail.com
Fernández, Gladys Vanesa
gvfernan07@gmail.com

Resumen: (Hasta 250 palabras)

Se describe el desarrollo y de una aplicación para el análisis semántico de una colección documental. La herramienta permite interactuar con otros sistemas de gestión documental que aporten una organización del conocimiento específica, a partir de los términos de indexación de sus documentos y definiendo de forma automatizada un espacio vectorial en función de los pesos semánticos de cada término evaluado a partir del factor TF-IDF. Como resultado a la consulta del usuario se construye un gráfico hipertextual, el que representa el nivel de presencia de cada concepto en una o varias colecciones, y facilitando la navegación de las bases de datos vinculadas y el acceso a sus objetos almacenados. La aplicación puede interactuar con cualquier sistema que permita la interoperabilidad, y pueda organizar los datos en correlación con los criterios establecidos en el complemento presentado. En este caso particular, la experiencia se trabajó bajo el sistema para repositorios DSpace.

Palabras clave:

Análisis semántico; representación del conocimiento; TF-IDF; Recuperación de información. Semantic analysis; Knowledge representation;TF-IDF; Information retrieval

1. Introducción

Los sistemas bibliotecarios en tanto artefactos diseñados y concebidos para la mediación en el acceso a la información (bibliográfica) han desarrollado y adoptado diversos métodos e instrumentos que faciliten (aunque no siempre) la satisfacción de la una demanda informativa. Por lo general y desde un enfoque técnico-instrumental el bibliotecario se ha valido de metodologías de análisis y representación del contenido y de sistemas de organización orientados a lenguajes que otorguen puntos de acceso controlados para la búsqueda y recuperación de información bibliográfica por parte del usuario. Las posturas

acerca de cuáles son las mejores opciones para la recuperación temática de los documentos han estado centradas sin embargo, más que en los métodos de análisis, en el antagonismo sobre la aplicación de sistemas de indización basados en lenguajes controlados o en lenguajes libres. La discusión nace hace varias décadas atrás, en el inicio de los años cincuenta, cuando comienzan los experimentos *Cranfield* destinados a generar diversas medidas y criterios para evaluar los sistemas de recuperación de información. Entre los resultados obtenidos en estas pruebas (discutidos y polemizados durante años por los especialistas) se obtuvo la comprobación empírica de que el lenguaje libre de interrogación tiene una leve ventaja por sobre los sistemas de organización estructurados y controlados (encabezamientos de materia, sistemas de clasificación), en lo que genéricamente se dio en llamar "eficiencia" en la recuperación. Aun así, a pesar de estas evidencias, la mirada o enfoque bibliotecológico (hoy lo denominamos "clásico") acerca de la organización, representación y recuperación de información estuvo orientado históricamente sobre los siguientes principios (Hjorland, 2008):

- El uso de vocabularios controlados.
- La regla de Cutter acerca de la especificidad.
- El principio de Hulme de la garantía literaria.
- La organización desde lo general a lo específico.

Y, deberíamos agregar a esta lista, de manera más contemporánea, la temprana adopción del modelo booleano de recuperación.

En paralelo al recorrido realizado desde el campo bibliotecológico sobre estas problemáticas fueron surgiendo diversos avances en el terreno de la recuperación de información (RI), sobre todo desde aquellas líneas de investigación emprendidas dentro del denominado paradigma físico (algorítmico). En particular, frente a la necesidad de encontrar respuestas a los problemas derivados del acceso, búsqueda, localización y recuperación de grandes volúmenes de datos/información, los estudios se orientaron hacia dos metodologías bien diferenciadas: el *querying* (interrogación) y el *browsing* (exploración/navegación).

Los sistemas basados en *querying* utilizan una ecuación de búsqueda como método de interrogación a un sistema de recuperación de información el cual realiza una equiparación entre la consulta y el fondo documental. El modelo de RI tradicionalmente utilizado ha sido el de equiparación exacta (booleano) el cual posee algunas limitaciones derivadas del modo en que resuelve una consulta, esto es, proporcionando resultados que cumplen exactamente con los términos propuestos para ejecutar la búsqueda (modo binario). En la década de los ochenta comienzan a aparecer nuevos enfoques derivados del modelo de RI denominado de equiparación parcial, particularmente desde los aportes realizados por Salton (1989) a partir de su modelo del espacio vectorial, introduciendo el concepto de relevancia en la

recuperación, determinada por la ordenación de los resultados de una búsqueda en función de su validez o pertinencia. En esencia, las distintas expresiones del modelo de equiparación parcial se sustentan en la posibilidad de:

- Transformar a los documentos o sus representaciones en expresiones numéricas.
- Ponderar los términos que los componen.
- Equiparar una consulta en función del peso que tienen los términos que componen los documentos pertinentes.

