TESIS DOCTORAL

SISTEMAS DE RECOMENDACIÓN EN CONTEXTO PARLAMENTARIO BASADOS EN TÉCNICAS DE APRENDIZAJE AUTOMÁTICO Y RECUPERACIÓN DE INFORMACIÓN

Autor:
Luis REDONDO EXPÓSITO

Directores:
Luis Miguel DE CAMPOS IBÁÑEZ
Juan Francisco HUETE GUADIX

Dpto. Ciencias de la Computación e Inteligencia Artificial
Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación
UNIVERSIDAD DE GRANADA

Programa de Doctorado en Tecnologías de la Información y la Comunicación

MARZO 2020

Editor: Universidad de Granada. Tesis Doctorales
Autor: Redondo Expósito, Luis
ISBN: 978-84-1306-626-4
URI: http://hdl.handle.net/10481/63893

Agradecimientos

Durante toda mi vida he estado acompañado por personas que me han querido y conducido a ser quien soy hoy en día. En este momento tan importante, tanto en mi carrera profesional como personal, como es la realización de una Tesis Doctoral, el valor de tener cerca a las personas que te quieren es aún más grande. Por eso me gustaría agradecerles en primer lugar a mis padres, Juani Expósito y Antonio Redondo, por haberme hecho quien soy y estar orgullosos de mí. A mi hermano Jose Antonio Redondo y su pareja Irene Salido, a mi tía Mari Expósito y a sus hijos, mis primos, con los que he crecido, Juana Mari Oliver y Simón Oliver. A mis abuelos y abuelas, Luis Redondo, Pepa Ruíz, José Expósito y Juana Muro, cuya pérdida fué un punto de inflexión en mi vida. A mis tíos, Jose Expósito y Simón Expósito, sus parejas y mis primos.

A mis amigos, que han demostrado estar cerca de mí cuando los he necesitado. A Francisco Román, el amigo más antiguo que tengo. A Ramón Sánchez, porque nuestra amistad siempre sea igual de estrecha, a su pareja Lidia Doncel y a Borja Ortiz por regalarme muchas tardes agradables. A mis compañeros de la carrera Rubén Sánchez, Germán Martínez, Jose Antonio González y Pablo Sánchez. A todos mis compañeros del Máster de Ciencia de Datos, por darme uno de los mejores años de mi vida. A Fernando Palacios y David Martínez porque en mis mejores anécdotas siempre aparecen ellos. A Juanjo Escobar, por haber estado siempre tan cerca. A mis compañeros y amigos del CITIC Luis, Eugenio, Juanma, Dani, Antares, Álvaro, Ofelia, Manolo, Salva y Carlos, profesionales excelentes y mejores personas. A mis directores, Luis M. de Campos y Juan Huete y a mi compañero Juanma Fernández. En definitiva, a todas las personas que no han dejado que haga este trayecto sólo.

Por último, aunque no menos importante, a mi pareja Melis Ozkan, que aunque ambos pertenecemos a lugares distintos del mundo y ni hablemos el mismo idioma, el destino quiso que nos encontráramos.

Dedicado:
A mi familia.

ÍNDICE

Página

Lista de Tablas

Tabla **Página**

Lista de Figuras

Figura **Página**

CAPÍTULO 1

INTRODUCCIÓN

1.1 Introducción

El término "Sociedad de la Información" fue acuñado con la intención de poder entender cómo la tecnología permite al individuo generar, acceder, manipular y distribuir cantidades ingentes de información [95]. Esta realidad ha ido desde entonces creciendo de forma exponencial hasta el punto en que se han desarrollado múltiples áreas interdisciplinares de estudio para poder lidiar con el problema de gestionar el increíble volumen de información con el que contamos hoy día y cómo un individuo accede a ella. Es aquí donde nos topamos con la idea de relevancia y el valor que le da a la información un individuo ante esta sobrecarga. La sobrecarga de información y cómo gestionarla de cara al usuario de un sistema donde se aloje dicha información es un hecho fácilmente visible en cualquier herramienta de búsqueda en internet de forma cotidiana, la cual puede devolver del orden de millares de resultados, pero apenas la primera decena de ellos es altamente relevante para el usuario.

En la actualidad, la Sociedad de la Información sigue evolucionando a un ritmo frenético día a día. Pero no fue hasta hace pocos años atrás cuando se concibió la idea de gestionar de manera eficiente la enorme cantidad de información de la que se dispone. En el año 2011, Eric Schmidt, CEO de Google, afirmó que la humanidad había producido hasta 2003 una cantidad de información equivalente a 5 exabytes, pues bien, en la actualidad ese mismo volumen de información se genera cada dos días. Los más de 255 millones de usuarios de Redes Sociales como Twitter, generan diariamente un promedio de 500 millones de tweets, en Facebook o Instagram se publican al día más de 83 millones de imágenes. Google en el año 1998 indexaba alrededor de unos 26 millones de páginas web, dos años después, en el año 2000, Google tenía indexados más de 1000 millones de URLs y en 2008 esa cantidad alcanzó la cifra de 1 billón de páginas web indexadas en sus servidores. Se podrían añadir muchos más ejemplos como Amazon, Netflix o Ebay. Pese a que pueda parecer en primera instancia que todas estas empresas se dediquen a

sectores económicos distintos, la verdad es que el éxito de todas ellas radica en la gestión eficiente de información, independientemente de cómo esté representada, y facilitar posteriormente la accesibilidad a la misma por parte de los usuarios.

Esta tendencia no parece que tenga la menor intención de cambiar en un futuro. En los próximos años, no solo la cantidad de contenido que se genera en la *World Wide Web* va a seguir creciendo de forma exponencial sino que, de la misma forma, lo hará la información que se extraiga a su vez de esta. Esto representará nuevos retos en el tratamiento de la información y su posterior accesibilidad llegando a cambiar la concepción que se tiene hoy en día en la gestión de datos. Ciertos autores, tanto científicos como periodistas entre otros, aseguran que estos acontecimientos suponen una nueva revolución industrial que modificará la sociedad tal y como es actualmente; nuevos empleos, un nuevo sistema económico, nuevas formas de entender la política o una nueva concepción de la sociedad entre otros muchos, son los retos a los que la Sociedad de la Información está abocada a enfrentarse en un futuro no muy lejano.

Es ante esta realidad donde los Sistemas de Recomendación entran en escena con el objetivo de amparar al usuario ayudándole a seleccionar la información que le es relevante en función de sus propios intereses. Un Sistema de Recomendación está compuesto por técnicas y herramientas implementadas con el objetivo de sugerir al usuario sobre algún tipo de decisión que se pueda tomar a partir de la experiencia y basándose en un conjunto de datos en los que argumentarla [7]. La aplicación práctica de los Sistemas de Recomendación puede ser encontrada en ámbitos completamente dispares gracias a que el tratamiento de la información es un ciencia completamente horizontal y transversal. He aquí algunos de los antecedentes más destacados de aplicación de estos sistemas.

- **Comercio electrónico**: En [133] hizo aparición el concepto de Sistema de Recomendación aplicado al comercio electrónico con el objetivo de ayudar al usuario a encontrar con mayor facilidad los productos que se ajustasen mejor a sus necesidades. A día de hoy, plataformas de comercio electrónico tan fuertes en el sector como lo puede ser Amazon han basado su fortaleza en la aplicación de estas técnicas. Más adelante en el tiempo, en [105] se planteó con éxito la aplicación de técnicas de Aprendizaje Automático para mejorar la calidad de las recomendaciones en plataformas como TripAdvisor.

