Experiencias y lecciones aprendidas sobre búsqueda de expertos y filtrado de documentos en un contexto parlamentario

Luis M. de Campos, Juan M. Fernández-Luna
Juan F. Huete, Luis Redondo-Expósito
(lci,jmfluna,jhg,luisre)@decsai.ugr.es
Dpto. de Ciencias de la Computación e I.A.
ETSI Informática y de Telecomunicación, CITIC-UGR,
Universidad de Granada
Granada, Spain

Carmen Tur-Vigil
Antonio Tagua-Jiménez
(mc.tur,aj.tagua)@parlamento-and.es
Servicio de Publicaciones Oficiales
Parlamento de Andalucía
Sevilla, Spain

RESUMEN

En este trabajo presentamos los trabajos que desde el grupo de investigación UTAI de la Universidad de Granada se han efectuado dentro del campo del *expert finding* o *búsqueda de expertos*, haciendo hincapié en las lecciones aprendidas bajo los enfoques de recuperación de información y clasificación documental. Dentro de un contexto parlamentario hemos tratado, por un lado, de identificar a aquellos diputados que serían capaces de solventar una determinada problemática planteada por un ciudadano y, por otro lado, distribuir un nuevo informe (o documento en general) entre los diputados que pudiesen estar interesados.

KEYWORDS

Expert finding, document filtering, profiles, recommender systems, information retrieval, document classification

1. INTRODUCCIÓN

En este trabajo pretendemos ofrecer una revisión muy general de la investigación que hemos llevado a cabo en los campos de la Búsqueda de Expertos (BE) y el Filtrado de Documentos (FD), con objeto de mostrar tanto las contribuciones a los campos como las lecciones aprendidas.

Cuando hablamos de BE nos referimos a la recomendación de expertos, es decir, a partir de un conjunto de expertos (científicos, políticos, etc.) y una necesidad de información por parte de usuarios expresada mediante una consulta, el problema consiste en determinar qué expertos son los más apropiados para satisfacer dicha necesidad [11]. Por otro lado, el FD es un problema parecido en el que llega un documento nuevo a una organización y, partiendo de dicho conjunto de expertos, debemos determinar qué expertos estarían interesados en ese documento [10]. Aún siendo problemas diferentes, tienen muchas características en común. En este trabajo presentaremos los enfoques de solución basados en Recuperación de Información (RI) y clasificación documental (CD) que hemos puesto en práctica para resolverlos.

En cualquiera de los dos problemas se deben capturar y almacenar los intereses de los usuarios/expertos y esto se hace mediante una estructura que se denomina perfil de usuario [9] y, que en su formato más simple, almacena un conjunto de palabras que representa sus preferencias. Una parte esencial de la investigación se ha centrado en determinar las estructuras más adecuadas para representar a los expertos.

El contexto de la investigación que hemos desarrollado desde la Universidad de Granada ha sido el parlamentario, en el que los expertos son los diputados del Parlamento de Andalucía. Y la razón es histórica, ya que en el pasado se han realizado desarrollos de RI estructurada con la colección de diarios de sesiones de la VIII legislatura marcados en XML. Esa colección es también apropiada para los problemas aquí considerados.

Para dar a conocer la investigación llevada a cabo en los campos de BE y FD y las lecciones aprendidas (a lo largo del texto se irán describiendo las conclusiones obtenidas en los distintos problemas que se aborden, indicadas mediante (Ci)), este trabajo quedará organizado de la siguiente forma: tras esta introducción se describirán las experiencias y lecciones aprendidas con el enfoque de RI, exponiendo el entorno de evaluación y las alternativas exploradas, así como su rendimiento. Brevemente también se describirá la experiencia basada en clasificación documental.