El método del *browsing*, en cambio, se sustenta en la idea de la exploración y la navegación (hipertextual) de los documentos, en una técnica que se orienta más a lo visual y a la posibilidad de explorar grandes volúmenes de información mediante interfaces que permitan la observación de estructuras, relaciones y conjuntos de documentos asociados por elementos comunes (Cove & Walsh, 1987 & 1988). Las técnicas de *browsing* cobraron fuerza en su aplicación a partir de la aparición de internet en el inicio de la década de los noventa introduciendo, para este entorno, nuevas formas de acceso basadas en la visualización de grandes volúmenes de información mediante interfaces gráficas.

Ambas técnicas, el *querying* y el *browsing*, pueden ser complementarias en un sistema de recuperación de información.

2. Contexto y objetivo

Las estructuras de datos formalizadas para el almacenamiento o captura de recursos de información y su posterior recuperación tienden cada vez más a profundizar los procesos de semantización de las consultas en los sistemas bibliográficos de información (Bosch y Manzanos, 2013). Esta realidad se ha hecho más evidente en bases de datos de objetos documentales surgidas por fuera de los sistemas integrados de bibliotecas orientados a los Web-OPACs ya que han adoptado con mayor facilidad los nuevos estándares propuestos por la web 3.0. Uno de estos casos lo constituyen los repositorios institucionales surgidos desde una filosofía completamente diferente a la de los tradicionales catálogos de biblioteca.

El caso que nos ocupa se circunscribe al repositorio institucional de la Facultad de Humanidades de la UNMdP sobre el cual se vienen desarrollando diversas aplicaciones en torno a las formas de organización y recuperación de los objetos digitales que en él se depositan (Liberatore, Hernández y Saya, 2014). El diseño de herramientas semánticas para la búsqueda y recuperación (por fuera de las formas tradicionales) en un repositorio institucional (RI) supone contemplar no solamente las características y constitución de él/los campos semánticos que lo conforman sino, además, las particularidades de la institución

desde donde se genera, ya que la misma determina las condiciones de producción, registro y comunicación del conocimiento almacenado. Este último factor condiciona particularmente el modo en que se articula y constituye un dominio, en este caso las “humanidades” en la UNMdP, ya que los productos intelectuales se encuentran sujetos a un plexo normativo que los determina en su forma y contenido y los categoriza según su ámbito de procedencia u origen dentro de las áreas de actuación académica: docencia, investigación, extensión y gestión. En paralelo es importante considerar la división interna que el propio espacio institucional posee en términos de las disciplinas que lo constituyen, aspecto este que no puede abordarse por lo general desde una perspectiva epistemológica de la división del conocimiento dentro del campo de las humanidades sino que, más bien, responde a una delimitación sujeta a variables históricas, políticas y sociales. En este punto resulta funcional el abordaje que aporta Hjorland (2008) desde su teoría de la organización del conocimiento en torno a las divisiones “cognitiva” y “social”, entendiendo por la primera a los sistemas conceptuales estructurados sobre visiones paradigmáticas de la ciencia y, en el segundo caso, a una división del conocimiento moldeada por variables institucionales, profesionales y de la propia práctica científica.

Atendiendo estos factores la aplicación que se describe en este trabajo se estructura sobre la idea de la representación y visualización de las distintas colecciones del RI por medio de las palabras clave que aportan los autores en el proceso de autoarchivo de sus producciones académicas. Básicamente se intenta aprovechar la riqueza y fidelidad (en términos de representación semántica) del lenguaje libre aportado en la descripción de los objetos digitales generando representaciones a partir del indicador TF-IDF (*Term Frequency - Inverse Document Frequency*) (Sparck Jones, 1972; Salton, 1989) por medio de interfaces gráficas de consulta.

3. Metodología

a. Aplicación del factor TF-IDF como medida para la ponderación de las palabras claves

La representación semántica de las colecciones implicadas requiere la intervención de un mecanismo de ponderación, el cual debe permitir identificar el peso de cada palabra clave en relación a la comunidad de términos en la que participa. El indicador TF-IDF (Term Frequency - Inverse Document Frequency) permite la construcción de modelos de recuperación de información avanzados de tipo vectorial, caracterizados por la representación de conceptos por medio de la vectorización de la colección, dando como resultado una matriz de datos con tantas columnas como términos, y tantas filas como documentos. El modelo vectorial, iniciado por Salton (1989), ha dado lugar al desarrollo de

innovadoras interfaces de búsqueda y recuperación y a numerosas investigaciones que han ido mejorando e incorporando alternativas a su forma inicial (Vargas Rosales, 2015; Ropero Montejo, 2014; Cobo Ortega, Rocha Blanco y Alonso Martínez, 2009; H. Wu and R. Luk and K. Wong and K. Kwok, 2008).

