Content-based recommendation for Academic Expert finding

César Albusac
Dpto. de Ciencias de la Computación e Inteligencia Artificial. ETSI Informática y de Telecomunicación. Universidad de Granada. CITIC-UGR
Granada, España
calbusac@ugr.es

Luis M. de Campos
Dpto. de Ciencias de la Computación e Inteligencia Artificial. ETSI Informática y de Telecomunicación. Universidad de Granada. CITIC-UGR
Granada, España
lci@decsai.ugr.es

Juan M. Fernández-Luna
Dpto. de Ciencias de la Computación e Inteligencia Artificial. ETSI Informática y de Telecomunicación. Universidad de Granada. CITIC-UGR
Granada, España
jhg@decsai.ugr.es

Juan F. Huete
Dpto. de Ciencias de la Computación e Inteligencia Artificial. ETSI Informática y de Telecomunicación. Universidad de Granada. CITIC-UGR
Granada, España
jhg@decsai.ugr.es

ABSTRACT

Nowadays it is more and more frequent that Web users search for professionals in order to find people who can help solve any problem in a given field. This is call expert finding. A particular case is when users are interested in scientific researchers. The associated problem is to get, given a query that expresses a topic of interest for a user, a set of researchers who are expert on it. One of the difficulties to tackle the problem is to indentify the topics in which a professional is expert. In this paper, we face this problem from a content-based recommendatation perspective and we present a method where, starting from the articles published by each researcher, and a query, the expert researchers are obtained. We also present a new document collection, called *PMSC-UGR*, specifically designed for the evaluation in the field of expert finding and document filtering

CCS CONCEPTS

• **Information systems → Information systems applications**; **Retrieval tasks and goals**; **Test collections**; **Relevance assessment**; *Document structure*;

KEYWORDS

Sistema de recomendación, Recomendación basada en contenido, Búsqueda de expertos

ACM Reference Format:
César Albusac, Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete. 2018. Content-based recommendation for Academic Expert finding. In *CERI 18: 5th Spanish Conference in Information Retrieval, June 26–27, 2018, Zaragoza, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3230599.3230607

CERI 18, June 26–27, 2018, Zaragoza, Spain

ACM ISBN 978-1-4503-6543-7/18/06…$15.00
https://doi.org/10.1145/3230599.3230607

1. INTRODUCCIÓN

Una de las tareas que se realizan actualmente en la Web y que está tomando más auge es buscar gente. Podemos hacerlo por múltiples razones: simplemente para satisfacer nuestra curiosidad o por necesidad de contactar con algún tipo de profesional que nos permita solucionar un problema concreto, entre otras. Un ejemplo del primer caso sería acceder a la biografía de un actor para conocer las películas que ha rodado; del segundo podría ser conocer con qué políticos tendríamos que ponernos en contacto para exponerles la idoneidad de que la sanidad pública asumiera un nuevo tratamiento para una enfermedad. Este tipo de búsqueda se encuadra en el contexto de la *Búsqueda de Expertos* [4]. El problema consiste en, dados un grupo de expertos y una consulta normalmente corta, que representa una necesidad de información, devolverle al usuario un conjunto de estos. Aquellos situados en los puestos más altos serían los más relevantes y los que de más valor le resultarían. Desde otra perspectiva, se está abordando un problema de recomendación [8].

Los expertos suelen estar representados por medio de las temáticas o tópicos que definen sus intereses obtenidos a partir diferentes fuentes. Por ejemplo, para un político se podría usar el conjunto de intervenciones en sesiones parlamentarias, o para un científico, el conjunto de trabajos publicados en revistas y congresos. En algunos casos, se emplea el concepto de perfil textual, construido de forma explícita a partir de la unión de los textos de esas fuentes documentales y usado como "documento" que alimentará un Sistema de Recuperación de Información (SRI). En este caso, se ha creado un documento ficticio que resume, de forma general, y por medio de palabras, los intereses del experto. En otros casos, los ítems textuales del profesional se introducen en el SRI de forma diferenciada. No se crea de forma explícita un perfil, aunque todos los documentos del experto se pueden ver como un perfil implícito.

Los autores de este trabajo han realizado un estudio extensivo en los campos de la recomendación basada en contenido [4, 5] y el filtrado de documentos [16][1]. Así, se desarrollaron metodologías de construcción de perfiles de usuarios que ya se aplicarían inicialmente a los problemas de búsqueda de expertos y de filtrado de

[1]Variación del problema de recomendación basada en contenido, donde, al llegar un nuevo documento al sistema, éste debe decidir a qué expertos enviárselo, teniendo en cuenta la temática del mismo y los intereses de los profesionales.

documentos. Por ejemplo, en [10] se diseñaron perfiles de usuario, de diferente tamaño y composición, que permitieron determinar la idoneidad para resolver los problemas entre manos en el contexto parlamentario, donde los expertos eran diputados del Parlamento de Andalucía y las fuentes documentales los diarios de sesiones de dicha cámara autonómica. Se utilizó un SRI para, mediante consultas simuladas de usuarios (normalmente cortas) o mediante la llegada de un nuevo documento al sistema (consultas largas), determinar qué diputados serían los más adecuados para recomendar. En este caso se construyeron perfiles monolíticos, es decir, creados a partir de todas las intervenciones de un mismo diputado. El trabajo [12] se centró en plantear métodos para determinar, de una forma lo más automática posible, el tamaño de los perfiles. En [11], se planteó una segunda alternativa en la construcción de perfiles: subperfiles temáticos. Partiendo del hecho de que los diputados suelen participar en diferentes comisiones, sesiones de menor tamaño que tratan temas concretos, cada experto tendría asociado varios perfiles, uno por temática. Esta alternativa presentaba un mejor rendimiento con respecto a los monolíticos. En el último trabajo que se ha sometido, [13], se optó por construir dichos subperfiles aplicando métodos de aprendizaje automático, más concretamente de agrupamiento. Los experimentos muestran que es una forma más precisa de construirlos. Como resultado de estas investigaciones, se ha comprobado que cuando se consideran perfiles monolíticos, las consultas adecuadas para tener un mejor rendimiento deben ser largas; por el contrario, con subperfiles, las mejores consultas deben ser cortas.

