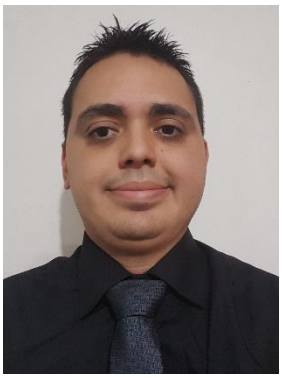
BÚSQUEDAS DISTRIBUIDAS SOBRE REPOSITORIOS HETEROGÉNEOS EN UN AMBIENTE NO-COOPERATIVO



AUTOR

Andrés Marenco Zúñiga

Faculty of Electrical Engineering, Mathematics and Computer Science,
University of Twente
Maestría en Ciencias de la Computación, 2014

Escuela de Computación,
Instituto Tecnológico de Costa Rica
Bachillerato en Ingeniería en Computación, 2010

INVESTIGADORES ASOCIADOS (PROFESORES SUPERVISORES):

Dr.ir. Robin Aly
University of Twente

Dr.ir. Djoerd Hiemstra
University of Twente

RESUMEN DE TESIS

Federated Aggregated Search (http://essay.utwente.nl/66435/)
University of Twente

INTRODUCCIÓN

El paradigma de los motores de búsqueda ha evolucionado aceleradamente con el fin de adaptarse a las exigentes necesidades de información de los usuarios: actualmente no es suficiente con responder simples documentos de texto que se ajusten al criterio de búsqueda, sino que se debe incluir una amplia combinación de archivos heterogéneos tanto en su contenido (noticias, blogs, artículos científicos…) como en su tipo (imágenes, videos, mapas…). Cada tipo de documento, almacenado en repositorios especializados conocidos como “verticales” y físicamente ubicados en diversos puntos geográficos han sido integrados en los modernos buscadores que utilizamos hoy en día.

Sin embargo, esta separación de repositorios y su subsecuente integración crea interesantes retos en el campo de la recuperación de la información: ¿es acaso necesario propagar cada una de las búsquedas del usuario a cada uno de los verticales que forman parte del sistema? Se debe tomar en cuenta que cada llamada a un vertical aumenta el tiempo de búsqueda, y muy probablemente algunos de ellos no contengan la respuesta que el usuario requiere.

Ilustremos con un ejemplo: el criterio de búsqueda "The Beatles" tiene mejores probabilidades de encontrar resultados valiosos en *YouTube* o en *Amazon*, que en fuentes de datos como *LinkedIn* o *Food Network*.

Así, para tener una apropiada selección de los repositorios para una búsqueda en particular, debemos primero contar con una representación lo suficientemente descriptiva de cada uno de los verticales la cual nos permita identificar el tipo de respuestas que podemos encontrar en el mismo (Figura 1). No obstante, la misma naturaleza heterogénea de los documentos y la falta de cooperación de los administradores de los repositorios pueden presentar un obstáculo en la creación de la representación del vertical.