La posibilidad que este modelo tiene, de construir un espacio de múltiples dimensiones, parte de su capacidad de cuantificar el peso de cada palabra en una colección específica. Diferenciándose del cálculo de frecuencia absoluta principalmente por incluir en sus cálculos la “Frecuencia de aparición del término TF" y la "Frecuencia inversa del documento para un término IDF”. Para comprender la funcionalidad de este cálculo de representación vectorial se debe prestar atención a los sub-factores que se incluyen en su fórmula:

- El factor TF calcula la capacidad de representación del término en un documento o colección a través de la obtención de su frecuencia de aparición. Su fórmula es:Tf(n)=∑D1(n). Frecuencia de aparición de un término (n) en un documento (D1), es la suma de sus ocurrencias.
- Factor IDF de cada término en cada documento o colección: Es el coeficiente que determina la capacidad discriminatoria del término con respecto a la colección.

Continuando con el desarrollo de la fórmula, esta se completa de la siguiente manera:

- IDF(n)= log10 N/DF(n)+1: Donde N es el número total de documentos, DF es el número de documentos donde aparece el término n. El logaritmo se utiliza para obtener un coeficiente bajo de fácil manejo, y el +1 funciona como factor correctivo del resultado.
- Cálculo de la Ponderación TF-IDF de cada término: Corresponde al producto de ambos factores. Los resultados son una representación de la importancia del término en cada documento y por consiguiente de la presencia del tópico en la unidad de análisis trabajada.

A los efectos prácticos, considerando la particularidad de la base de datos fuente de esta investigación, se tomaron como elementos participantes de la fórmula los siguientes:

- Términos de análisis: las palabras claves, provenientes del lenguaje libre, incluídas en cada uno de los documentos almacenados
- Texto de los documentos: en este caso se consideraron el título en forma conjunta con el resumen, por ser estos la abstracción del texto completo con mayor

extensión en palabras en comparación con la información guardada en otros campos: Y, probablemente, la que mejor describe el contenido de todo el documento incluyendo los conceptos de mayor significado en el dominio de estudio.

A modo de ejemplo, para ampliar la explicación del factor se pueden tomar los siguientes datos:

- Un documento que contiene 100 palabras en el que el término "taxonomía" aparece 5 veces. La frecuencia del término (es decir, tf) para el caso "taxonomía" es entonces (5/100) = 0,05.
- Si la colección tiene 200 documentos y la palabra "taxonomía" aparece en treinta de ellos. Entonces, la frecuencia de documento inversa (es decir, idf) se calcula como log (200 / 30) = 0,82.
- De este modo, el peso Tf- idf es el producto de estas cantidades: 0,05 * 0,82 = 0,041. Y aplicando el corrector +1, quedaría en 1,041.

b. Fuente de datos y esquema de trabajo

En este caso en particular se tomó como universo de aplicación el Repositorio de la Facultad de Humanidades de la Universidad Nacional de Mar del Plata (HUMADOC) creado y gestionado con el software DSpace.

DSpace es un sistema de información con arquitectura de repositorio digital que captura, almacena, ordena, preserva y distribuye material de investigación digital con el propósito de garantizar que se preserve y distribuya toda la producción intelectual generada al interior de las instituciones que hacen uso de este. Esta herramienta está desarrollada bajo la filosofía del opensource por lo cual es gratuita y se puede personalizar según las necesidades. Es un proyecto conjunto de las bibliotecas del MIT (Massachussets Intitute of Technology) y Hewlet-Packard Co. El modelo de datos de Dspace está organizado en comunidades, subcomunidades, colecciones e items. La arquitectura de DSpace se encuentra estructurada en

1- Una capa de almacenamiento (esta capa es la base de datos, donde se almacena toda la información descriptiva del material depositado).
2- Una capa lógica de negocios (esta capa media entre la base de datos y el usuario).
3- Una capa de aplicación (es la capa que visualiza el usuario y que le permite interactuar con el sistema).

El esquema de trabajo se basó en la extracción de las palabras clave en lenguaje libre asignadas por los autores de las tesis depositadas en el repositorio por medio del proceso de autoarchivo. En este caso, las palabras clave son almacenadas en una base de datos PostgreSQL (un sistema de gestión de bases de datos relacional orientado a objetos y libre) más precisamente en el campo 57 de de la tabla metadatavalue. También, a los efectos del cálculo TF-IDF, se extrajeron otras partes del registro como título y resumen.

Debido a que en este caso experimental sólo se tomó la colección de tesis, se realizó una ecuación que combinara las palabras clave con el id almacenado referente a dicha colección. Como paso siguiente generó una lista de recuperación con estos datos representadas en umbrales ya que debido a que la representación gráfica puede resultar poco accesible cuando todos los términos de una colección son mostrados en su totalidad. Cada término enlaza con un conjunto de registros que fueron descriptos con la palabra clave en cuestión.