En este trabajo planteamos una variación de este problema en el que los expertos son científicos y las fuentes documentales los artículos científicos que han escrito en revistas o congresos.

El problema de la recomendación de autores, tiene una característica principal que lo configura como un problema interesante, diferente al hasta ahora abordado de la recomendación de políticos: en la recomendación de políticos, los documentos son entidades independientes entre sí. Sin embargo, los artículos científicos representan un ecosistema donde cada documento se interrelaciona con los demás, de forma explícita a través de las referencias bibliográficas, que se pueden considerar como evidencias explícitas de un interés. Pensamos que su explotación directa puede ser de mucho provecho en el futuro.

El objetivo de esta investigación es comprobar el rendimiento de una aproximación simple basada en técnicas de RI con objeto de conocer las peculiaridades del problema entre manos y establecer métodos *baselines* para próximas investigaciones. Así, en una primera aproximación a este problema, en este trabajo no se van a construir perfiles de usuario de los autores, sino que se considerarán de forma independiente todos los artículos y a partir de ellos y de una consulta se obtendrá una ordenación de autores dispuesta de forma decreciente según el grado de relevancia (modelo que se ha visto que funciona mejor [4]).

Otra contribución importante de este trabajo es la presentación de la colección *PMSC-UGR*, la cual ha sido generada a partir de artículos científicos descargados de PubMed y Scopus para tareas de búsqueda de expertos y filtrado de documentos. Analizando las colecciones de prueba existentes en estos campos de investigación se detectaron ciertas limitaciones en ellas, por lo que se decidió crear un corpus con un número elevado de artículos y autores, identificados estos últimos de forma unívoca, conteniendo además las citas de los trabajos. Estas características configuran a esta fuente documental como una opción bastante interesante para evaluar modelos y técnicas.

Con objeto de mostrar este estudio preliminar sobre búsqueda de expertos, este trabajo quedará organizado de la siguiente forma: la Sección 2 describirá el proceso de recomendación de autores. La Sección 3 presentará la nueva colección de prueba para evaluar técnicas en este campo. La Sección 4 será la encargada de mostrar el diseño experimental, así como analizar los resultados obtenidos en la experimentación. En la Sección 5 se hará una breve revisión de algunos trabajos que tienen relación con este que se presenta, y por último, finalizaremos este trabajo con la Sección 6 en donde se expondrán las conclusiones y se esbozarán las próximas líneas de trabajo.

2. RECOMENDACIÓN DE AUTORES

La aproximación que se ha seguido para resolver inicialmente el problema descrito en la introducción está basada en RI y tiene como objetivo servir como referencia para posteriores estudios. Partimos de una fuente documental de artículos científicos y el objetivo es devolver una lista de autores que sean relevantes a una consulta formulada por el usuario.

El primer paso, una vez que se dispone de la colección documental, será la indexación de los artículos por parte del sistema de RI. Seguidamente, se procederá a realizar la recuperación propiamente dicha: dada una consulta sometida al sistema de RI, éste devolverá una ordenación de artículos de forma decreciente según su grado de relevancia. En este *ranking*, cada artículo se replicará tantas veces como autores tenga, asociándole el identificador de cada uno de ellos. Podríamos decir que la nueva ordenación quedaría formada ahora por pares $(idAutor_i, idArticulo_j)$, cada uno con el mismo grado de relevancia que se obtuvo para el $idArticulo_j$. Pero, como nuestro objetivo es recomendar autores, hay que convertir el *ranking* de pares, donde pueden aparecer los mismos autores varias veces en pares distintos, en uno formado exclusivamente por autores, por lo que se tiene que aplicar un método de combinación / fusión de resultados para asignar un nuevo grado de relevancia a cada autor y proceder a reordenar los autores. En este trabajo se han probado los métodos *CombMax*, *CombSum* y *CombLogDCS*, presentados en [11], y que pasamos brevemente a describir:

- *CombMax*: dado un autor idAutor, se selecciona el par $(idAutor, idArticulo_j)$ más alto en el *ranking*, es decir, aquel con máximo grado de relevancia, y se asocia dicho valor al autor.
- *CombSum*: dado un autor idAutor, se le calcula un nuevo grado de relevancia sumando todos los valores de los pares $(idAutor, idArticulo_j)$ que aparecen en la ordenación.
- *CombLgDCS*: esta estrategia calcula un único grado de relevancia para un autor concreto, idAutor, agregando los valores de los pares $(idAutor, idArticulo_j)$, pero devaluados logaritmicamente considerando sus posiciones en la ordenación. Así, se penaliza más a aquellos que estén en las posiciones más bajas.

3. LA COLECCIÓN DE PRUEBA: *PMSC-UGR*

El corpus documental *PMSC-UGR*[2] ha sido confeccionado por los autores de este trabajo a partir de un subconjunto amplio de artículos científicos de MEDLINE/PubMed, en inglés, con objeto de que sea usada principalmente en evaluaciones relacionadas con la búsqueda de expertos y el filtrado de documentos.

El conjunto de datos contiene originalmente un total de 26.759.991 artículos en inglés, los cuales ofrecen información sobre el identificador del trabajo, PubMedId, el título, revista y datos de publicación, resumen, autores, incluyendo un identificador único de cada uno de ellos, y las listas de palabras clave y de descriptores del tesauro MeSH.