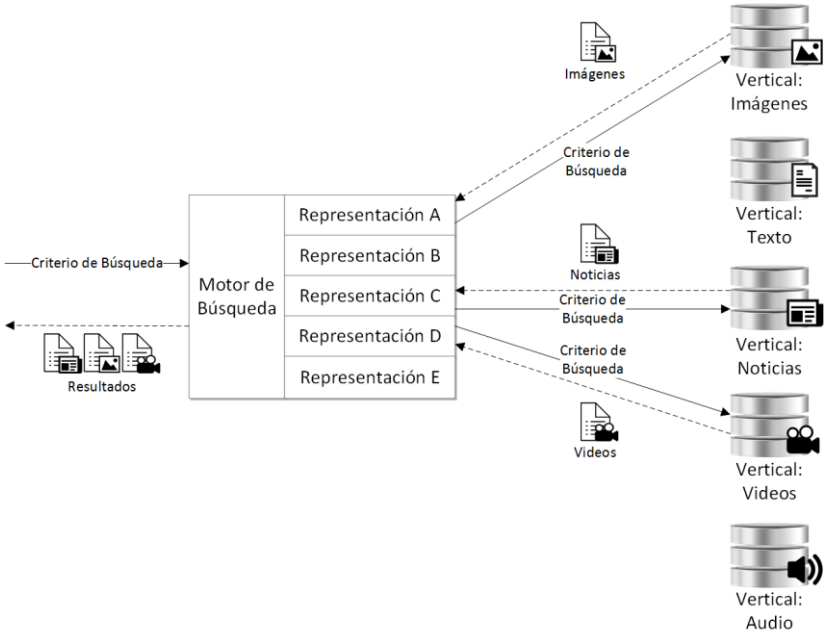


FIGURA 1: BÚSQUEDA DISTRIBUIDA ENTRE VERTICALES

Esta tesis investiga y evalúa el uso de Wikipedia como fuente externa para enriquecer las representaciones textuales de los verticales en un intento por depurar el proceso de selección del repositorio más apropiado para la búsqueda.

METODOLOGÍA

Como es de esperar, la literatura nos expone como diferentes técnicas de muestreo son el camino correcto para obtener una representación inicial de cada repositorio. Para nuestros experimentos, se tomó la colección de datos del TREC FedWeb 2013[1]: una muestra de aproximadamente 1,878,613 documentos, obtenidos de 157 verticales, en los que se incluyen 24 diferentes tipos de contenido (imágenes, noticias, videos…). Para medir la efectividad de nuestra propuesta, el FedWeb también provee una lista de 50 criterios de búsqueda y los verticales más apropiados para cada una de ellos.

Como fue mencionado anteriormente, la heterogeneidad de los documentos en la muestra presenta un gran reto al seleccionar el repositorio adecuado para el usuario. Esto se debe a que los algoritmos tradicionales de búsqueda

[1] https://sites.google.com/site/trecfedweb/2013-track

están parcializados hacia documentos con más texto. Por lo tanto, si el vertical era mayormente de imágenes, su representación textual va a ser muy limitada, y tiende a “perder” contra un vertical con contenido mayormente de noticias o artículos científicos. Igualmente, verticales en los que su representación tenga excesivo texto, tales como foros de opinión, pueden generar “ruido”, el cual es igualmente perjudicial para nuestros propósitos.

Para encontrar un poco más de justicia en la etapa de selección debemos homogenizar la representación: este proceso será un enriquecimiento de la representación de los verticales y se enfocará en identificar cuáles artículos de Wikipedia se relacionan directamente con el contenido del repositorio. Para esto utilizaremos técnicas de modelado estadístico que nos permita categorizar los “temas” que se pueden encontrar en una colección de documentos. En particular nos enfocaremos en la técnica conocida como *Latent Dirchlet Allocation (LDA)*.

Cada uno de los temas del repositorio son simplemente una lista de términos que están relacionados semánticamente (por ejemplo, un tema puede incluir: supernova, emisión, radio, galáctico, física…) y cada repositorio tendrá una cantidad predefinida de temas (ver Figura 2).

T1: networks random systems disordered neural boolean information critical mat phase transition lyapunov cond mutual exponents percolation localization
T2: supernova remnant snr emission shell radio galactic snr thermal evolution ph astro study ray physics
…
T100: algorithm cs problem complexity data structures time log approximation computational distributed graph computing number

FIGURA 2: EJEMPLO DE TEMAS GENERADOS CON LDA. LISTA OBTENIDA AL MODELAR 200 TEMAS DE ARXIV.ORG

Teniendo el modelado completo, obtenemos una descripción más enfocada del vertical. Aunque en primera instancia el posible ruido que tenga la muestra ha sido eliminado y parece razonable utilizar la lista de temas como representación final, esto sería un error: los temas encontrados no nos dicen más que lo dicho por la misma muestra, pero utilizando mucho menos texto, lo cual perjudica el algoritmo de selección.

Por lo tanto, cada uno de los temas modelados (la lista de términos) se utiliza como criterio de búsqueda en Wikipedia. Los primeros resultados se asumen como los artículos más relevantes y por ello, los que mejor describen al repositorio. Tal como se ilustra en la Figura 3, estos artículos de Wikipedia serán la nueva representación del vertical.