2. EXPERIENCIAS CON EL ENFOQUE RI

Estos problemas pueden resolverse mediante el uso de técnicas de RI: consideraremos que los perfiles de usuario de los diputados que se han construido a partir de sus intervenciones son los documentos que alimentan un sistema de RI (asumimos que puede haber más de un perfil por experto). Así, se construye un índice con ellos y este es consultado para obtener una ordenación de diputados a partir de una consulta efectuada al sistema (normalmente una corta en la búsqueda de expertos y más larga en el filtrado de documentos). Previamente a esta salida final, tenemos que tener en cuenta que el índice está formado por perfiles de usuario y que la ordenación original dada la consulta estará compuesta de perfiles, no de expertos. Por tanto, se tendrá que llevar a cabo un cálculo del valor de relevancia de cada experto a partir de las posiciones de sus perfiles y producir una ordenación final de expertos. Se han probado varias técnicas de fusión de ordenaciones para tal fin, siendo en [3] donde se presentaron y se evaluaron.

2.1. La colección de prueba

La colección con que se ha trabajado en esta línea de investigación es la de diarios de sesiones de la VIII Legislatura del Parlamento de Andalucía. Cada documento está marcado en XML y contiene tanto los metadatos de la sesión, como la transcripción del conjunto

http://irutai2.ugr.es/ColeccionPA/legislatura8.tgz

de iniciativas que se han discutido en pleno o en una comisión en la fecha del diario (cada iniciativa está formada por las intervenciones de uno o más diputados en el debate sobre algún tema). Concretamente son un total de 5.258 iniciativas, compuestas por 12.633 intervenciones de diputados (2,4 intervenciones por iniciativa en media), ocupando un total de 148 MB. En total está formada por 136.209 párrafos, 19.429.148 palabras y 73.443 términos distintos (después de eliminar palabras vacías y realizar *stemming*).

Las ventajas de esta colección para los problemas de búsqueda de expertos y filtrado de documentos son: en primer lugar, al incluir las intervenciones de los diputados en las iniciativas se puede asumir que los discursos de éstos ofrecen evidencias textuales de las temáticas en que están interesados y, por tanto, son susceptibles de ser empleados para la construcción de sus correspondientes perfiles de usuario. En segundo lugar, al estar marcada en XML posibilita un acceso muy eficiente a las iniciativas e intervenciones. En tercer lugar, posee un tamaño manejable, lo que permite conocerla bien y comprender el porqué de los resultados que se están obteniendo. Por el contrario, las desventajas podrían ser que en la VIII legislatura sólo había tres partidos políticos representados en el Parlamento de Andalucía y varios diputados que participan en muchos plenos y comisiones. Esto causa que haya perfiles muy genéricos que puedan introducir ruido en los resultados. Y también que es una colección formada sólo por documentos, careciendo de consultas y sus correspondientes juicios de relevancia (cómo hemos abordado este problema lo mostramos en las secciones 2.3 y 2.4).

2.2. Particionamiento: entrenamiento y prueba

La colección anterior se ha dividido en dos particiones aleatorias: una de entrenamiento del 80 % de las iniciativas y otra de prueba con el 20 % restante y se ha repetido este proceso de particionamiento cinco veces en total, aprendiendo los perfiles con cada partición de entrenamiento y calculando las medidas de evaluación con cada una de prueba, obteniendo las medias de los resultados obtenidos.

2.3. Consultas

En el caso de la BE y con objeto de simular las posibles consultas que formularía un ciudadano que estuviese interesado en localizar a un experto (en este caso, diputado) hemos empleado los extractos (títulos formados por un conjunto pequeño de términos) de las iniciativas que componen el conjunto de prueba. Por otro lado, en relación con el problema del FD, las consultas deben ser mucho más grandes, por lo que hemos empleado las iniciativas completas (haciendo las veces de documentos que llegan a la institución) [2]. Así, para cada iniciativa en test, creamos dos consultas: una corta formada por el extracto y una larga formada por la iniciativa en sí.