Dado que se pretende usar para la recomendación de expertos, en la fase de construcción de la misma se prestó especial atención a la desambiguación de autores, con objeto de tener una colección libre de problemas con los nombres de los creadores de los trabajos y para tal fin se empleó el identificador Open Research and Contributor ID (ORCID). Así, se tuvieron en cuenta sólo aquellos trabajos con autores que contaban con él en los registros de MEDLINE. El problema que se planteaba a continuación era que se podían descartar artículos de autores en donde no figuraban sus ORCIDs, normalmente más antiguos, y sólo considerar los artículos con esos mismos autores donde sí estaban incluidos, normalmente más modernos. Este hecho hacía que se redujera considerablemente el tamaño de la colección, por lo que se pensó en emplear Scopus como fuente para resolver autorías: se asoció el ORCID de cada autor con el identificador de Scopus y se realizó una búsqueda de artículos por autor. El resultado se confrontó con los trabajos incluidos en PubMed y se seleccionaron todos aquellos que aparecían en el subconjunto descargado de esta biblioteca digital.

Para completar la colección, se añadieron citas, las cuales no están disponibles en MEDLINE/PubMed. Y para tal fin se volvió a echar mano de Scopus. En este caso, y para cada trabajo de la colección, se buscaron los artículos que en sus referencias lo incluían. Tras comprobar que los trabajos que lo citaban estaban en el subconjunto de MEDLINE que estábamos manejando, se incluían como citas, incorporando a su vez los ORCIDs de sus autores. El 66,78 % de los artículos de la colección contiene citas.

A modo de resumen, en la Tabla 1 presentamos los tamaños de la colección de prueba creada, en cuanto a número de artículos, autores únicos y citas se refiere y número medio de citas por artículo y referencias por artículo.

# Artículos	# Autores	# Citas	# medio Autores/Artículo
762.508	20.406	3.593.931	1,1
# medio Citas/Artículo		# medio Referencias/Artículo	
7,06		6,10	

Cuadro 1: Principales datos sobre la colección *PMSC-UGR*

Como se puede apreciar en la imagen 1, que representa el número de trabajos por autor, la gran mayoría tienen un número muy bajo de autores y muy pocos artículos tienen un número elevado.

En las colecciones de prueba clásicas, además del conjunto de documentos se suministra una relación de consultas y el subconjunto de documentos que son relevantes para cada una de ellas. El principal problema de la creación de estas colecciones es el alto costo que conlleva generar los juicios de relevancia, por lo que han surgido las denominadas Pseudo Colecciones de Prueba (*Pseudo Test Collections*), en donde, de forma automática, se generan las consultas o se determinan los documentos relevantes para consultas dadas. En este sentido, la colección *PMSC-UGR* no contiene un conjunto explícito de consultas, ni consecuentemente la especificación de los documentos relevantes, sino que las consultas pueden modelarse a partir de cualquier elemento estructural de un artículo, es decir, su título, resumen, palabras clave o materias de un tesauro.

De esta forma, el conjunto de autores relevantes a dicha consulta, el *gold standard*, sería directamente el formado por los firmantes del mismo. Incluso este conjunto podría extenderse con el grupo de autores que, desde otros artículos, citan al original, los cuales también podríamos decir que son expertos relacionados. Asumimos que los juicios de relevancia generados de esta forma no son exhaustivos ya que puede haber más autores a los cuales les pudiera interesar, además de los propios autores y los autores citadores. De cualquier forma, en [9, 20] se indica que esta situación en la que los juicios no son completos no es un problema para comparar diferentes sistemas siempre y cuando el número de consultas sea grande, como es el caso de esta colección descrita.

En la literatura especializada se encuentran enfoques basados en la generación automática de consultas a partir de un conjunto de documentos conocidos como salida de un buscador [2]; a partir de los textos de los enlaces a documentos Web (los documentos a los que apuntan los enlaces son relevantes al texto de los mismos) [1]; considerando *hashtags*, los tuits indexados por ellos [6] serían relevantes; en colecciones documentales anotadas por medio de palabras clave o materias de un tesauro, los documentos anotados bajo una palabra clave serían relevantes a ellos [7]. Alternativamente, también se pueden encontrar estudios que tienen como objetivo encontrar los documentos que satisfacen una consulta estudiando los clics realizados sobre una ordenación de documentos, como por ejemplo, el presentado en [17], o el más reciente [21].

En nuestro caso, por tanto, asumimos que el conjunto de documentos relevantes de cada consulta (una o varias partes de un artículo dado) son sus propios autores. Por tanto, son juicios implícitos que no hay que inferir ni generar de ninguna forma.

Centrándonos en el contexto del corpus, existen varias colecciones apropiadas para experimentar en el marco de la búsqueda de expertos. Por ejemplo, LExR [19], ArnetMiner [23], CERC [3] y W3C [15]. Las principales diferencias de *PMSC-UGR* con respecto a ellas son que es bastante más grande que W3C y CERC, el número de consultas es mayor que el de cualquiera de las anteriores (asumiendo una división típica del 80 % de documentos en entrenamiento y 20 % de prueba, nos encontramos con más de 150.000 consultas), además incluye citas (sólo ArnetMiner las incluye también) y que no hay lugar a duda sobre la atribución de autores a trabajos por el estricto proceso de identificación de autores que se ha realizado en *PMSC-UGR* para su creación.