- **Ocio web**: En [52], se presentan los algoritmos y técnicas que hacen a Netflix uno de los Sistemas de Recomendación híbridos más potentes en la actualidad. En estos sistemas se utilizan técnicas propias de Recuperación de Información como es la obtención de un ranking personalizado basado en el perfil de usuario y algoritmos de Aprendizaje Automático con el objetivo de discernir la relevancia o no de un elemento a un usuario en función de los intereses de otros usuarios similares.

- **Redes sociales**: Las grandes potencias en la construcción de perfiles de usuario y la correspondiente gestión mediante Sistemas de Recomendación como son Twitter y Facebook

basan en gran medida su funcionamiento en las técnicas de etiquetado de elementos, las cuales son llevadas a cabo por los propios usuarios del sistema de forma colaborativa. De esta forma se crea una red propiamente dicha de usuarios en torno a las mismas etiquetas, la cual tiene el propósito de facilitar y reforzar la calidad de predicción de los algoritmos usados en la recomendación de nuevo contenido. Esta aproximación se explica con mayor detalle en [142].

- **E-gobierno**: En [34] se presentan a los Sistemas de Recomendación como un enfoque más que viable en el acercamiento de los ciudadanos a sus gobernantes, además de considerarlos como el núcleo de la actividad política en las emergentes *Smart Cities*. En otro orden, la labor que se desarrolla en este momento en la propia Universidad de Granada sobre la temática de Sistemas de Recomendación en el ámbito parlamentario continúa progresando con interesantes avances, los cuales dan paso a la realización de esta Tesis Doctoral [39–42].

Cómo particularidad específica, la cual se pretende estudiar en esta Tesis Doctoral, los representantes políticos de la sociedad en su actividad parlamentaria también se ven afectados por esta casuística; una cantidad abrumadora de información (noticias de prensa, sesiones parlamentarias, peticiones ciudadanas...) y una necesidad apremiante de estar completamente al día en los temas actuales del territorio donde ejercen su labor política y, sin embargo, no toda la información con denotaciones políticas les puede interesar directamente. Por ejemplo, un Miembro del Parlamento cuya actividad se vea enfocada al área de agricultura probablemente no esté directamente interesado en lo acontecido en una sesión parlamentaria donde se trataron temas referentes al área de sanidad.

La enorme cantidad de información cotidiana que se genera a cada instante globalmente y más concretamente en el ámbito político desemboca en la no trivialidad de gestionarla de forma eficiente de modo que un usuario pueda estar correctamente informado. La compleja labor de este proceso no solo radica en que el usuario reciba la mayor cantidad de información posible, sino que esta sea solo y exclusivamente la más relevante para el mismo. Aunque la eficiencia de los Sistemas de Recomendación avanza en virtud de este propósito a pasos agigantados día a día con notables progresos en lo que respecta a su funcionalidad y estructura, la creación de forma exponencial de nueva información y la necesidad apremiante de acceso a esta por parte de los usuarios sigue ganando la carrera.

En este punto, surge el planteamiento de cómo se puede construir un Sistema de Recomendación que sea capaz de trabajar a partir de la información generada en un parlamento donde la principal fuente de información proviene de los discursos parlamentarios. A partir de esta cuestión entra en escena el concepto del perfilado de usuarios y la forma en la que un individuo debe estar representado de cara al sistema. Un perfil de usuario es la representación de ciertas características que definen a los usuarios que componen un Sistema de Recomendación y, encontrar una correcta representación de la información y características de un usuario, es

determinante a la hora de obtener un mejor rendimiento en las tareas de recomendación y filtrado. Un perfil de usuario se puede construir de diversas formas en función de la información de la que se disponga, de cómo esté esta representada y, en definitiva, el propósito para el cual el Sistema de Recomendación vaya a ser utilizado.

En el caso de estudio específico de esta Tesis Doctoral, el ámbito político y parlamentario, el problema cobra una añadida sensibilidad. Un parlamento fue concebido como un extracto fiel de la sociedad a la que representa y es por esto que debe estar al corriente de cualquier acontecimiento acaecido en el mismo momento en el que ocurra para poder desempeñar su propósito de la forma más eficiente posible. Sin embargo, en el simple periodo de un día, en la vida política y social, se pueden producir una infinidad de eventos que oscilan entre los más insignificantes a los más trascendentales. Es en este escenario donde los Sistemas de Recomendación entran en juego con el objetivo de gestionar de forma coherente la información que se produce en un territorio hablando desde el punto de vista político, permitiendo que cualquier detalle proveniente de la sociedad, por minúsculo que sea pueda llegar al receptor adecuado sin que este a su vez se vea bombardeado sistemáticamente con asuntos que, o bien no son de su competencia, o bien están fuera de su jurisdicción. La finalidad de esta idea es la de acercar y conectar a la política con la sociedad a la que representa estableciendo puentes para que todos trabajemos en la misma dirección.

Dicho esto, la hipótesis principal que se plantea en esta Tesis Doctoral es que dada la estructura de un parlamento y todo lo que gira en torno a este, los Sistemas de Recomendación pueden convertirse en un elemento fundamental para mejorar el desempeño de la actividad política independientemente del marco en que esté situada. Para ello, se plantea el estudio de la viabilidad de nuevas técnicas específicas que contribuyan a incrementar la correcta actuación de estos sistemas en el contexto político. Además de esto, se proponen nuevas formas de representación de la información de la que se dispone sobre los Miembros del Parlamento con el objetivo de determinar cuáles definen los intereses de estos de forma más precisa.

1.2 Objetivos

El problema que se plantea en este estudio es el de mejorar las prestaciones de los Sistemas de Recomendación y Filtrado en el ámbito parlamentario. Dicho esto, desde el punto de vista de la Recomendación en el ámbito parlamentario se plantea el problema como Búsqueda de Expertos [60], es decir, dada una consulta de entrada en el sistema, ya sea en forma de petición ciudadana, un concepto político en particular, etc. el sistema devuelve al usuario un conjunto con los diputados que puedan ser más afines a dicha consulta. Por otro lado, en lo que respecta al tratamiento del sistema como elemento de Filtrado, lo que se pretende es que, dado un documento como entrada ya sea una nueva iniciativa parlamentaria, una noticia de prensa, etc. el sistema debe filtrar de entre todo el conjunto de diputados aquellos para los cuales dicho documento

pueda tener alguna relevancia en función de sus intereses políticos.

Por lo tanto, el objetivo primordial de esta Tesis Doctoral es el de explorar y proponer nuevas posibles estructuras e implementaciones de Sistemas de Recomendación y Filtrado en un contexto parlamentario basándose en el uso de técnicas de Aprendizaje Automático y de Recuperación de Información. Para esta finalidad se pretenden abordar una serie de tareas específicas:

- **O1**: Revisión exhaustiva del estado del arte de la materia en lo referente a los Sistemas de Recomendación y estudio y comprensión de los trabajos previos que dan pie a la realización de esta Tesis Doctoral.
- **O2**: Estudio y comparativa de distintas aproximaciones basadas en Aprendizaje Automático y Recuperación de Información para la recomendación y filtrado de documentos en un contexto parlamentario. En este punto, el perfil de los diputados está representado inicialmente de dos formas; un documento por intervención del diputado y un documento único con todas las intervenciones unidas [50].
- **O3**: Investigación sobre una representación alternativa del perfil de los Miembros del Parlamento basada en unir intervenciones con temáticas similares haciendo uso de diferentes técnicas de clustering y valorando distintas aproximaciones para establecer el número de clusters [124].
- **O4**: Abordaje del problema desde el punto de vista del Aprendizaje Automático. En este punto, la representación de los perfiles va a estar definida por modelos Support Vector Machine (SVM) [35] donde las intervenciones propias de los diputados van a constituir los ejemplos positivos en el entrenamiento y las intervenciones del resto de los diputados los ejemplos negativos. Con el objetivo de mejorar esta representación se plantea el diseño de una nueva técnica de Positive Unlabeled Learning [22] basada en una modificación del algoritmo K-means cuya función será eliminar del conjunto de ejemplos negativos de entrenamiento aquellas intervenciones que pese a no ser del Miembro del Parlamento en cuestión, tratan la misma temática. Por último, en esta fase, se propone el calibrado de los modelos SVM ajustando de forma individual el umbral que define cuando un documento es relevante para cada uno de los diputados.
- **O5**: Aplicación del algoritmo Latent Dirichlet Allocation [14] con el objetivo de encontrar una nueva representación del perfil de los diputados basada en términos, la cual se adapte mejor a nuestro caso de estudio en el ámbito parlamentario.