2.4. Juicios de relevancia

Para cada consulta, en primer lugar hemos considerado como relevantes aquellos diputados que han participado en la iniciativa correspondiente. Estamos asumiendo que los métodos de recomendación ofrecerán un mayor rendimiento en tanto en cuanto mejor sea su capacidad de identificar a los propios participantes (que consideramos interesados/expertos en la temática tratada). Esta puede ser una restricción considerable en los juicios de relevancia al asumir que no hay ningún otro diputado relevante a una consulta dada, siendo este hecho de especial importancia cuando consideramos problemas de filtrado. Por tanto una segunda alternativa para definir los juicios de relevancia es considerar solo aquellas iniciativas del conjunto de test correspondientes a alguna comisión parlamentaria (reuniones de un número menor de parlamentarios que tratan temas específicos como cultura, sanidad, educación, etc.). Así, para una iniciativa de este tipo consideramos que todos los miembros de la comisión en la que fue discutida son relevantes (hayan intervenido o no en la misma).

(C1) Nuestra hipótesis inicial a la hora de considerar los juicios de relevancia (un modelo es mejor que otro cuanto mayor sea su capacidad de identificar a los participantes en la iniciativa) se ha mostrado acertada en los problemas de filtrado. En este caso, aunque los valores de las métricas son más altos cuando extendemos los juicios de relevancia a todos los miembros de la comisión, el *ranking* entre los distintos modelos es prácticamente el mismo con unos juicios o con otros.

2.5. Medidas de evaluación

Se ha considerado llevar a cabo la evaluación en los primeros resultados en la ordenación empleando como medidas principales las clásicas exhaustividad, precisión y *Normalized Discounted Cumulative Gain* , esto es *recall@K*, *precision@K* y *NDCG@K*, respectivamente, tomando K los valores 10 y nr, siendo nr el número (variable) de diputados relevantes para la consulta.

(C2) A partir del análisis de resultados experimentales vemos que todas las medidas están altamente correladas, ofreciendo las mismas tendencias dentro de cada uno de los contextos de experimentación considerados, aunque exhiben diferencias cuando comparamos entre BE y FD. Esto nos ha llevado a concluir que son problemáticas esencialmente distintas, aunque ambas se puedan abordar de forma similar. De igual modo, hemos visto que parece más interesante evaluar el problema del FD con *recall@nr*, porque en este contexto se tiene la intención de enviar los documentos que llegan al sistema al mayor número posible de expertos potencialmente interesados (cuantos más, mejor, y de forma relativamente independiente de su posición en la ordenación) y este hecho lo refleja mejor la medida de exhaustividad. Por otro lado, pensamos que la medida *NDCG* es la más apropiada para el contexto de la BE, ya que en este caso nos interesa que estén los verdaderos expertos en posiciones altas del orden.

2.6. Modelos básicos de referencia

El primer tipo de perfil con que hemos trabajado, y el más básico, es el denominado monolítico: compilar todas las intervenciones de un diputado y crear un macrodocumento que represente de esta forma su perfil de usuario. Así, cada experto queda representado por un único perfil en el que encuentran mezcladas todas las temáticas de interés del diputado correspondiente. El segundo, considerar las iniciativas como documentos aislados (por lo que no se puede hablar de perfil propiamente dicho capaz de representar los intereses del diputado). Estos dos métodos de referencia representan los dos posibles extremos que podemos tener: todo junto o todo separado.

(C3) En la comparación inicial, concluimos que el uso de perfiles monolíticos es más interesante que el de iniciativas sueltas para el problema de FD, ocurriendo justo al contrario para BE.

2.7. Composición del perfil

Cuando hablamos de la composición de perfil nos estamos refiriendo a cuatro elementos principales (tratados en [2]):

El tipo de elemento léxico que compone el perfil: Un perfil puede estar compuesto de los términos que aparecen directamente en las intervenciones, la materias de un tesauro que han sido asignadas a las iniciativas por documentalistas, o las temáticas que se han generado de forma automática analizando el texto con métodos como LDA, por ejemplo. Estos dos últimos tipos se podrían enmarcar en una categoría más conceptual o semántica, ya que representan una idea o concepto tratada en la intervención o iniciativa.