[2]Esta colección puede descargarse del enlace http://irutai2.ugr.es/PMSC-UGR

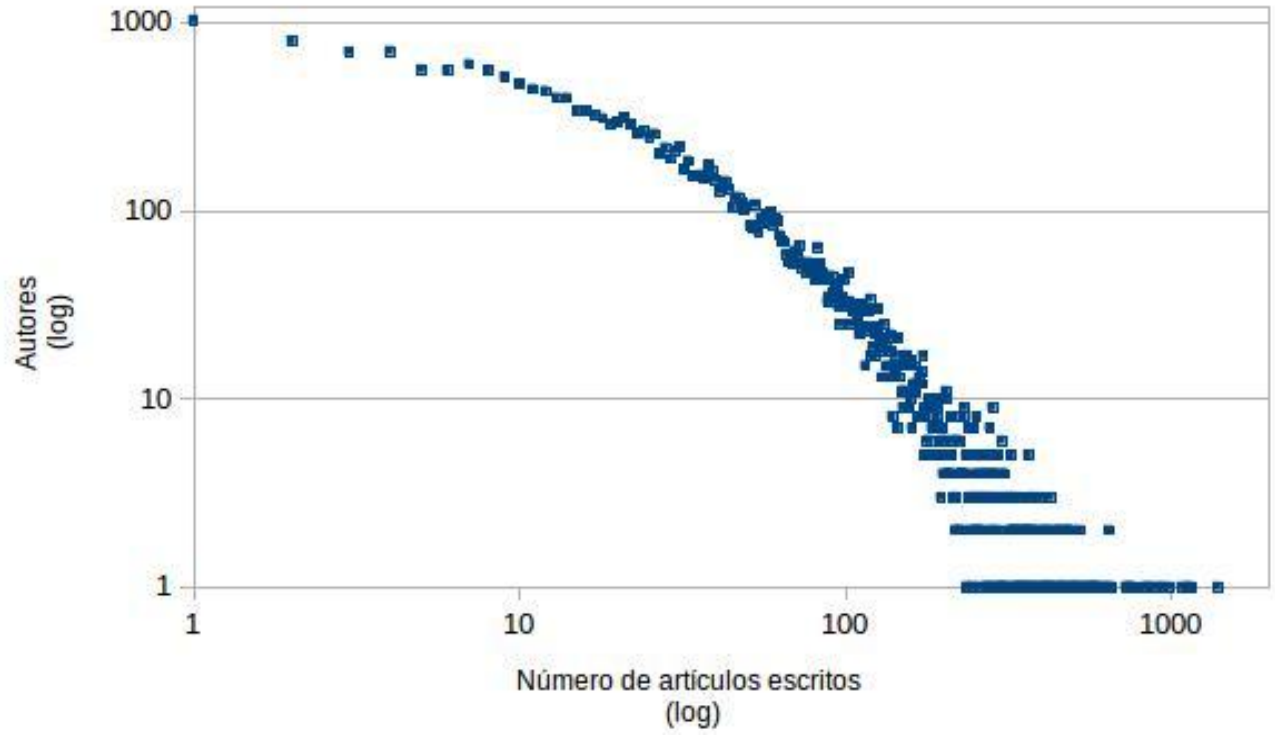

Figura 1: Representación gráfica del número de autores por trabajo

4. EVALUACIÓN

En esta sección describiremos los objetivos de la evaluación y plantearemos el diseño experimental que se ha realizado, así como los resultados obtenidos.

4.1. Objetivos

La presente evaluación tiene como objetivo general plantear unos experimentos preliminares que determinen si la aproximación presentada es de interés y que establezca las bases para futuras líneas de trabajo en el campo de recomendación de autores de trabajos científicos, considerando como base la nueva colección de prueba con que se han llevado a cabo.

De manera más específica, estas serán las preguntas de investigación que nos plantearemos:

- PI1: ¿Cuál es la mejor forma de representar los intereses de los autores?
- PI2: ¿Qué parte de un artículo tiene mayor capacidad de recuperar autores?
- PI3: ¿Qué combinación de los dos anteriores es más adecuada?
- PI4: ¿Que método de combinación de pares (idAutor, idArtículo) consigue un mejor rendimiento?
- PI5: ¿Se pueden considerar los citadores como juicios de relevancia?

4.2. Diseño experimental

Se ha particionado la colección de artículos en 80 % para el entrenamiento (610.006 artículos con 676.715 autores) y el 20 % restante para prueba (152.502 artículos con 168.956 autores). El número medio de autores por trabajo, tanto en entrenamiento como en prueba es 1,1. También se ha llevado a cabo la eliminación de palabras vacías y *stemming* para el idioma inglés.

Lucene ha sido la biblioteca empleada para desarrollar el sistema de recomendación de autores y su implementación del Modelo Espacio Vectorial el método de recuperación usado. Aprovechando la información sobre la estructura de un artículo, se han indexado en campos separados, para cada artículo $idArticulo_j$), su título, resumen, palabras clave y materias de MeSH (PCyM). Además, se ha creado un campo con la unión del texto de los tres anteriores. Esta creación de campos en Lucene es equivalente a representar el perfil de un autor, a partir de la estructura de los artículos, de diferentes formas. En particular, la unión no considera estructura ninguna.

En cuanto a las consultas, considerando que normalmente las que realiza un usuario en este contexto suelen ser cortas, se ha optado por simularlas mediante los siguientes elementos estructurales de cada artículo del conjunto de prueba: título, resumen, palabras clave y materias de MeSH (cada una separadamente) y la unión en una única consulta de los textos de las tres anteriores. Cada tipo de consulta se ha sometido contra el campo de Lucene análogo, es decir, consultando por título, por ejemplo, se ha utilizado el campo título para recuperar (Consulta Título → Campo Título), pero también

se ha lanzado contra el campo que engloba el título, el resumen y la lista de palabras clave y materias (Consulta Título → Campo U-TRPCyM). En este sentido y considerando la posibilidad que ofrece Lucene de consultar por campos, se ha probado una variación de la consulta completa, de tipo estructurado, que consultaría cada elemento del artículo en el campo correspondiente y quedarían unidas todas esas subconsultas por el operador lógico OR: ¨ *"Campo Título: Consulta Título" OR "Campo Resumen: Consulta Resumen" OR "Campo PCyM: Consulta PCyM"*. Igualmente se ha probado el mismo tipo de consulta estructurada pero con el operador AND.