1.3 Revisión histórica

En el artículo *As We May Think* [21], el ingeniero estadounidense Vannevar Bush introdujo públicamente por primera vez la idea de la extracción de fragmentos de información relevante

de manera automática. A partir del asentamiento de esta idea como una ciencia, comienza a considerarse el área de estudio de la Recuperación de Información, dando lugar unos años más tarde a los primeros Sistemas de Recuperación de Información automatizados a finales de los años 60. Después de eso, durante años, los progresos en la Recuperación de Información quedaron prácticamente paralizados, y no fue hasta el año 1992 que este campo de estudio se volvió a reactivar cuando, el Departamento de Defensa de E.E.U.U en colaboración con el Instituto Nacional de Estándares y Tecnología, financiaron la Conferencia de Recuperación de Texto (TREC). Este hecho tuvo como consecuencia que el conjunto de ingenieros y científicos que hasta la fecha se había dedicado a la investigación, sin futuro por aquél entonces de la Recuperación de Información, pudieran reunir el apoyo y financiación suficientes como para poder desarrollar una infraestructura con el objetivo de diseñar y evaluar distintas metodologías innovadoras para la extracción de información textual en colecciones documentales a gran escala.

A partir de ahí, comenzaron a diseñarse los primeros motores de búsqueda, los cuales eran más complejos y eficientes que sus predecesores. Algunos de los motores de búsqueda más famosos y más utilizados operan sobre una de las fuentes de información más grandes que jamás haya existido; Internet. Google, Bing, Yahoo o Lycos son algunos de los nombres de los Sistemas de Recuperación de Información más potentes en la actualidad. De entre los investigadores más importantes que centran sus estudios en torno a la Recuperación de Información, cabe destacar autores como W. Bruce Croft [36]. Keith Van Rijsbergen [120] y Ricardo Baeza-Yates [7].

Dentro de las distintas implementaciones de los Sistemas de Información, existen los Sistemas de Recomendación y Filtrado, los cuales trabajan con entidades de información con distintas características (películas, música, libros, etc.), pero considerándolas como una representación textual de las mismas. Para evaluar los resultados de las distintas aproximaciones que se proponen en esta Tesis Doctoral se ha diseñado e implementado un Sistema de Recomendación y Filtrado específico que se construye a partir de la información contenida en un perfil de usuario [1], en este caso, las intervenciones en los debates parlamentarios de los Miembros del Parlamento. De este modo, un usuario podría lanzar un consulta con el objetivo de obtener un ranking con los Miembros del Parlamento que mejor se ajusten a dicha consulta.

Este Sistema de Recomendación y Filtrado se desarrolla a partir de los progresos de investigaciones previas del equipo de investigación *Uncertainty Treatment in Artificial Intelligence* (UTAI). En 2015, se publicó el artículo *A lazy approach for filtering parliamentary documents* [39], donde se planteaba la primera versión del Sistema de Recomendación y Filtrado en el ámbito parlamentario sobre en el que posteriormente se iban a evaluar las distintas metodologías que en esta tesis se plantean. En esta primera publicación, se trataba el problema de recomendar Miembros del Parlamento, o filtrar documentos para los Miembros del Parlamento, en base a la construcción de unos perfiles basados en la conjunción de todas sus intervenciones. Es, por tanto, la primera vez que en el estado del arte se plantea el concepto de perfil en un contexto político, donde la actividad parlamentaria y temáticas en las que un Miembro del Parlamento puede

estar interesado están bien definidas en las intervenciones del propio Miembro del Parlamento en los debates parlamentarios como pueden ser Educación, Sanidad, Empleo, Igualdad, Obras Públicas, Vivienda, etc. A raíz de este primer artículo, con posterioridad, se evaluó la posibilidad de definir dos formas distintas para la representación de los perfiles de los diferentes Miembros del Parlamento [41].

1.4 Organización de la Tesis Doctoral

Los siguientes capítulos de esta Tesis Doctoral estan organizados en función de los avances que se han producido el desarrollo de la línea de investigación. En primer lugar, en el Capítulo 3 se realiza una comparativa entre, construir un Sistema de Recomendación y Filtrado con un método basado en Aprendizaje Automático y posteriormente clasificando las consultas, y con las técnicas clásicas del campo de la Recuperación de Información. A raíz de este primer estudio se plantea, por un lado, en el Capítulo 4 cómo la aplicación de técnicas de Positive Unlabeled Learning pueden ayudar a mejorar un Sistema de Recomendación y Filtrado basado en Aprendizaje Automático, y por otro lado, en el Capítulo 5 se introduce el uso de perfiles multifacéticos basados en técnicas de clustering como alternativa a la representación de los perfiles de los Miembros del Parlamento donde se usaba un enfoque basado en Recuperación de Información. En el Capítulo 6 se presenta otra forma alternativa de representación de perfiles de Miembros del Parlamento, usando un enfoque de Recuperación de Información basado en términos asociados a temáticas específicas, donde se emplea la técnica de Latent Dirichlet Allocation (LDA). En cada uno de los capítulos mencionados se puede encontrar una revisión del estado del arte más exhaustiva y una serie de trabajos relacionados de forma más detallada con respecto a lo que en el capítulo en cuestión se trate. Finalmente, aunque de la misma forma, cada capítulo dispone de una sección de conclusiones propia, en el Capítulo 7 se pueden ver una serie de conclusiones generales de todo el trabajo de investigación.

CAPÍTULO 2

PRELIMINARES

2.1 Preliminares

El problema que se pretende abordar en esta Tesis Doctoral consiste en que dado un Sistema de Recomendación y Filtrado en un contexto parlamentario, el cual se construye a partir de las intervenciones en los debates parlamentarios de los Miembros del Parlamento, se propone el estudio de la aplicación de diversas técnicas de Aprendizaje Automático y de Recuperación de Información con el objetivo de mejorar el sistema. Dicho esto, en primer lugar se va a realizar una contextualización de todos los aspectos que giran en torno a este problema.

2.2 Miembros del Parlamento, iniciativas e intervenciones

Los Miembros del Parlamento, a partir de ahora MPs, son el eje central de esta investigación. Esta denominación la reciben todos aquellos individuos que alguna vez hayan participado de forma activa en los debates parlamentarios, es decir, no solo los diputados electos de los distintos partidos sino también ciertas personalidades o expertos que hayan podido acudir al parlamento con el propósito de participar en los debates. En total, en los procesos experimentales de esta investigación se trabajan con un total de 132 MPs, aunque cabe destacar que en realidad son algunos más, para ser prácticos se han descartado aquellos MPs que han intervenido un número muy reducido de veces, más específicamente, aquellos que hayan intervenido en los debates parlamentarios menos de 10 veces. La mayoría de estos individuos solo han acudido una única vez al parlamento para intervenir sobre un tema muy concreto y, por lo tanto, no se podría construir un perfil fiable ni descriptivo de estos, lo que finalmente contribuiría a añadir ruido al Sistema de Recomendación y Filtrado.