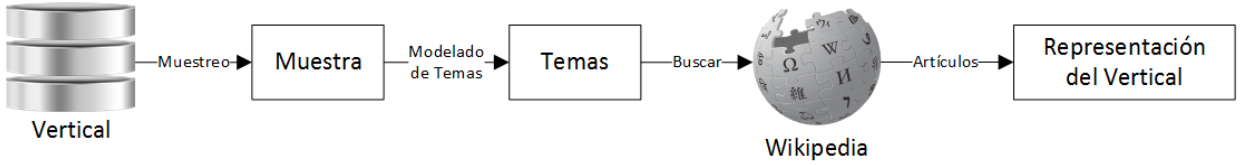


FIGURA 3: WIKIPEDIA COMO REPRESENTACIÓN DEL VERTICAL

Nótese como simplemente estamos cambiando la representación del vertical con el fin de poder mejorar la selección de los mismos. Una vez identificados los más apropiados para la búsqueda, esta se hará directamente en el repositorio como tal. Así, el modelado y la generación de la representación se ejecutará una única vez y será integrada dentro del motor de búsqueda descrito en la Figura 1.

RESULTADOS

Los primeros resultados fueron poco satisfactorios. Basados en la métrica utilizada para la evaluación (*nDCG@20*), nos encontramos como al tomar únicamente el primer artículo de Wikipedia (por cada tema modelado) en todos los repositorios perjudicó el proceso de selección (ver

Tabla 1).

	NDCG@20	TOTAL DE DOCUMENTOS	DOCUMENTOS POR VERTICAL
WIKI	0.12464	28,459	181.26
BASE	0.24244	1,878,613	11,965.68

TABLA 1 REPRESENTACIÓN BASADA EN WIKIPEDIA

Sin embargo, el experimento resulta un poco simplista: un método no debe ser aplicado de igual manera a todos los repositorios, ya que cada uno de ellos presenta características particulares que deben ser tomadas en consideración. Analizando la

Tabla 2 podemos observar como el número de términos en la muestra original de los verticales varía considerablemente. Basado en nuestras premisas, hay 2 grupos que analizar: a) los verticales que tienen pocos términos, ya que su representación es muy limitada con respecto a los otros y esto los hace poco propensos a ser seleccionados, y b) los verticales con demasiados términos, ya que puede existir demasiado ruido y esto los hace muy propensos a ser seleccionados erróneamente.

Número de términos sin repeticiones	*Número de verticales*
0-499,999	23
500,000-999,999	15
1,000,000-1,999,999	33
2,000,000-3,999,999	38
4,000,000-5,999,999	24
6,000,000-7,999,999	16
8,000,000-9,999,999	8

TABLA 2 NÚMERO DE TÉRMINOS SIN REPETICIONES POR VERTICAL

Utilizando los puntos del análisis anterior realizamos dos nuevos experimentos: a) reemplazamos la muestra en los verticales cuyos número de términos sin repeticiones esté dentro del rango de 0 a 499,999 (Tabla 3), y b)

reemplazamos la muestra en los verticales cuyos términos sin repeticiones esté dentro del rango de 8,000,000 a 9,999,999 (Tabla 4).

	NDCG@20
WIKI	0.24244
BASE	0.24244

TABLA 3 APLICACIÓN DEL ALGORITMO EN REPOSITORIOS CON POCOS TÉRMINOS

	NDCG@20
WIKI	0.28131 [2]
BASE	0.24244

TABLA 4 APLICACIÓN DEL ALGORITMO EN REPOSITORIOS CON GRAN CANTIDAD DE TÉRMINOS

CONCLUSIÓN

Los datos soportan la hipótesis de que al usar técnicas de modelado de temas sobre una colección de documentos y asociando la colección a documentos externos resulta en mejoras sobre el proceso de selección: se reducen posibles "ruidos" en la muestra, se ayuda a enfocar el contenido del vertical y se reduce la heterogeneidad entre los diferentes repositorios. Sin embargo, el éxito del algoritmo propuesto depende de variables como la cantidad de términos en la muestra y la correcta aplicación del modelado inicial.

Se propone como trabajo futuro continuar experimentando con dichas variables con el fin de encontrar valores óptimos para nuestro propósito. Igualmente, las características de ciertos repositorios en particular pueden indicarnos que su muestra es lo suficientemente efectiva para una correcta selección y por lo tanto no se debe modificar.

Finalmente, el conjunto de datos utilizados fue el de FedWeb 2013. La utilización de una nueva colección con los mismos métodos aquí descritos sería un paso más en la confirmación de nuestros descubrimientos.

[2] El resultado es significantemente mejor sobre la base ($p < 0.05$)