IFLA WLIC 2013
SINGAPORE

Enviado en: 19/02/2014

Conexión de los contenidos de la biblioteca utilizando minería de datos y análisis de texto en datos estructurados y no estructurados

Traducción al español del documento original: " Connecting library content using data mining and text analytics on structured and unstructured data"

Chee Kiam Lim
Technology and Innovation, National Library Board, Singapore
E-mail address: chee_kiam_lim@nlb.gov.sg

Balakumar Chinnasamy
Technology and Innovation, National Library Board, Singapore.
E-mail address: balakumar_chinnasamy@nlb.gov.sg

TRADUCTOR/A: Luisa María Landáburu Areta, Biblioteca Nacional de España

Resumen:

Con tantos datos disponibles, ¿cómo pueden los ocupados usuarios encontrar los bits adecuados de información?. La honorable búsqueda hace un gran trabajo, pero ¿es suficiente?

Una búsqueda típica en un popular motor de búsqueda de Internet devuelve miles de resultados para que los filtre el usuario. Después de encontrar un artículo pertinente, continua este tedioso proceso para encontrar el siguiente y el siguiente hasta que son solventadas las necesidades de información de los usuarios (o se dan por vencidos).

Deberíamos enviarles paquetes de información pertinente, en vez de hacer que los usuarios repitan el tedioso proceso de búsqueda y filtrado. Y para hacer esto, debemos conectar nuestro contenido.

Los avances en tecnologías Big Data nos muestran importantes oportunidades para conectar la enorme y cada vez mayor cantidad de recursos de información en nuestros repositorios.

Mediante el uso de técnicas de minería de datos y análisis de texto, y tecnologías Big Data, la National Library Board (NLB) de Singapur ha conectado nuestro contenido estructurado y no estructurado. Esto nos ha permitido ofrecer una información completa, relevante y de confianza a nuestros usuarios.

Palabras clave: minería de datos, análisis de texto, Mahout, Hadoop, National Library Board de Singapur.

---

1 CONEXIÓN DE DATOS ESTRUCTURADOS

La amplia red de bibliotecas de la National Library Board de Singapur, tiene una colección de más de un millón de títulos físicos. Estos títulos generan más de 30 millones de préstamos anuales.

Mediante el uso de técnicas de minería de datos sobre las operaciones de préstamo pasadas y los registros bibliográficos de los libros, la NLB ha conectado correctamente nuestros títulos y puso en marcha nuestro propio servicio de recomendación de títulos en 2009.

Nuestro servicio de recomendación de títulos puede encontrarse en muchos de los sitios web y portales de la NLB.

Entre ellos:

"Read On" (http://readon.sg)

Read On" es un micrositio amigable con motor de búsqueda de la NLB, que permite que los buscadores rastreen e indicen con facilidad las colecciones de libros de la NLB.

Figura 1.1 Recomendaciones de títulos en el micrositio "Read On"de la NLB

"Library in Your Pocket" (LiYP) (http://m.nlb.gov.sg) LiYP es un portal bibliotecario para dispositivos móviles de la NLB

Figura 1.2: Recomendaciones de títulos en el sitio móvil LiYP de la NLB

Portal de búsqueda "SearchPlus" (http://searchplus.nlb.gov.sg)

El buscador "SearchPlus"de NLB para títulos físicos y recursos digitales, incluyendo libros electrónicos y bases de datos electrónicas.

Figura 1.3 Recomendaciones de títulos en el buscador "SeachPlus" de NLB

Servicio al usuario en línea "Check Your Account" ("Comprueba tu cuenta") del portal de las Bibliotecas públicas de la NLB (http://www.pl.sg)

"Check Your Account" del portal de las Bibliotecas públicas de la NLB permite a nuestros usuarios comprobar la información de su cuenta, tal como ejemplares en préstamo, fechas de vencimiento y las multas pendientes.

Figura 1.4 Recomendaciones de título del portal de las Bibliotecas públicas de la NLB

El Servicio de recomendaciones de títulos también disponible para que desarrolladores externos lo utilicen a través de la iniciativa de la NLB Open Data / Web (http://nlblabs.sg)

La minería de datos es el proceso de descubrir patrones en grandes volúmenes de conjuntos de datos. A continuación se describe el método y proceso adoptado por NLB.

La implementación del servicio de recomendación de títulos de la NLB es un sistema híbrido de recomendación que combina tanto el filtrado colaborativo como el basado en el contenido.