Como se ha indicado en la Sección 2, se han empleado los métodos *CombMax*, *CombSum* y *CombLogDCS*. Para los dos últimos, no se ha considerado la ordenación completa sino los 250 primeros resultados.

Como juicios de relevancia, para cada consulta, es decir, para cada artículo del conjunto de prueba, se han considerado dos conjuntos de autores relevantes: por un lado, los propios autores del trabajo, y por otro, los mismos autores más aquellos que han citado a ese artículo, es decir, los autores citadores. De esta forma se estaría considerando que el hecho de que un experto cite un trabajo automáticamente lo convertiría también en relevante de ese artículo. La idea es determinar si tomando unos u otros el comportamiento del sistema cambia y se pueden o no considerar a los citadores como juicios de relevancia válidos para este problema.

En resumen, la metodología de experimentación que se ha llevado a cabo es la siguiente:

- Particiones: 80 % de artículos en entrenamiento, 20 % en prueba.
- Biblioteca de RI usada: Lucene.
- Indexación: documentos representados por los artículos, asociándoles la información textual de los elementos de los artículos en los siguientes campos de Lucene:
  - Título (T),
  - resumen (R),
  - palabras clave y materias (PCyM),
  - la unión del texto de todas estas unidades (U-TRPCyM) y
  - todos los campos individuales para consultas estructuradas (E-TRPCyM).
- Modelo de recuperación: Espacio Vectorial.
- Consultas:
  - Título (T),
  - resumen (R),
  - palabras clave y materias (PCyM),
  - la unión del texto de todas estas unidades (U-TRPCyM),
  - Estructurada con AND (E-AND)y
  - Estructurada con OR (E-OR).
- Combinación de consultas → campos:
  - Título → Título,
  - Resumen → Resumen,
  - PCyM → PCyM,
  - U-TRPCyM → U-TRPCyM
  - Título → U-TRPCyM,
  - Resumen → U-TRPCyM,
  - PCyM → U-TRPCyM,
  - E-AND → E-TRPCyM,
  - E-OR → E-TRPCyM.
- Método de combinación de ordenaciones: CombMax, CombSum y CombLgDCS.
- Juicios de relevancia:
  - Autores de cada artículo sometido como consulta y
  - Autores + Citadores (autores de los trabajos que lo citan).

4.3. Medidas de rendimiento

Para evaluar el rendimiento de las diferentes alternativas, se ha empleado el ampliamente conocido programa *trec_eval* [3] y de las medidas de evaluación que calcula, dados una ordenación y un conjunto de documentos relevantes, para esta experimentación hemos seleccionado las ampliamente conocidas Exhaustividad (*Recall – R*), Precisión (P), como medidas que determinan la calidad de recuperación, y RPrecisión (RPrec, precisión al número de relevantes) y *Normalized Discounted Cumulative Gain* (NDCG), centradas en medir la calidad del *ranking*. Concretamente, nos fijaremos en los 10 primeros autores de la ordenación para *Recall*, R@10, y Precisión, P@10, y en el *ranking* completo para RPrec y NDCG. Consideramos que estas dos medidas son especialmente interesantes en un contexto de búsqueda de expertos ya que el sistema debería devolver el mayor número de autores relevantes colocados en las posiciones más altas de la ordenación.

4.4. Resultados

En las Tablas 2 y 3 se presentan los resultados para las cuatro medidas de evaluación que hemos probado y para los Autores como juicios de relevancia y Autores + Citadores, respectivamente. En ellas podemos apreciar las diferentes combinaciones entre consultas y campos del índice de Lucene, así como los valores para la correspondiente medida considerando los tres métodos para combinar autores.

[PI1:] ¿Cuál es la mejor forma de representar los intereses de los autores?

Comparando los valores de las medidas de rendimiento para el campo que contiene la unión de los textos de los elementos estructurales de un artículo (U-TRPCyM) con respecto a cada uno de ellos individualmente, podemos destacar que no hay prácticamente diferencia cuando lo comparamos con el resumen y con las palabras clave y las materias, pero sí existe negativamente con los títulos, lo cual indica que éstos contienen un texto escaso y en muchas ocasiones poco representativo del contenido del documento, por lo que son poco útiles para describir el contenido del artículo. Por tanto, podemos concluir que la mejor forma de expresar los intereses de un autor, y de forma independiente de la medida de rendimiento empleada, es aquella que contiene todo el texto o, al menos, el resumen de sus trabajos.

[PI2:] ¿Qué parte de un artículo tiene mayor capacidad de recuperar autores?

Analizando todas las tablas, en prácticamente todos los casos la consulta formada por la unión de los elementos estructurales del artículo se configura como el tipo de consulta que mejores resultados obtiene con respecto al resto de unidades. Básicamente, cuanta más

[3] http://trec.nist.gov/trec eval/

P@10		Combinación		
Consulta	**Campo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,0541	0,0583	0,0590
	U-TRPCyM	0,0635	0,0658	0,0679
Resumen	Resumen	0,0778	0,0778	0,0810
	U-TRPCyM	0,0784	0,0778	0,0813
PcyM	PcyM	0,0613	0,0638	0,0665
	U-TRPCyM	0,0608	0,0643	0,0666
U-TRPCyM	U-TRPCyM	0,0789	0,0777	**0,0816**
E-OR	E-TRPCyM	0,0759	0,0759	0,0791
E-AND		0,0736	0,0726	0,0759