Las iniciativas parlamentarias son los elementos donde se recoge lo que se ha dicho en los debates parlamentarios. Las iniciativas son documentos con una estructura en formato XML

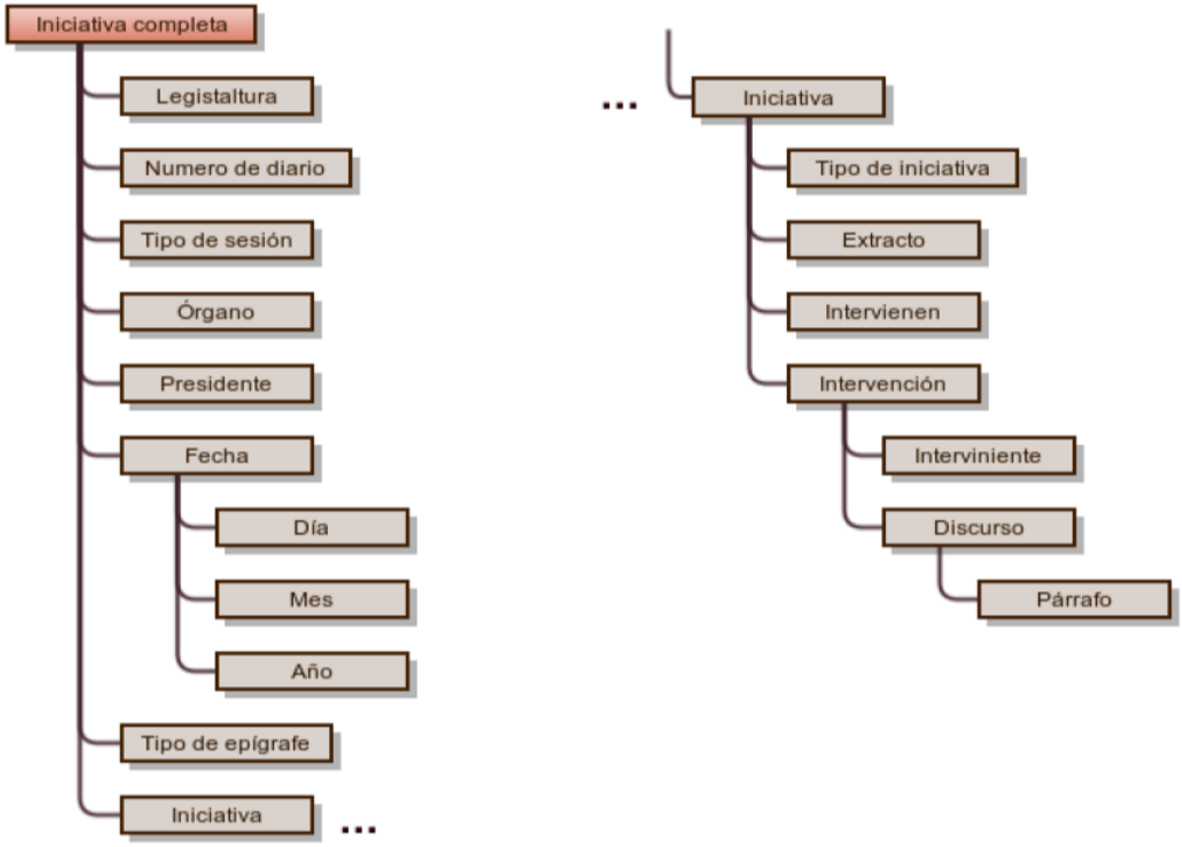

Figura 2.1: Estructura XML de una iniciativa del Parlamento de Andalucía.

donde se transcriben literalmente las intervenciones de los distintos MP. El Parlamento de Andalucía tiene un estructura específica para construir las transcripciones de las iniciativas parlamentarias, pero una de las virtudes de este Sistema de Recomendación y Filtrado es que es invariable frente a distintos tipos de representación, por lo cual se puede extrapolar y escalar a otros ámbitos políticos. Puesto que la experimentación se va a llevar a cabo con la colección de iniciativas del Parlamento de Andalucía, en la Figura 2.1 se muestra la estructura de estas.

Cabe destacar que las iniciativas se puede dividir a su vez en dos grupos disjuntos, por un lado las iniciativas parlamentarias asociadas a plenos y, por otro lado las iniciativas parlamentarias asociadas a comisiones. En un pleno se debate sobre temas diversos contenidos en la agenda del orden del día y, en las iniciativas de las comisiones, se debaten temas específicos de la comisión en sí (Sanidad, Educación, Empleo, Comercio, etc.). Independiente del grupo al que pertenezca cada iniciativa, en la estructura común caben destacar ciertas etiquetas. El *Extracto* es un nodo que contiene un breve resumen sobre lo que se va a debatir en la iniciativa parlamentaria, la información contenida en este nodo se utiliza en los experimentos como consultas para evaluar el Sistema de Recomendación y Filtrado desde el punto de vista exclusivo de la recomendación de expertos. En el nodo *Intervienen* figura un listado con todos los MPS que van a participar

en la iniciativa, esta información se tiene en cuenta para determinar los juicios de relevancia de los experimentos. El nodo $Intervencion$ por sí solo no tiene ninguna información, pero todo lo que cuelga de él se utiliza en los experimentos como consultas para evaluar el Sistema de Recomendación y Filtrado desde el punto de vista exclusivo del filtrado de documentos. Por último, los nodos $discurso$ contienen las transcripciones literales de las intervenciones de los MPs en la iniciativa.

Las intervenciones, tal y como la palabra lo define, se corresponden con lo que un MP haya aportado verbalmente en los debates parlamentarios pero, en este caso de estudio, cada una de estas intervenciones pasa a ser la unidad mínima de información de las que se puede disponer de los MPs. Cada vez que un MP interviene, cabe esperar que en dicha intervención refleje todo aquello en lo que tiene algún interés político o, en definitiva, información sobre su labor parlamentaria. Por tanto, a partir de estas intervenciones se puede extraer la información suficiente para poder perfilar a un MP con el objetivo de encontrar un tipo de representación que pueda ser aplicada sobre un Sistema de Recomendación y Filtrado.

2.3 Unidades de indexación

Como norma general, los Sistemas de Búsqueda no trabajan directamente con los documentos como tales, ni tampoco con las consultas tal y como el usuario las introduce en el sistema. Para ello, se utilizan ciertas técnicas y estrategias, en función del propósito del sistema, para representar los aspectos semánticos tanto de los documentos a indexar como de las consultas, de cara a que estos elementos sean interpretados correctamente por parte del sistema. A este proceso de extraer y representar de forma apropiada la información semántica de un documento o consulta se le denomina indexado.

El objetivo principal de la indexación es representar el contenido semántico de documentos o consultas mediante una serie de características que puedan ser indexadas. Normalmente, estas características varían en función del formato en el que la información de origen esté representada, ya sean términos en el caso de documentos textuales, notas musicales en una canción o partitura, códigos de color en una imagen o gráfico, etc. A estos elementos se les llama unidades de indexación.

Centrándonos en los documentos textuales para ejemplificar la situación, las unidades de indexación se pueden representar de diversas formas en función de lo que se quiere indexar y las características en sí del propio documento. De forma general, y en esta tesis se ha considerado como tal, es muy común determinar que las unidades de indexación en un documento de texto sean las propias palabras o términos, es decir, toda cadena de caracteres entre espacios en blanco es una unidad de indexación de manera individual. Sin embargo, existen otras estrategias para definir las unidades de indexación, como ejemplo una de las más recurrentes, los *n-gramas* [97], que son básicamente subcadenas de n caracteres, o bien cadenas de n términos para poder

encontrar así palabras compuestas por varios términos con significado completo ("Parlamento de Andalucía", "Comisión de Sanidad", etc).