(C4) Se han realizado pruebas para determinar qué elemento léxico es más apropiado y la conclusión es que el uso de términos derivados directamente del texto (con *stemming*) es la mejor opción. La mayor amplitud del vocabulario de términos parece que es un factor positivo en este sentido.

(C5) Por otro lado, se ha visto que es aconsejable para nuestro problema el considerar todas las categorías gramaticales, en lugar de tratar de buscar perfiles compuestos de sólo sustantivos, sustantivos y verbos, etc. ya que pierde capacidad expresiva.

Otro elemento importante es el tamaño del perfil, es decir, el número de términos que lo componen: tras calcular el peso de cada término y ordenarlos de forma decreciente, establecer un corte e incluir en el perfil sólo los que estén por encima de él. En [2] se llevó a cabo un estudio sobre el rendimiento del uso de diferentes tamaños fijos para todos los expertos (50, 250, 500, 750 y 1000).

(C6) Se descubrió que para consultas medias y largas (filtrado), mejor perfiles grandes, mientras que para las cortas (búsqueda de expertos), perfiles pequeños.

En un trabajo posterior [4], se llevó a cabo un estudio exhaustivo de un total de cuatro métodos para realizar selección de términos y crear perfiles con diferente tamaño, personalizado según la información disponible del experto. De la experimentación podemos concluir que siempre, y para los dos problemas, es mejor usar métodos de selección variable, pero no se puede determinar de forma categórica una técnica concreta.

La importancia de las palabras, materias o temáticas incluidas en el perfil también ha sido tratada en nuestras investigaciones. Se han propuesto tres métodos de ponderación: la frecuencia del término en la intervención (*tf*) y su combinación con la frecuencia documental inversa (*TfIdf*) y la diferencia (*Diff*), que establece la diferencia entre la frecuencia normalizada del término en un documento concreto y la frecuencia normalizada fuera del documento.

(C7) A tenor de los resultados obtenidos, consideramos que siempre es mejor tener en cuenta esquemas de ponderación que consideren también información de la colección, siendo las dos últimas las más apropiadas, sin haber diferencias significativas de rendimiento.

Por último nos queda hablar de la forma de generar los perfiles una vez que se tiene la lista de términos seleccionados. Se han planteado dos métodos: uno replicar un término en el perfil el número de veces que aparece en la iniciativa y otro hacerlo de forma proporcional al peso.

(C8) Ambos son totalmente válidos y sin diferencias destacables.

2.8. Representación de los perfiles

El paso siguiente es construir subperfiles que reflejen separadamente los intereses en temáticas independientes (uno para agricultura, otro para sanidad y un tercero para economía para un experto que trabaje o tenga interés en estos tres campos, por ejemplo). Una primera opción para llevar a cabo esta construcción de subperfiles temáticos es aprovechar la participación de los diputados en comisiones parlamentarias. Si un diputado participa en una o varias comisiones, asumimos que es de su interés la temática principal de la misma por lo que se puede crear un subperfil por comisión. Así, el experto quedará representado por un subperfil por comisión creando un perfil estructurado.

(C9) En el campo de BE, en [3] se demostró empíricamente que, en general, los subperfiles son una mejor opción para representar los intereses del usuario, pues al dividir las temáticas de interés se obtiene un mejor rendimiento, aún a costa de aumentar ligeramente el tamaño del índice que los almacena.

Aunque esta forma de construir subperfiles puede ser útil si nos centramos en las comisiones, no será válida cuando consideramos las intervenciones de los diputados en los plenos (que representan aproximadamente la mitad de los documentos de la colección), por lo que se está desaprovechando información que puede ser relevante. De igual forma, esta aproximación no será válida en aquellas situaciones en las que no se disponga de la información explícita sobre la temática a la que pertenece un documento. Por tanto, sería mucho más apropiado crear subperfiles de una forma automática y en donde pudieran quedar delimitados mucho mejor los intereses de los expertos. Así, mediante la técnica del agrupamiento documental, se han construido subperfiles mucho más puros, haciendo la propuesta más genérica.