P@10		Combinación		
Consulta	**Campo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,0826	0,0915	0,0925
	U-TRPCyM	0,1046	0,1097	0,1141
Resumen	Resumen	0,1301	0,1305	0,1378
	U-TRPCyM	0,1323	0,1312	**0,1392**
PCyM	PCyM	0,0998	0,1062	0,1107
	U-TRPCyM	0,0981	0,107	0,1104
U-TRPCyM	U-TRPCyM	0,1330	0,1303	0,1387
E-OR	TRPCyM	0,1250	0,1258	0,1323
E-AND		0,1218	0,1207	0,1277

R@10		Combinación		
Consulta	**Campo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,491	0,5294	0,5354
	U-TRPCyM	0,5758	0,5961	0,6152
Resumen	Resumen	0,7029	0,7018	0,7305
	U-TRPCyM	0,7078	0,7019	0,7334
PCyM	PCyM	0,5581	0,5807	0,6045
	U-TRPCyM	0,554	0,5852	0,6055
U-TRPCyM	U-TRPCyM	0,7147	0,7043	**0,7389**
E-OR	E-TRPCyM	0,6878	0,6883	0,7163
E-AND		0,6663	0,6575	0,6875

R@10		Combinación		
Consulta	**Objetivo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,3666	0,3959	0,4017
	U-TRPCyM	0,4438	0,4567	0,4761
Resumen	Resumen	0,5388	0,5331	0,5626
	U-TRPCyM	0,5453	0,5348	0,5666
PCyM	PCyM	0,4175	0,4333	0,4540
	U-TRPCyM	0,4131	0,4374	0,4546
U-TRPCyM	U-TRPCyM	0,5610	0,5463	**0,581**
E-OR	TRPCyM	0,5336	0,5312	0,5581
E-AND		0,5126	0,5023	0,5309

RPrec		Combinación		
Consulta	**Campo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,2681	0,2607	0,2913
	U-TRPCyM	0,3191	0,2880	0,3501
Resumen	Resumen	0,4708	0,378	0,4901
	U-TRPCyM	0,4731	0,3724	**0,4905**
PCyM	PCyM	0,2818	0,2592	0,3194
	U-TRPCyM	0,2637	0,2602	0,3104
U-TRPCyM	U-TRPCyM	0,4726	0,3608	0,4873
E-OR	E-TRPCyM	0,4388	0,3508	0,4593
E-AND		0,4285	0,3367	0,4452

RPrec		Combinación		
Consulta	**Objetivo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,2278	0,2254	0,2486
	U-TRPCyM	0,2796	0,2549	0,3017
Resumen	Resumen	0,3856	0,3202	0,3986
	U-TRPCyM	0,3899	0,3174	0,4009
PCyM	PCyM	0,2466	0,2318	0,2747
	U-TRPCyM	0,2342	0,2331	0,2682
U-TRPCyM	U-TRPCyM	0,394	0,3147	**0,4022**
E-OR	TRPCyM	0,3655	0,3054	0,3799
E-AND		0,3550	0,2918	0,3664

NDCG		Combinación		
Consulta	**Campo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,4269	0,4358	0,4545
	U-TRPCyM	0,4917	0,4823	0,5216
Resumen	Resumen	0,6229	0,5756	0,6428
	U-TRPCyM	0,6264	0,5731	0,6444
PcyM	PcyM	0,4686	0,4643	0,504
	U-TRPCyM	0,4587	0,4668	0,5005
U-TRPCyM	U-TRPCyM	0,6292	0,5683	**0,6455**
E-OR	E-TRPCyM	0,6009	0,5562	0,6219
E-AND		0,5825	0,5330	0,5996

NDCG		Combinación		
Consulta	**Objetivo**	**CombMax**	**CombSum**	**CombLgDCS**
Título	Título	0,3837	0,3915	0,4066
	U-TRPCyM	0,4558	0,4465	0,4797
Resumen	Resumen	0,5642	0,5253	0,5798
	U-TRPCyM	0,5698	0,5254	0,5835
PCyM	PCyM	0,4304	0,4273	0,4589
	U-TRPCyM	0,4225	0,4302	0,4561
U-TRPCyM	U-TRPCyM	0,5774	0,5277	**0,5896**
E-OR	TRPCyM	0,5488	0,5135	0,5659
E-AND		0,5284	0,4889	0,5419

Cuadro 2: Tablas de resultados para las cuatro medidas de rendimiento y los autores como relevantes

Cuadro 3: Tablas de resultados para las cuatro medidas de rendimiento y los autores y citadores como relevantes

información se emplee para consultar, mejor. El segundo sería el resumen, normalmente muy similar en rendimiento, seguidamente la unión de palabras clave y materias. Y el peor, el título. Claramente se aprecia que las medidas de rendimiento reducen los valores conforme se va haciendo más pequeña la consulta. Podríamos decir que frente a una consulta basada en el resumen, al incorporar el título y las palabras clave y las materias no se mejora considerablemente. El resumen tiene, por tanto, un poder recuperador alto y los otros dos elementos, quizá porque haya algún término con un poder discriminador alto, añaden algunos autores relevantes más al listado final. En el caso de las materias, consideramos que son buenos elementos para describir el contenido de un artículo, pero no tanto para describir autores.

Si nos fijamos seguidamente en las consultas estructuradas conectadas por OR y AND, respectivamente, el rendimiento es menor que el de la consulta unión de los tres elementos estructurales del

artículo y que el del resumen de forma individual. Sí superan a las consultas formadas por el título y por palabras clave y materias. La expresión conectada por el operador OR es siempre mejor que la del AND. Aquí también se intuye un peso considerable del resumen en los resultados de estas consultas estructuradas.

Como conclusión podríamos decir que el usar todo el texto de los elementos del artículo como consulta es la mejor opción y en su defecto, sólo el resumen es igualmente competitivo.