Dado un título, el servicio de recomendación de títulos genera 2 tipos de recomendaciones: “Our patrons also borrowed the following” ("Nuestros usuarios también pidieron prestado lo siguiente") y "Quick Picks” (“Selección rápida”).

Figura 1.5 Recomendaciones: “Our patrons also borrowed the following” " Nuestros usuarios también pidieron prestado lo siguiente" y "Quick Picks” (“Selección rápida”).

La categoría “Our patrons also borrowed the following” ("Nuestros usuarios también pidieron prestado lo siguiente") se basa principalmente en un filtro colaborativo de recomendaciones. El filtrado colaborativo analiza los patrones de lectura dentro de los cientos de millones de registros de préstamo para elaborar las recomendaciones. Se basa en la idea de que tenemos una tendencia a regirnos por recomendaciones de personas con los mismos intereses. Para generar las recomendaciones solamente se consideran los préstamos de los últimos 3 años, de esta forma se minimizan las recomendaciones que pudieran ser anticuadas.

En los casos en que el filtrado colaborativo falla debido al bajo o nulo préstamo, el algoritmo vuelve al filtrado basado en contenido usando los registros bibliográficos.

Por lo tanto, es un sistema de recomendación robusto y detallado que tiene en cuenta los préstamos reales de los usuarios, y como resultado, también ajustará sus recomendaciones si hay cambios en los patrones de lectura de estos.

Figura 1.6 Filtrado colaborativo y basado en contenido

Las recomendaciones "Quick Picks" se basan únicamente en el filtrado basado en contenido para proponer recomendaciones utilizando los registros bibliográficos. Pero estas recomendaciones se hacen a partir de una lista de libros seleccionados por los bibliotecarios en lugar de con toda la colección.

Actualmente las recomendaciones a nivel de usuario se implementan además de las recomendaciones a nivel de título Los últimos títulos prestados del usuario se utilizan para proporcionar recomendaciones que luego se fusionaran en una única lista.

Figura 1.7 Recomendaciones a nivel usuario

Tabla 1.1 muestra algunas estadísticas sobre la cobertura de recomendación

Los títulos más prestados son los de ficción y, por ello, el filtrado colaborativo genera recomendaciones para el 59% del total de títulos de ficción, en comparación con un sólo un

28% de porcentaje de éxito para los menos prestados, títulos de no ficción.

	Filtro en colaboración	Combinado con filtrado simple basado en contenido
Ficción	59%	89%
No ficción	28%	53%
Total	34%	59%

Tabla 1.1 Cobertura de recomendación de la NLB

El filtrado simple basado en el contenido aplicado aquí, sólo utiliza los campos de autor e idioma. Como puede verse, incluso combinando filtrado colaborativo con un muy sencillo algoritmo de filtrado basado en contenido mejora significativamente la cobertura de recomendación, elevando el porcentaje de éxito de ficción de un aceptable 59% a un excelente 89% y de no ficción de un bajo 28% a un aceptable 53%.

La cobertura de recomendación para las recomendaciones a nivel usuario es aún mayor, ya que estas son la combinación de las recomendaciones a nivel de título basadas en los préstamos recientes al usuario.

2 CONEXIÓN DE DATOS NO ESTRUCTURADOS

Los datos no estructurados constituyen una parte enorme y creciente de los fondos de la NLB . La conexión de estos datos nos permitirá hacer mucho más visibles estos contenidos.

Uno de nuestros primeros intentos de conectar estos contenidos implica el uso de análisis de texto para realizar la "extracción de la frase clave". Las frases clave extraídas se utilizan entonces para realizar búsquedas que revelan otro contenido relacionado. Estas conexiones garantizan que están, a tan solo 2 clics de distancia, los siguientes descubrimientos sobre los temas clave.

El micrositio amigable con motor de búsqueda de NLB "Infopedia" (http://infopedia.nl.sg) es una popular enciclopedia electrónica de la historia, cultura, gente y acontecimientos de Singapur. Con menos de 2.000 artículos, atrae más de 200.000 páginas vistas al mes.

Hace uso de este enfoque junto con las recomendaciones manuales de los bibliotecarios. Estas recomendaciones manuales de los bibliotecarios se añaden como Recomendaciones del bibliotecario y proporciona con 1 clic acceso al contenido relacionado.

Figura 2.1: Conexiones Infopedia través de frases clave extraídas y recomendaciones manuales

Sin embargo, como las recomendaciones manuales son una tarea laboriosa y requieren mucho tiempo, no todos los artículos tienen *Recomendaciones del bibliotecario.*

Con la mayoría de las otras colecciones, muchas veces más grandes que la colección Infopedia,

la NLB necesita un enfoque alternativo para generar automáticamente buenas recomendaciones para su contenido no estructurado.