Una vez seleccionada la estrategia de indexación para determinar qué información se desea propiamente indexar y en que forma va a estar representada de cara al sistema, en el caso de los documentos de texto, que es el que acontece en esta Tesis Doctoral, se lleva a cabo el proceso de indexación como tal, el cual se desarrolla en una serie de etapas.

2.4 Procesamiento de textos

La mayoría de los Sistemas de Recuperación de Información actuales se basan en el indexado automático de documentos y consultas. Para llevar a cabo este indexado, de forma general, en el proceso de construcción de un índice de documentos textuales se proceden una serie de pasos que se explicarán con más detalle en las siguientes subsecciones.

2.4.1 Análisis de la estructura y tokenización

En este primer paso se debe observar las estructura del documento a indexar para determinar si es un documento de texto plano, un documento con estructura XML, etc. Una vez detectada la estructura del documento en sí, se debe determinar qué información se desea indexar, es decir, el documento completo, ciertas etiquetas si se encuentra en formato XML, etc, en definitiva, determinar qué información es la relevante de cara a la construcción del Sistema de Recuperación de Información y cual vaya a ser su propósito. A continuación, se *parsea* la información relevante de modo que se puedan extraer las unidades de indexación, a las cuáles en este punto se les llamará *tokens*. Aunque este parezca un proceso trivial, las decisiones sobre cómo *tokenizar* un texto serán de vital importancia para todo el proceso de indexación y el posterior funcionamiento del Sistema de Recuperación de Información. Qué hacer con los caracteres especiales, los símbolos numéricos, los símbolos de puntuación o la capitalización de ciertos términos son aspectos ciertamente importantes a tener en cuenta en este punto. En definitiva, este paso conlleva qué características, en este caso términos, van a definir nuestro conjunto de datos.

2.4.2 Eliminación de stopwords

Una vez se tienen los términos en forma de tokens se procede a eliminar del conjunto aquellas palabras que no aportan gran contenido semántico. Concretamente en castellano, estas palabras son los determinantes (el, la, los, las), preposiciones (a, en, por), conjunciones (y, o), pronombres (yo, tu, él, ella) y ciertas formas verbales auxiliares que no aportan información para este propósito (haber, ser, estar). A estas palabras carentes de significado semántico se las denomina *stopwords* y el principal propósito por el que se eliminan, entre otros, es por el gran número de apariciones que tienen en cualquier texto, lo cual las dota de una gran importancia, pero sin embargo no aportan información relevante y, por otro lado, evitar así que las consultas se emparejen con los

documentos indexados a partir de esos términos, lo cual conllevaría a recomendaciones un tanto inconsistentes. Las *stopwords*, en definitiva, sólo aportan una cantidad de ruido considerable al conjunto de datos que sólo contribuye negativamente a las tareas de Recuperación de Información puesto que no sirven para determinar si un documento es relevante o irrelevante para una consulta. Por otro lado, la eliminación de *stopwords* contribuye a una notable reducción del espacio de características que van a ser indexadas de en torno al 30% y 50% del total de términos.

Por otro lado, se pueden observar en función de las características de la colección que se esté tratando, que existen ciertos términos que, aunque tengan un claro contenido semántico, en el contexto en el que se presentan pueden ser irrelevantes. Para ilustrar estos casos se va a proceder con un ejemplo: es obvio que los términos "parlamento", "institución", "comisión", "iniciativa" o "señoría" hacen referencia a una realidad y por lo tanto tienen un significado semántico bastante claro, sin embargo, en un contexto político, como el que se trata en esta Tesis Doctoral, esos términos aportan muy poca información puesto que son sistemáticamente repetidos a modo de jerga más que por su significado en sí. Dicho esto, una buena estrategia es determinar qué términos pese a tener significado no aportan más información que ruido al conjunto de datos y en consecuencia eliminarlos. Generalmente, una buena forma de detectar este tipo de términos es observar si tienen una frecuencia muy superior al promedio de la frecuencia del resto de términos.

2.4.3 Normalización morfológica

Este paso del proceso de indexación es uno, si no el más importante de todo el procedimiento. En este punto, se dispone de un conjunto de términos (*tokens*) más o menos consistente y representativo de la colección de documentos que se pretende indexar. Sin embargo, en este punto se debe aplicar cierta normalización morfológica con el objetivo de unificar, en cierto modo, aquellas variaciones de una misma palabra que hacen referencia a una misma realidad en una misma raíz. A este proceso se le llama *stemming*, y consiste en identificar la raíz morfológica de una palabra y convertir todas las variaciones de esa palabra a dicha raíz. De esta forma se evita que palabras distintas pero que hacen referencia a la misma realidad sean consideradas como palabras distintas. Por ejemplo, los términos "económico", "economía" y "economista" son palabras distintas pero que hacen referencia a la misma realidad, por lo tanto, tras el proceso de *stemming* estas palabras pasarían a estar representadas por la raíz morfológica "econom" no habiendo así distinciones entre ellas y conservando así su significado. La aplicación de procedimientos de *stemming* es sin duda una de las técnicas más utilizadas para la normalización morfológica en el ámbito de la Recuperación de Información y utilizar esta herramienta para reunir palabras con una misma raíz morfológica contribuye de manera positiva a un mejor rendimiento en las tareas de Recuperación de Información

2.4.4 Representación del peso de los términos

Tal y cómo se ha dicho anteriormente, un Sistema de Recuperación de Información fragmenta un texto en un conjunto de términos, a continuación elimina las palabras más frecuentes y las que carecen de contenido semántico y determina las raíces morfológicas de cada término para obtener como resultado las unidades de indexación. En este punto, se podría establecer un esquema de indexado binario donde un documento estaría representado por un conjunto de términos igualmente relevantes y las consultas serían emparejadas con los documentos con los que tuvieran términos en común. Pero ciertamente, en un documento, no todos los términos aparecen el mismo número de veces, lo cual los hace más relevantes, ni tampoco aparecen el mismo número de veces en la colección completa. Para representar esto, a cada uno de los términos se le debe aplicar un peso que determine su relevancia, no solo en un documento de texto en sí, sino también en toda la colección completa, de esta forma se crea una distinción entre términos y mejora la flexibilidad de la indexación. Por otro lado, otro aspecto a tener en cuenta para determinar la importancia de un término es la longitud del documento en el que se encuentra [131], no representa la misma importancia que un término aparezca tres veces en un documento y suponga un 1% del total de términos que, en otro documento aparezca las mismas tres veces y suponga un 70% del total de términos.

Para lidiar con esta problemática, aunque esto puede variar según el modelo, una de las medidas más utilizadas en el estado del arte para determinar la relevancia de un término es la medida *tf-idf* [69] la cual expresa cómo de relevante es un término dentro de una colección de documentos independientes de distinta longitud. Esta medida se utiliza a menudo como un factor de ponderación en la Recuperación de Información y la minería de textos. El valor *tf-idf* aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras. Esta medida está compuesta por el producto de dos estadísticos; la frecuencia del término en el documento y la frecuencia inversa del documento. En primer lugar, la frecuencia del término en el documento determina el número de veces que aparece un término en cuestión en un documento entre la frecuencia del término que más ocurrencias tiene en dicho documento y, en segundo lugar, la frecuencia inversa de documentos determina la importancia del término sobre toda la colección de documentos dividiendo el número total de documentos entre el número de documentos que contienen el término y se toma el logaritmo de este cociente. De esta forma, lo que finalmente se indexa es una representación del documento como un vector de pesos correspondiente a cada uno de los términos que lo componen.