[PI3:] ¿Qué combinación de los dos anteriores es más adecuada?

Claramente el uso del resumen o la unión de todos los textos de los elementos estructurales del artículo como consultas en combinación con el campo de Lucene que aglutina el texto completo de los elementos estructurales ofrece el máximo rendimiento, prácticamente sin diferencias entre ellos. En general también, la opción Título → Título es la menos recomendable.

[PI4:] ¿Que método de combinación de pares (idAutor, idArtículo) consigue un mejor rendimiento?

En todos los casos, el método *CombLgDCS* es el que ofrece mejores resultados para identificar autores. El integrar en la medida, para cada autor concreto, las posiciones de los pares $(idAutor_i, idArticulo_j)$ es una opción mucho más precisa que las dos basadas simplemente en el máximo y en la suma. Cabe destacar que con el conjunto de relevantes formado por autores se alcanza un valor de R@10 de 0,73. Esto indica que prácticamente se recupera el 75 % de los relevantes entre los diez primeros, medida bastante alta. En el caso de autores más citadores, el valor baja al 0,58, pero sigue siendo bastante aceptable. Por otro lado, y fijándonos en la RPrec, se alcanza con sólo autores un valor 0,49, lo que indica que se alcanza a recuperar la mitad de autores relevantes en una ordenación cuya longitud es el número de éstos, valor también bastante interesante.

En cuanto a CombMax y CombSum, su rendimiento es muy parecido y las diferencias entre ellas son realmente bajas en P@10 y R@10. En la RPrec y NDCG se observa que el máximo es generalmente mejor y que las diferencias entre ambas son mayores. Esto implica que CombMax y CombSum recuperan el mismo conjunto de relevantes, pero el primero los sitúa en posiciones más altas que el segundo.

[PI5:] ¿Se pueden considerar los citadores como juicios de relevancia?

Como cabía esperar, en el momento en que se incrementa considerablemente el número de autores relevantes al añadir los citadores (se pasa de 168.956 autores relevantes para todas las consultas de la partición de prueba a 584.320 al añadir los autores citadores), las medidas R@10, RPrec y NDCG decrecen, mientras que la precisión, que siempre se mantiene muy baja, se incrementa levemente. A parte de este comportamiento normal, el sistema de recomendación mantiene la misma tendencia en los resultados, por lo que se puede considerar que los citadores pueden actuar también como juicios de relevancia.

En este sentido, y con la idea de validar la capacidad del modelo para identificar otros autores que están interesados en la temática (o desde otro punto de vista, analizar en qué grado las citas pueden considerarse como juicios de relevancia) hemos hecho un estudio más detallado de los resultados considerando su rendimiento únicamente sobre el conjunto de citadores, centrándonos en U-TRPCyM → U-TRPCyM. En el estudio obtenemos que el modelo es capaz de recuperar el 55 % de los autores que han mostrado su interés en el trabajo (sobre citas validadas) y que estas aparecen en posiciones relativamente altas del *ranking* (RPrec de 0,40). Además, es de destacar que no existen diferencias entre el comportamiento del modelo considerando únicamente autoría como juicios de relevancia o añadir también los citadores (mas allá de los valores puntuales de las métricas). De todo esto se puede concluir que autoría y citadores constituyen un conjunto homogéneo y razonable de juicios de relevancia, proporcionando estas últimas una evidencia externa sobre el interés de un autor en un determinado trabajo.

5. TRABAJOS RELACIONADOS

Uno de los enfoques más habituales para abordar la recomendación de autores es la utilización de métodos de análisis de enlaces para combinar la relevancia de contenido de un autor a partir de una consulta con su importancia en la red de autores generada de alguna forma a partir de citas o referencias. Ahora bien, en el contexto de este trabajo, vamos a comentar sólo aquellas aproximaciones que están cercanas a la nuestra, y que la mayoría de las veces se emplean como *baselines*. En el trabajo [14], se pone en práctica una técnica básica de recomendación de autores de mensajes de correo electrónico en dos fases: en primer lugar, se recuperan mensajes dada una consulta, y seguidamente se genera una ordenación de autores a partir del número de mensajes de cada uno que aparecen en la ordenación. Este mismo método también se emplea como *baseline* en [18]. Con el mismo fin, pero ya en el contexto de búsqueda de expertos, en [22] se ordenan los documentos asociados a expertos y luego se genera una ordenación de estos a partir de los grados de relevancia de sus documentos por medio de una suma. Con una aproximación similar a la construcción de perfiles de autores, en [24] se genera para cada autor un macrodocumento uniendo todos los documentos asociados. Mediante un SRI se obtiene un ranking de autores.

Como se puede observar, los enfoques anteriormente citados y el presentado en este trabajo tienen una base común. Las diferencias fundamentales radican en la forma de llegar a la ordenación de autores, pues nosotros además de la suma empleamos dos métodos de combinación más, y en el manejo de artículos, pues se consideran individualmente y no de forma global.

6. CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo se ha presentado la colección *PMSC-UGR* para realizar evaluaciones en el marco de la búsqueda de expertos y el filtrado documental. Es una colección con un volumen alto de artículos, a los que se les ha añadido sus citas. Además, los autores están correctamente identificados por medio de su ORCID.

También se ha propuesto un método básico para realizar recomendación por contenido de autores basándose en las publicaciones científicas de las que son autores. Cada par (idAutor, idArtículo) es indexado en Lucene y dada una consulta se obtiene una ordenación de pares que luego será convertida a una ordenación final de autores. Se ha probado con diferentes tipos de consultas y campos

de indexación, así como métodos de combinación de resultados. La principal conclusión que obtenemos es que se alcanzan unos valores de rendimiento interesantes fundamentalmente cuando en la consulta está contenido el resumen del artículo y cuando se indexa la unión de los textos de los elementos estructurales del artículo, y además se combina con el método CombLgDCS.