El proceso de construcción, presentado en [7] es el siguiente: dado el conjunto de intervenciones de todos los diputados se aplica un algoritmo de agrupamiento el cual creará grupos con intervenciones que traten de la misma temática. Un diputado tendrá tantos subperfiles como en grupos diferentes se hayan agrupado sus iniciativas (agrupamiento global). Cada subperfil estará formado por todas las iniciativas que hayan caído en el mismo grupo. También se puede hacer el agrupamiento exclusivamente de las iniciativas de cada diputado (agrupamiento local). Un elemento muy relevante es el número de grupos que se generarán. Lo ideal es que de forma automática se pudiera determinar el número óptimo de clústeres a crear pero eso no es fácil, por lo que se han empleado métodos que dependen del número de iniciativas y de los términos que aparecen en ellas.

(C10) De los experimentos realizados, en los que se han barajado diversos métodos de agrupamiento (jerárquicos, basados en centroides, en modelado de temáticas y mapas auto-organizativos), podemos destacar que, en general, los subperfiles construidos de esta forma son más naturales y mejoran los basados en comisiones, monolíticos e iniciativas individuales, aunque la decisión del algoritmo de *clustering* y del número de grupos tiene un gran impacto sobre el rendimiento en los dos problemas considerados.

(C11) En cuanto a la elección de agrupamiento global o local, parece que depende del problema y el tamaño de los superfiles: para la búsqueda de expertos, mejor el local, ya que las consultas cortas

rinden mejor con subperfiles más pequeños, mientras que para filtrado de documentos, mejor el global, pues ocurre al contrario.

El siguiente paso consiste en detectar que una iniciativa puede simultáneamente tratar de diferentes temáticas, y por tanto se puede dividir en diferentes subdocumentos más homogéneos, y así tratar de construir subperfiles más puros. Y esta es la tarea que se ha llevado acabo en [8] aplicando la técnica *Latent Dirichlet Allocation* [1]: se dividen los documentos originales en subdocumentos que están asociados a diferentes temáticas obtenidas mediante el algoritmo LDA y se distribuyen los términos en estos subdocumentos, a partir de los cuales se construyen los subperfiles. Se pasa de un dominio de términos a temáticas y seguidamente a términos de nuevo.

(C12) Trabajar con subperfiles de temáticas, aunque se reduce considerablemente la dimensionalidad del problema, no se ha demostrado como una solución efectiva.

En [8] se han planteado varias técnicas de distribución de los documentos en subdocumentos (para evitar una atomización excesiva de los documentos). Y aunque sólo se ha aplicado al problema de la BE, podemos indicar que la aplicación de LDA y estas técnicas de distribución son en su gran mayoría adecuadas para este problema.

2.9. Gestionando la ordenación de subperfiles

Los resultados obtenidos por el sistema de RI no son necesariamente una ordenación de expertos únicos, sino que pueden encontrarse en esa ordenación varias ocurrencias de un mismo experto relacionadas con distintos subperfiles del mismo. Por esto se deben aplicar estrategias de fusión para agregar los subperfiles de un mismo experto con alguna función de agregación.

(C13) Se obtienen mejores resultados empleando, en lugar de funciones de agregación clásicas, una función que aplique un descuento logarítmico a los valores de relevancia en función de su posición en el ranking [3].

3. EXPERIENCIAS CON EL ENFOQUE DE CLASIFICACIÓN DOCUMENTAL

Para abordar estos problemas mediante el aprendizaje automático en lugar de usar RI, podríamos utilizar clasificación multietiqueta. Una de las posibilidades es construir un clasificador documental binario, con clases relevante y no relevante, para cada diputado, que represente sus intereses a partir de los textos de las intervenciones que ha realizado. Cuando un nuevo documento o consulta llegue al sistema, se pueden emplear los clasificadores aprendidos para determinar qué expertos recibirán el documento o podrán ser aptos para resolver la consulta, asignando un valor numérico que establezca el grado de pertenencia a cada clase y generar así una ordenación de diputados, recomendando aquellos que estén en las posiciones más altas.