2.5 Modelos de Recuperación de Información

Para definir lo que es un Modelo de Recuperación de Información y cómo funciona, en primer lugar, es necesario determinar cómo las unidades de información, en este caso documentos de texto, y las consultas están representadas de cara al sistema y cómo se comparan para obtener una lista ordenada (*ranking*) de elementos recuperados. Para ello, en esta sección, se van a presentar diversos modelos como el modelo de recuperación booleano, el cual fue la primera y más rudimentaria aproximación a la Recuperación de Información. Por otro lado, en la Sección 2.5.2 y la Sección 2.5.3 se presentaran respectivamente el paradigma del espacio vectorial para la Recuperación de Información y, por otro lado, algunos modelos probabilísticos para el mismo propósito. Estos dos últimos modelos se corresponden con aproximaciones más actuales en Recuperación de Información y, de forma general arrojan mejores resultados en las tareas de recuperación que el modelo booleano.

2.5.1 El modelo clásico de recuperación booleano

El modelo booleano fue el primer modelo desarrollado en Recuperación de Información, el cual cuenta con una larga trayectoria de aplicación [31], aunque en la actualidad sólo se utiliza en ámbitos muy específicos para propósitos muy específicos. En este modelo, los documentos están representados como un conjunto de términos a menudo pertenecientes a un vocabulario establecido previamente y extraídos mediante un proceso de indexación manual [5]. Por lo tanto, para escribir una consulta al sistema, el usuario tiene que transformar sus necesidades de información a una sentencia lógica usando los mismos términos indexados por el modelo y los operadores booleanos AND, OR y NOT. Así, siguiendo el ejemplo de la Tabla 2.1, la consulta "term2 AND term5" recuperaría los documentos *doc2* y *doc3* mientras que la consulta "term3 OR term5" recuperaría los documentos *doc2*, *doc3* y *doc4*.

	Términos Indexados				
Documentos	term1	term2	term3	term4	term5
doc1	1	0	0	0	0
doc2	1	1	0	1	1
doc3	0	1	1	0	1
doc4	1	0	1	0	1

Tabla 2.1: Índice booleano.

Sin embargo, con el objetivo de obtener una respuesta rápida por parte de este modelo, la información del índice no se almacena internamente como se ha representado en la Tabla 2.1. De hecho, a medida que la colección va creciendo, el número de términos indexados va creciendo de manera exponencial hasta el punto de que se pueden llegar a tener varios millones de términos indexados para una colección que consta de unos pocos miles de documentos. Por tanto, para obtener una respuesta más rápida, el sistema almacena un índice invertido de documentos en los

que aparecen los términos indexados [164]. La Tabla 2.2 muestra el índice invertido de la Tabla 2.1.

Términos Indexados	Documentos
term1	doc1, doc2, doc4
term2	doc2, doc3
term3	doc3, doc4
term4	doc2
term5	doc2, doc3, doc4

Tabla 2.2: Índice invertido.

La estructura del modelo booleano presenta ciertos inconvenientes que han hecho que se quede en cierto modo obsoleto en comparación con los modelos basados en espacio vectorial y los modelos probabilísticos. Quizás, la desventaja más importante es que los documentos que se recuperan no forman una lista ordenada por relevancia, en otras palabras, no se obtiene un ranking con los resultados ordenados de forma decreciente en función de como se asemejen a la consulta, y por tanto todos los documentos se consideran igualmente relevantes. Por ejemplo, considerando el ejemplo anterior de la Tabla 2.1 y considerando además la consulta "term3 OR term5", sería razonable que el primer resultado se correspondiese con aquellos documentos que albergan ambos términos, en este caso *doc3* y *doc4*, que aquellos documentos que sólo alberguen uno de ellos, cómo puede ser *doc2* en este ejemplo. Otro notable inconveniente es que en este modelo no es posible determinar si un término especifico en la consulta de un usuario es realmente importante dentro de la colección indexada o sin embargo es un término marginal, puesto que, al estar todos los términos considerados de la misma forma se hace imposible discernir cómo de importantes son dentro de la colección. Otra problemática es que el modelo booleano no puede recuperar documentos que respondan de manera parcial a una consulta, por ejemplo, la consulta "term2 AND term3" sólo devolverá *doc3* y obviará *doc2* y *doc4* aún siendo en cierto modo relevantes. Por último, otra de las principales desventajas a destacar del modelo booleano es que la correcta formulación de la consulta y los términos que la componen afecta en gran medida a los resultados que se van a obtener. Una consulta demasiado restrictiva puede conllevar a la pérdida de documentos relevantes en la recuperación y, de la misma forma, una consulta más permisiva puede derivar en una sobrecarga de información haciendo imposible detectar qué documentos de los que se han recuperado son realmente relevantes.

2.5.2 El modelo de espacio vectorial

En este modelo de Recuperación de Información [130, 131], tanto los documentos como la consultas se indexan siguiendo los pasos descritos en la Sección 2.4. De esta forma, la representación de los documentos viene determinada por un conjunto de términos con un peso que determina cómo de relevantes son dentro de la colección y los cuáles son indexados. De esta forma, al contrario

que en el modelo booleano explicado en la sección anterior, el usuario no necesita expresar sus necesidades de información usando operadores lógicos, sino más bien las consultas se realizan en lenguaje natural, mucho más intuitivo y práctico. Este tipo de modelo destaca por dar un acceso a la información más amigable para el usuario.

En el modelo de espacio vectorial, los documentos y consultas están representados por vectores en un espacio de alta dimensionalidad en el que cada término indexado se corresponde con una dimensión. Los elementos de los vectores pueden ser binarios, lo que indica la presencia o ausencia del término en cuestión, o tener un peso que indica la importancia relativa del término en el documento o en la consulta. De esta forma, el conjunto de dimensiones definidas por los términos forma así una base ortogonal (vectores linealmente independientes) y los documentos se representan como un vector en dicho espacio que empieza en el origen y termina en el punto que los valores de los pesos de los términos determinan. Así, se consigue también que los términos sean independientes entre sí, de modo que la ocurrencia de un término en un documento no implica nada sobre la presencia o ausencia de otros términos tal y como pasaba en el modelo booleano.

Por lo tanto, determinar la relevancia de un documento para una consulta viene determinado por la amplitud del ángulo que existe entre el vector que define la consulta y los vectores de los distintos documentos. Y así se puede obtener también un ranking ordenado de documentos, siendo el más relevante aquel documento que presenta una menor amplitud de ángulo con respecto a la consulta y el menos relevante aquel que presente una mayor amplitud en el ángulo con respecto a la consulta. De hecho, mediante este método se puede cuantificar cual es la similitud entre un documento y una consulta mediante la medida coseno (véase la ecuación 2.1), la cual se aplica calculando el valor que toma el coseno del ángulo que forman un documento y una consulta, siendo el valor 1 una coincidencia exacta entre documento y consulta y el valor 0 la ausencia de coincidencia alguna entre documento y consulta.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.1}$$

2.5.3 El modelo probabilístico

En lo que respecta a los modelos probabilísticos [123], la recuperación de documentos es vista en este caso como un proceso de clasificación. Para cada una de las consultas, el sistema arroja como resultado un valor de entre dos clases: relevante e irrelevante. Por lo tanto, dado un documento, el modelo probabilístico tiene que estimar la probabilidad de que el documento pertenezca a la clase relevante o a la clase irrelevante para una consulta dada. De esta forma la decisión de si un documentos es relevante y por lo tanto debe ser recuperado radica en que el modelo determine que la probabilidad de que el documento sea relevante sea mayor que la probabilidad

de que sea irrelevante. En general, los modelos probabilísticos determinan que un conjunto de documentos pertenecientes a una colección puede representarse mediante la indexación de sus términos. En este conjunto de documentos hay un subconjunto ideal con los documentos que son relevantes para una necesidad de información en forma una consulta lanzada por el usuario y que el sistema la maneja en forma de secuencia de términos ponderados. Por lo tanto, el sistema calcula la similitud entre cada documento de la colección y la consulta y a continuación muestra los resultados ordenados en función de la probabilidad de ser relevante que exista entre la consulta y cada uno de los documentos. De esta forma, a diferencia del modelo booleano, en este tipo de modelos no se lleva a cabo una comparación exacta, es decir, comprobar que todos los términos de la consulta existan o no en los documentos y, además, aportan la funcionalidad de que los usuarios puedan llevar a cabo un proceso de retroalimentación valorando de manera subjetiva los resultados del ranking de documentos recuperados devuelto por el sistema. De esta forma, el sistema es capaz de calcular las probabilidades de relevancia en las siguientes consultas, determinando así qué documentos recuperados son relevantes o no en función de si los términos de la consulta son relevantes o no.