En este caso se aplicó la herramienta sobre un repositorio pero, debido a su flexibilidad, es posible utilizarla a cualquier sistema que cuente con una arquitectura similar, (SIGB, revista electrónica, etc.) que permita la interoperabilidad y pueda organizar los datos en correlación con los criterios establecidos en el complemento presentado, más precisamente que almacene los metadatos en una base relacional y que esta sea susceptible de interrogar, acceder y exportar.

c. Tecnologías aplicadas en el procesamiento y modelado de la información

El algoritmo desarrollado permite confeccionar un gráfico en el mismo momento en el que el usuario realiza su consulta al sistema, facilitando de esta forma el acceso a resultados de búsqueda por medio de conexiones hipertextuales y también mostrando perfiles semánticos de las colecciones expuestas. La secuencia de rutinas que incluye son:

- Acopio de listado de palabras claves y todos los demás metadatos necesarios
- Procesamiento y tratamiento previo de los textos
- Cálculo de un valor tf-idf para cada palabra incluida
- Determinación de los diferentes umbrales de presentación, generando un núcleo de las palabras claves con mayor valor TF-IDF y por consiguiente participantes del gráfico resultante.
- Construcción de las imágenes del dominio consultado.

Para ello se trabajó con lenguaje de programación PHP. Como gestores de bases de datos se utilizó el nativo del sistema Dspace, postgreSQL; y se reforzó el aspecto de gestión de datos con MySQL. La interface se trabajó en HTML5, en combinación con CSS3 y el uso de la herramienta Google Chart para la confección de grafos que trabajen la tecnología SVG para el desarrollo de espacios vectoriales en entornos web.

4. Consideraciones finales

Como se mencionaba en la introducción de este trabajo las formas tradicionales o canónicas de representación y organización de la información en el campo bibliotecológico han producido en nuestros profesionales cierto conservadurismo respecto a los métodos e instrumentos de aplicación dentro de este campo. Frente a nuevas formas de producción, almacenamiento y acceso a la información resulta importante repensar y adecuar formas alternativas que ofrezcan una recuperabilidad distinta de aquello que se busca o se desconoce, atendiendo a contextos y situaciones particulares. No se trata de desechar lo tradicional (la técnica utilizada aquí tiene una considerable antigüedad) sino más bien el de resignificar nuestro rol de mediadores en los actuales escenarios informativos.

5. Bibliografía

- Bosch, M. & Manzanos, N. (2013). De los registros a los objetos: Semántica y comportamiento de los documentos: el desafío de la Web 3.0. *Palabra Clave (La Plata)*, 2(1), 51-60.
- Cobo, A., Rocha, R. & Alonso, M. (2009). Descubrimiento de conocimiento en repositorios documentales mediante técnicas de minería de texto y swarm intelligence. Revista Electrónica de Comunicaciones y Trabajo de Asepuma, 10, 105-124.
- Cove, J. F., & Walsh, B. C. (1987). Browsing as a means of online text retrieval. *Information Services & Use*, *7*(6), 183-188.
- Cove, J. J., & Walsh, B. C. (1988). Online text retrieval via browsing. *Information Processing & Management*, *24*(1), 31-38.
- Hjorland, B. (2008). What is Knowledge Organization (KO)? *Knowledge Organization*, 35 (2-3), 86-101.
- Liberatore, G., Hernández, A. & Saya, J. (octubre, 2014). *Diseño e implementación de un sistema de organización del conocimiento en un entorno institucional multidisciplinar: el caso del repositorio de la Facultad de Humanidades de la UNMdP.* X ENCUENTRO DE DIRECTORES Y IX DE DOCENTES DE ESCUELAS DE BIBLIOTECOLOGÍA Y CIENCIAS DE LA INFORMACIÓN DEL MERCOSUR, Biblioteca Nacional, Buenos Aires, Argentina.
- Ropero Montejo, F. T. (2014) Método para la evaluación automática de la organización de textos argumentativos. Tesis para la obtención del título

Magister en Ingeniería de Sistemas y Computación. Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Bogotá, Colombia.

- Salton, G. (1989). Automatic text processing: the transformation, analysis and retrieval of information by computer. Reading, MA: Addison Wesley.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation,* 28(1), 11-21.
- Vargas Rosales, A. A. (2015). Desarrollo de una herramienta que permita la extracción de una taxonomía de un conjunto de documentos de un dominio específico usando CFinder para la extracción de conceptos clave. Tesis para la obtención del título de Ingeniero Informático. Facultad de Ciencias e Ingeniería, Pontificia Universidad Católica del Perú, Lima, Perú.
- Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS), 26(3), 13.