Apache Mahout es un software de código abierto ampliable de la Fundación Apache que implementa una amplia gama de algoritmos de minería de datos y aprendizaje de la máquina (http://mahout.apache.org). Es compatible con cuatro casos de uso principales, en concreto minería de recomendación, agrupamiento, clasificación y minería del conjunto de elementos frecuentes.

Usando análisis de texto para encontrar artículos similares en Apache Mahout, en menos de 5 minutos, generamos con éxito recomendaciones para la colección completa de Infopedia.

Figura 2.2 Muestra de resultados del análisis de texto de Infopedia

Usando el mismo artículo de Infopedia en la Figura 2.1 (titulado "King Edward VII College of Medicine "), la figura 2.2 muestra que 3 de las recomendaciones manuales se encuentran entre los mejores 4 recomendaciones generadas automáticamente por Mahout.

Los análisis de texto Mahout se basan en algoritmos matemáticos establecidos y rigurosos (como el Euclideo, Coseno, Tanimoto, Manhattan y muchos otros para la medición de distancias, frecuencia de términos en el documento y su inversa, etc). Las recomendaciones de Mahout son, por tanto, estadísticamente relevantes.

Estas recomendaciones generadas automáticamente nos permiten conectar artículos pertinentes para su descubrimiento inmediato. Ya no hay necesidad de esperar a recomendaciones manuales antes de que dicha conectividad esté disponible.

.

Las siguientes cifras de Infopedia en entorno de ensayo de NLB, ilustran la capacidad de las recomendaciones generadas automáticamente para complementar las recomendaciones manuales de los bibliotecarios. En los casos en los que las recomendaciones manuales no están todavía disponibles, enriquecen el contenido y de inmediato permiten el descubrimiento de otro contenido relacionado.

Figura 2.3 Complementando un artículo de Infopedia con recomendaciones manuales

Figura 2.4 Enriqueciendo un artículo de Infopedia con recomendaciones no manuales

El siguiente conjunto de datos que hemos completado con éxito con un proceso similar, es la colección de memoria personal del portal Singapore Memory (http://singaporememory.sg) y estamos en proceso de incrementar esto, para mejorar las actuales recomendaciones de filtrado basadas en el contenido.

Superando Infopedia y Singapore Memory, buscamos perfeccionarlo y completamos con éxito otra prueba de concepto (PoC) de unos 58.000 artículos de prensa de nuestra colección NewspaperSG.

NewspaperSG (http://newspapers.nl.sg) es el recurso en línea de NLB para periódicos de Singapur y Malasia publicados entre 1831 y 2009. Hay que destacar la inversión de la NLB para digitalizar los periódicos a nivel de artículo, y ahora hay más de 18 millones de artículos

en NewspaperSG

Subir de 2.000 a 58.000 resultó relativamente fácil. Completamos el proceso en aproximadamente 18 horas.

Los resultados demostraron ser muy prometedores. Es interesante señalar que si organizamos los artículos recomendados en un orden cronológico, podemos descubrir la progresión de una noticia y ver cómo se desarrolla la historia.

Figura 2.5 Muestra de resultados de PoC de newspaperSG (58.000 artículos)

Con la exitosa aplicación del análisis de texto en Infopedia, Singapore Memory y un año con 58.000 artículos de periódicos, nos animaron a dar el paso y hacer frente a un conjunto de datos mucho mayor. Decidimos coger todos los artículos de 2 periódicos y generar recomendaciones para ver si podíamos descubrir nuevas perspectivas de una noticia cuando los artículos de otro diario aparecen en las recomendaciones.

Trabajando en los 6 millones de artículos de NewspaperSG de los 2 periódicos (The Straits Times 1845-2009 y The Singapore Free Press 1925-1962) tuvimos nuestro primer problema de escalabilidad real. El tratamiento duró más de una semana antes de que nos quedáramos sin espacio de almacenamiento en los discos.

En el procesamiento de los artículos NewspaperSG también surgió otro tema difícil. Los artículos de prensa se extraen del uso del software de reconocimiento óptico de caracteres (OCR) durante la digitalización de los microfilmes de periódicos históricos. Los errores del OCR son comunes debido a la naturaleza histórica de los periódicos, en particular para los números más antiguos. Estos errores de OCR introdujeron 'ruido' en el conjunto de datos, y también aumentaron significativamente la complejidad de la computación, dando lugar a un procesamiento largo y a la necesidad de una gran cantidad de espacio de almacenaje en discos intermedios.