Este enfoque fue puesto en práctica por los autores en el trabajo [6] usando la misma colección de pruebas. Uno de los principales problemas de este enfoque es que, si bien los ejemplos positivos de cada diputado están claros (son sus intervenciones), podríamos considerar negativos todos aquellos en los que no intervienen. Pero podría darse el caso de que otros documentos de temática similar a los ejemplos positivos de un diputado sean considerados negativos, lo que podría hacer que el clasificador pueda confundirse. Como solución que intente paliar el problema, hemos propuesto el uso de la técnica *Positive Unlabeled Learning (PUL)* [12], con la cual se podría identificar un conjunto de ejemplos probablemente negativos (de entre los que no están en realidad etiquetados, los documentos correspondientes a intervenciones de otros diputados) y aprender los clasificadores con los dos conjuntos de una manera más fiable. Se ha propuesto un método PUL basado en una modificación de la técnica de agrupamiento *K-Means*.

(C14) Como conclusión podemos decir que siempre que se empleen estas técnicas PUL previamente y métodos que permitan el balanceo de los datos, la clasificación podrá verse como una alternativa viable al enfoque de RI.

4. TRABAJOS FUTUROS

Pensamos que para seguir trabajando en estos problemas y avanzar, debemos considerar nuevas colecciones que permitan afrontar la BE y el FD en combinación con nuevas técnicas. Para tal fin, considerando colecciones de artículos científicos, se podrían incorporar métodos de análisis de redes de investigación (coautorías o citas) o la inclusión de índices de calidad a la hora de recomendar expertos. Es por esto que se ha generado la colección PMSC-UGR a partir de artículos científicos de PubMed completados con datos de Scopus [5]. También nos dará pie a incluir la dimensión temporal en estos procesos y determinar si los perfiles temporales ayudan a una mejor recomendación.

ACKNOWLEDGMENTS

Este trabajo ha sido cofinanciado por el Ministerio de Economía y Competitividad español mediante el proyecto TIN2016-77902-C3-2-P y el programa FEDER.

REFERENCIAS

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. of Machine Learning Research 3:993-1022, 2003.
[2] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, Profile-based recommendation: A case study in a parliamentary context, J. of Inf. Sci. 43(5):665-682, 2017.
[3] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, Committee-based profiles for politician finding, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 25(Suppl. 2):21-36, 2017.
[4] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, On the selection of the correct number of terms for profile construction: Theoretical and empirical analysis, Information Sciences 430–431:142–162, 2018.
[5] C. Albusac, L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, PMSC-UGR: A test collection for expert recommendation based on PubMed and Scopus, LNCS 11160:34-43, 2018.
[6] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, L. Redondo-Expósito, Positive unlabeled learning fro building recommender systems in a parliamentary setting, Information Sciences, 433-434:221-232, 2018.
[7] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, L. Redondo-Expósito, Automatic construction of multi-faceted user profiles using text clustering and its application to expert recommendation and filtering problems, Knowledge-Based Systems 190, article number 105337, 2020.
[8] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, L. Redondo-Expósito. LDA-based term profiles for expert finding in a political setting. Submitted 2020.
[9] S. Gauch, M. Speretta, A. Chandramouli and A. Micarelli, User profiles for personalized information access, LNCS 4321:54–89, 2007.
[10] U. Hanani, B. Shapira, P. Shoval, Information filtering: Overview of issues, research and systems, User Model. User-Adapt. Interact., 11(3):203-259, 2001
[11] S. Lin, W. Hong, D. Wang, T. Li, A survey on expert finding techniques, Journal of Intelligent Information Systems 49:255-279, 2017.
[12] B. Zhang , W. Zuo , Learning from positive and unlabeled examples: a survey, in: Proc. 2008 Int. Symposiums on Inf. Process. 650–654, 2008.