Una vez realizado este estudio preliminar que nos deja un conjunto de valores que se podrán emplear en el futuro como valores de rendimiento mínimo deseados, se nos plantean varias líneas de investigación principales en este campo de búsqueda de expertos académicos:

- Incorporar el análisis de enlaces, por ejemplo empleando el algoritmo PageRank, a partir del grafo de citas de autores o de artículos para calcular la relevancia de cada autor y cada artículo, respectivamente, y poder incorporarla al proceso de recomendación.
- Considerar, para cada autor, las publicaciones del conjunto de autores que lo citan y/o que cita para realizar la recomendación. Este método añadiría al perfil implícito del autor perfiles implícitos y extendidos de autores parecidos que podrían incluirse también en el mecanismo de búsqueda de expertos.
- Calcular medidas bibliométricas de cada autor y diseñar métodos de recomendación que las tuvieran en cuenta.

Apoyándonos en la colección *PMSC-UGR*, aparecen también nuevos campos de trabajo. Por ejemplo, citamos dos de ellos:

- Desarrollar un sistema de recomendación de revistas científicas que los autores podrían usar para determinar dónde publicar según la temática del artículo.
- Aunque los autores de esta colección están identificados de forma unívoca mediante su ORCID, se podrían explorar métodos de desambiguación de autores con el mismo nombre a partir de sus trabajos.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad mediante el proyecto de investigación TIN2016-77902-C3-2-P y los Fondos Europeos de Desarrollo Regional (FEDER).

REFERENCIAS

[1] N. Asadi, D. Metzler, T. Elsayed, J. Lin Pseudo test collections for learning web search ranking functions. SIGIR, 1073–1082, 2011.
[2] L. Azzopardi,M. de Rijke, K. Balog Building simulated queries for known-item topics: An analysis using six European languages. SIGIR 2007, 455–462, 2007.
[3] P. Bailey, N. Craswell, I. Soboroff, A.P. de Vries. The CSIRO enterprise search collection. SIGIR Forum 41(2):42-45, 2007.
[4] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, L. Si. Expertise Retrieval. Foundations and Trends in Information Retrieval 6(2-3):127-256, 2012.
[5] J. Beel, B. Gipp, S. Langer, C. Breitinger. Research-paper recommender systems: a literature survey. International Journal on Digital Libraries 17(49):305-338, 2016.
[6] R. Berendsen, M. Tsagkias, W. Weerkamp, M. de Rijke. Pseudo test collections for training and tuning microblog rankers. SIGIR, 2013
[7] R. Berendsen, M. Tsagkias, M. de Rijke, E. Meij. Generating Pseudo Test Collections for Learning to Rank Scientific Articles. CLEF 2012, LNCS 7488, pp. 42–53, 2012.
[8] J. Bobadilla, A. Hernando, O. Fernando, A. Gutiérrez. Recommender systems survey. Knowledge Based Systems 46:109-132, 2013.
[9] B. Carterette, V. Pavlu, E. Kanoulas, J. Aslam, J. Allan. Evaluation over thousands of queries. Proceedings of the 31st ACM SIGIR Conference, pp. 651-658, 2008.
[10] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete. Profile-based recommendation: A case study in a parliamentary context. Journal of Information Science, 43(5), 665-682, 2017.
[11] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete. Committee-based profiles for politician finding. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 25(2), 21–36, 2017.
[12] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete. On the selection of the correct number of terms for profile construction: Theoretical and empirical analysis. Information Science, 430-431, 142-162, 2018.
[13] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete. Luis Redondo-Expósito. Automatic Construction of Multi-faceted User Profiles using Text Clustering and its Application to Expert Recommendation and Filtering Problems. Submitted to Knowledge-based Systems.
[14] Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, Byron Dom. Expertise Identification using Email Communications. CIKM, 528-531, 2003.
[15] N. Craswell, A.P. de Vries, I. Soboroff. Overview of the TREC 2005 Enterprise Track. Proceedings of the 14th TREC Conference, 2005.
[16] U. Hanani, B. Shapira, P. Shoval. Information filtering: Overview of issues, research and systems. User Modelling and User-Adapted Interaction 11:203-259, 2001.
[17] B. Huurnink, K. Hofmann, M. de Rijke. Simulating searches from transaction logs. SIGIR 2010 Workshop on the Simulation of Interaction, 2010.
[18] M. Kolla y O. Vechtomova. In Enterprise Search: Methods to identify argumentative discussions and to find topical experts. TREC, 2006.
[19] V. Mangaravite, R.L.T. Santos, I.S. Ribeiro, M.A. Gonçalves, A.H.F. Laender. The LExR Collection for Expertise Retrieval in Academia. In Proceedings of the 39th ACM SIGIR Conference, pp. 721-724, 2016.
[20] M. Sanderson, J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. Proceedings of the 28th ACM SIGIR Conference, pp. 162-169, 2005.
[21] A. Schuth, F. Sietsma, S. Whiteson, M. de Rijke. Optimizing Base Rankers Using Clicks. A Case Study Using BM25. ECIR 2014, LNCS 8416, 75–87, 2014.
[22] P. Serdyukov, H. Rode, D. Hiemstra University of Twente at the TREC 2007 Enterprise Track: Modeling relevance propagation for the expert search task. TREC, 2007.
[23] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su. ArnetMiner: extraction and mining of academic social networks. Proceedings of the 14th ACM SIGKDD Conference, pp. 990-998, 2008.
[24] Z. Yang, L. Hong, B. Davison. Topic-driven Multi-type Citation Network Analysis. RIAO, 2010.