Se necesita una plataforma de computación estable, extensible y distribuida.

Apache Hadoop es un sistema que soporta el proceso distribuido de grandes conjuntos de datos, a través de un grupo de ordenadores de consumo que utilizan modelos de programación simples (http://hadoop.apache.org). Es la plataforma de software de código abierto de-facto para el procesamiento de cantidades grandes de datos, que cuenta con usuarios como Facebook, Twitter, Yahoo, LinkedIn y EBay.

Muchos de los algoritmos básicos de Apache Mahout se implementan sobre Apache Hadoop usando el mapa/reducir paradigma para permitir la ampliación de los conjuntos de datos razonablemente grandes.

Tras mucha experimentación y ajustes, finalmente establecimos un grupo completo de Apache Hadoop gestionando Apache Mahout en nuestro centro de datos secundario. Hemos creado un total de 13 servidores virtuales en 3 hosts de máquinas virtuales.

Figura 2.6 Configuración de Apache Hadoop en NLB

Para obtener más información sobre nuestra configuración, no dude en contactar con nosotros.

Para afrontar la cuestión de los errores de OCR en los artículos de prensa, sintonizamos los parámetros para el algoritmo de análisis de texto que ignoran las ocurrencias de palabras poco frecuentes. Esto ha resultado en la eliminación de una porción significativa de los errores de OCR antes de que comenzase el verdadero proceso de Mahout.

Con el grupo Apache Hadoop instalado y la aplicación de parámetros de algoritmos más agresivos, hemos sido capaces de generar recomendaciones similares para más de 150.000 artículos. El cambio dinámico en los valores de los parámetros es instrumental al reducir una gran parte del tiempo requerido para el proceso. De ser capaz de procesar 58.000 artículos en alrededor de 18 horas, estamos ahora en condiciones de procesar 150 mil artículos en menos de 30 minutos.

Sin embargo, todavía estamos en proceso de resolver problemas a medida que avanzamos hacia nuestra meta de procesar 6 millones de artículos.

3. PRÓXIMOS PASOS

Actualmente estamos trabajando con Apache Mahout en la prueba de conceptos adicional, específicamente en las áreas de enriquecimiento de contenido, agrupación y clasificación

En el ámbito del enriquecimiento de los contenidos, hay para nosotros un montón de ideas interesantes y posibilidades para explorar y dominar? Esperamos enriquecer el contenido con información semántica para que este quede conectado semánticamente en lugar de sólo textualmente. También esperamos enriquecer el contenido con traducciones de lenguas para explorar las posibilidades de conexión de contenidos en distintos idiomas, teniendo en cuenta que contamos con cuatro idiomas oficiales en Singapur.

En primera línea del grupo, trabajando con memorias personales enviadas por usuarios del portal Singapore Memory, Apache Mahout agrupa las memorias en 43 grupos (Figura 8). Los términos clave entre las memorias dentro de cada grupo pueden ser analizados para identificar el tema clave para los grupos.

Figura 3.1:Tamaños del grupo Singapore Memory Cluster PoC

Figura 3.2: Muestra de la parte superior

Estamos explorando las tecnologías de visualización que permitan a los usuarios navegar a través de las agrupaciones generadas.

Para la clasificación, actualmente estamos cotejando datos de grupos identificados por el usuario. Estos datos servirán de datos de formación que los algoritmos de clasificación utilizan para clasificar automáticamente nuevos envíos de los usuarios en uno de los grupos identificados. Esto nos permitirá reducir drásticamente la cantidad de esfuerzo que se requiere para mantener y actualizar los grupos actuales que se muestran en el portal de Singapore Memory.

El valor de las colecciones de una biblioteca aumenta con el número de conexiones dentro y a través del contenido de las colecciones, un resultado del bien conocido *fenómeno de efecto de red*. Por tanto, la capacidad de conectar de forma automática el contenido es esencial para que las bibliotecas desentierren los valores ocultos en sus colecciones cuidadosamente custodiadas, sobre todo para las colecciones importantes que van más allá de miles de artículos. La tecnología y el software de minería de datos y análisis de texto, ahora son más maduras y de fácil acceso a través de las licencias de código abierto. El momento es propicio para las bibliotecas, para utilizarlos, para resaltar los grandes valores de sus colecciones.

Existen muchas posibilidades que estudiaremos en el futuro para hacer más y mejores conexiones - conexiones que permitan a los usuarios descubrir una información completa, relevante y de confianza.

Nosotros hacemos el trabajo para que nuestros usuarios no lo hagan.