Uno de los mayores retos que plantea la Recuperación de Información es la incertidumbre que existe en los problemas que ésta plantea resolver. En contraposición con sistemas de bases de datos donde las necesidades de información pueden ser resueltas de forma plena mediante una consulta, pues tanto consulta como sistema tienen una estructura bien definida [66], en los Sistemas de Recuperación de Información, la casuística es mucho más compleja. La información que gestionan este tipo de sistemas se encuentra de manera no estructurada y, por lo tanto, no se puede aspirar a que una consulta represente al completo las necesidades de información reales del usuario y tampoco existe un único procedimiento capaz de determinar cuáles son dichas necesidades de información [45]. Por tanto, ante estas circunstancias, los modelos probabilísticos han sido los que mejor se han ajustado para tratar los problemas de incertidumbre relacionados con la Recuperación de Información.

Dicho esto, con respecto al ámbito de los modelos probabilísticos, se han desarrollados dos grandes enfoques [46]: los modelos clásicos, donde se asigna un grado de relevancia a los documentos con respecto a una consulta, y por otro lado, el enfoque de Van Rijsbergen [121]. Cabe destacar que existen múltiples implementaciones de modelos de Recuperación de Información probabilísticos, pero de forma particular, algunos de los más relevantes en la actualidad y que se han utilizado en ciertas fases de la experimentación de esta investigación serían:

- **Okapi BM25**: En un Sistema de Recuperación de Información, BM25 [122] es una función de ranking que asigna un valor de relevancia a los documentos del sistema dada una consulta por el cual estos documentos pueden ser ordenados, siendo el primer documento recuperado del ranking el más relevante a la consulta. BM25 basa su funcionalidad en los modelos probabilísticos clásicos de la Recuperación de Información, más concretamente se basa en el modelo de *bag of words* [98] para representar los documentos que se quieren or-

denar en función de la relevancia de la consulta. Por tanto, dada una consulta Q compuesta por una serie de términos $q_1,...,q_n$, el valor de relevancia para el documento D será:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{\text{avgdl}})} \tag{2.2}$$

donde $f(q_i,D)$ es la frecuencia de aparición de los términos que aparecen en la consulta Q en el documento D , $|D|$ es el número de términos del documento D, y $avgdl$ es la longitud media de los documentos en la colección sobre la cual se está realizando la búsqueda. Los valores de la función k_1 y b se corresponden con unos parámetros que permiten ajustar la función a las características de la colección de documentos sobre la que se esté aplicando. Aunque el valor de estos parámetros viene definido en cierto modo por las características de la colección, la experiencia en el uso del modelo BM25 determina que los mejores valores para establecer dichos parámetros son $k_1 = 1.2$ o $k_1 = 2.0$ y $b = 0.75$. La función $IDF(q_i)$ es el valor de idf de las palabras de la consulta Q y se calcula, de forma general para esta función como:

$$IDF(q_i) = \log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{2.3}$$

siendo para esta última función N el número total de documentos de la colección, y $n(q_i)$ el número de documentos que contienen el término q_i de la consulta.

- **Language Model**: Un modelo de lenguaje estadístico (LM) es básicamente una distribución de probabilidad sobre un conjunto de palabras representadas de manera secuencial [158]. Por tanto, si se tiene una secuencia de palabras o términos de longitud m, LM asigna una probabilidad a cada uno de estos términos de forma que se obtiene la distribución $P(w_1,\ldots,w_m)$. Aunque LM se aplica en diversas áreas como el reconocimiento de voz, traducción automática o el reconocimiento de escritura entre otras aplicaciones, también se utiliza de manera particular en la Recuperación de Información y el etiquetado de discursos. En el contexto de la Recuperación de Información, LM genera una distribución de probabilidad sobre cada documento de la colección, de esta forma los documentos se clasifican sobre la base de la probabilidad de la consulta Q en el modelo del lenguaje del documento $P(Q|M_d)$. Existen diversas formas de generar los rankings de documentos más similares a una consulta a partir de modelos estadísticos de lenguaje, pero los más utilizados en el estado del arte y que se han utilizado en la experimentación de esta Tesis Doctoral son el *Jelinek-Mercer Smoothing Method*:

$$score(Q,D) = \sum_{w\in Q\cap D} c(w,Q)log(1+\frac{(1-\lambda)\times c(w,D)}{\lambda\times p(w|C)\times|D|}) \tag{2.4}$$

donde se realiza la suma sobre todos los términos de la consulta que coinciden con D, $|D|$ representa la longitud del documento, $c(w,D)$ es el número de veces que aparece el término w en el documento D, $c(w,Q)$ es el número de veces que aparece el término w en la consulta Q, λ es el parámetro de suavizado, y $p(w|C)$ es la probabilidad del término w estimada por el modelo de lenguaje para la colección completa. Y, por otro lado, el método *Dirichlet Prior Smoothing*:

$$P_\mu(w|\hat{\theta}) = \frac{c(w,D) + \mu p(w|C)}{|D| + \mu} \tag{2.5}$$

donde $c(w,D)$ es el número de veces que aparece el término w en el documento D, μ es el parámetro de suavizado, y $p(w|C)$ es la probabilidad del término w estimada por el modelo de lenguaje para la colección completa.

2.6 Aprendizaje Automático

El Aprendizaje Automático es un área de estudio derivada de la Inteligencia Artificial. El principal objetivo de esta ciencia es el de encontrar o detectar patrones y poder reproducirlos de forma automática por parte de un computador. Básicamente, hacer que las máquinas aprendan a partir de representaciones de la realidad que puedan entender. Para este propósito, la labor del Aprendizaje Automático se basa en el diseño de algoritmos y heurísticas para extraer información útil a partir de muestras de datos. Dentro del área de estudio del Aprendizaje Automático existen diversas categorías, pero en particular hay dos de ellas que se podrían considerar como las principales y que en consecuencia van a ser las que se van a estudiar también como aproximaciones a los problemas que se plantean en esta Tesis Doctoral.

- **Aprendizaje supervisado**: Los algoritmos de aprendizaje supervisado, de forma general, construyen un modelo estadístico a partir de un conjunto de datos cuyas observaciones tienen que estar compuestas de un conjunto de características que las definan y una etiqueta de clase que determine al conjunto al que pertenecen [127]. Para ello, el algoritmo en cuestión tiene que disponer a modo de entrada de un conjunto de entrenamiento que esté correctamente etiquetado y a partir del cual se va a aprender. De esta forma, el modelo estadístico representa a cada una de las observaciones del conjunto de entrenamiento como un vector de características, quedando el conjunto completo de entrenamiento representado como una matriz, donde cada fila representa una observación (dato) y cada columna una característica de esa observación. Así, mediante un proceso de optimización iterativa, los algoritmos de aprendizaje supervisado aprenden una función que modela cómo las características del conjunto de entrenamiento definen la clase a la que pertenece cada una de las observaciones, para predecir así la clase a la que pueden pertenecer posibles

nuevas observaciones [100]. Los algoritmos de aprendizaje supervisado abordan a su vez dos distintos enfoques: clasificación y regresión [3]. Los algoritmos de clasificación se aplican principalmente cuando en el problema que se está tratando, el número de clases del conjunto de datos está limitado a un conjunto finito de valores y, en contrapartida, los algoritmos de regresión se aplican sobre problemas donde el valor de clase puede pertenecer a un intervalo de valores numéricos reales.

- **Aprendizaje no supervisado**: Esta familia de algoritmos, también conocidos como algoritmos de detección de grupos, trabajan sobre un conjunto de datos que solo tienen definidas sus características pero no tienen definido ningún valor de clase a priori [24]. Así, los algoritmos de aprendizaje no supervisado basan su funcionamiento en encontrar una estructura en los datos de entrada a partir de la agrupación de las observaciones. Estas observaciones se agrupan a partir de ciertas similitudes que encuentra el algoritmo sobre el conjunto de datos. Este tipo de enfoque se puede definir como un problema de optimización multi-objetivo. En este tipo de algoritmos se necesita establecer una configuración previa de parámetros como la función de distancia, el numero de grupos a encontrar, el umbral de densidad, entre otros en función del tipo de algoritmo, y estos parámetros se deben estimar de forma individual para cada conjunto de datos. El proceso de aprendizaje con los algoritmos no supervisados no es tanto una tarea automática, como en el caso del aprendizaje supervisado, sino más bien un proceso iterativo que implica ensayo y error.

2.7 Medidas de evaluación

Las medidas de evaluación de un Sistema de Recuperación de Información se usan para medir cómo de bien el sistema es capaz de satisfacer las necesidades de información por parte de un usuario. Generalmente, estas medidas se aplican sobre el ranking de salida que produce el sistema a partir de una consulta de entrada y evalúa si los mejores resultados del ranking son efectivamente aquellos que se debían recuperar por ser verdaderamente relevantes a la consulta. Sin embargo, para poder aplicar las diferentes medidas de evaluación, previamente es necesario establecer unos juicios de relevancia, es decir, determinar de algún modo un criterio para que un elemento del ranking sea relevante. A continuación se presentan algunas de las medidas más utilizadas en el estado del arte [114] referente a la Recuperación de Información y en sucesivos capítulos donde se usen de forma expresa se definirá un recordatorio sobre las mismas para situar el contexto.

- ***precision***: En el contexto de los Sistemas de Recuperación de Información la medida *precision* se define sobre un conjunto de documentos recuperados por el sistema a partir de una consulta y un conjunto de documentos verdaderamente relevantes para la consulta en cuestión. Esta medida básicamente evalúa qué porcentaje de los documentos que se han

recuperado cumplen los juicios de relevancia establecidos. De forma más específica, esta medida se presenta como el cociente entre la intersección de documentos relevantes sobre los documentos recuperados y todos los documentos recuperados.

$$precision = \frac{documentos\ relevantes \cap documentos\ recuperados}{documentos\ recuperados} \tag{2.6}$$

- ***recall***: La medida de *recall*, en el contexto de la Recuperación de Información se define, de la misma forma que la *precision*, sobre un conjunto de documentos recuperados a partir de una consulta por el sistema y un conjunto de documentos verdaderamente relevantes para la consulta en cuestión. Esta medida evalúa qué porcentaje del total de documentos relevantes han sido recuperados por el sistema. De forma más concreta, esta medida se presenta como el cociente entre la intersección de documentos relevantes sobre los documentos recuperados y todos los documentos que son relevantes.

$$recall = \frac{documentos\ relevantes \cap documentos\ recuperados}{documentos\ relevantes} \tag{2.7}$$

  Cabe destacar que en esta fórmula, si se recuperan todos los documentos, el valor de *recall* va a ser siempre 1, por tanto siempre es necesario establecer un corte en el ranking a partir del cual los documentos dejan de ser relevantes y no son tenidos en cuenta para la evaluación.

- **F-measure**: Esta medida se utiliza para poder obtener una visión global de las medidas de *precision* y *recall* y poder disponer así de un valor único que contenga ambos resultados. Para ello se calcula la media armónica entre *precision* y *recall*.

$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{2.8}$$

  Esta medida toma aproximadamente el valor promedio de *precision* y *recall* cuando ambas medidas son similares, pero se ve penalizada cuando existe una diferencia notable entre ellas. De esta forma, esta medida premia a un sistema que arroje resultados balanceados, respetando al mismo tiempo la *precision* y el *recall*.

- **Normalized Discounted Cumulative Gain** (NDCG): La premisa sobre la que se basa esta medida es que los documentos relevantes que no aparecen en las partes altas del ranking se ven penalizados de forma logarítmica en función de su posición en el ranking. Esta medida evalúa en cierto modo la calidad del ranking en sí y, a diferencia de las anteriores medidas, el orden en el que se presentan los resultados del ranking es importante.

Dicho esto, la NDCG se calcula como el cociente entre el cálculo de la medida DCG como la premisa anterior y la DCG del ranking óptimo donde todos los resultados relevantes estarían a la cabeza del ranking (IDCG).

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.9}$$

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.10}$$

$$\mathrm{NDCG_p} = \frac{DCG_p}{IDCG_p} \tag{2.11}$$

Donde p representa el número de posiciones que se consideran en el ranking, rel_i representa el valor de relevancia del elemento del ranking en la posición i y REL_p representa la lista de documentos relevantes ordenados de manera decreciente por valor de relevancia en las primeras p posiciones del ranking.

2.8 Sistemas de Recomendación y Filtrado

La base sobre la que se sostiene toda la investigación de esta Tesis Doctoral, es un Sistema de Recomendación y Filtrado [119], por tanto hay que definir en primer lugar en qué consiste la Recomendación y el Filtrado y las diferencias que existen entre ambos enfoques.

En primer lugar, el enfoque de Recomendación se define, de forma general, como el proceso por el cual se extrae información de los elementos que conforman el sistema para definir un perfil que englobe las características importantes de cada uno de los elementos, y éste perfil sirve para, dada un consulta, obtener que elementos del sistema se asemejan más a dicha consulta y, si se ajustan, el sistema devuelve el elemento como resultado. De forma más específica, en este caso de estudio la Recomendación está enfocada a la Búsqueda de Expertos, es decir, el sistema contiene los perfiles de todos los Miembros del Parlamento y los usuarios externos al sistema lanzan consultas concretas (generalmente cortas) con el objetivo de obtener un listado de los Miembros del Parlamento que se ajusten en mayor medida a la consulta realizada. Cabe destacar que desde el enfoque de la Recomendación, el sistema debe ser lo más preciso posible, es decir, el usuario espera recibir un conjunto pequeño de Miembros del Parlamento que sean unívocamente relevantes a la consulta que se haya realizado (mayor *precision*).

Por otro lado, el enfoque de Filtrado tiene como objetivo seleccionar qué información debe llegar a los Miembros del Parlamento que esté relacionada con sus intereses o su actividad

política. En este caso de estudio, el objetivo es el de filtrar documentos de modo que el conjunto de todos los Miembros del Parlamento tengan un flujo constante de información pero sin llegar a estar saturados con información que pueda no interesarles. Generalmente, para este enfoque, las consultas suelen ser mucho más largas (documentos completos) y en este caso, al igual que en el enfoque anterior, se obtiene una lista con aquellos Miembros del Parlamento a los que les puede ser relevante la información, y en qué grado les puede resultar relevante. Cabe destacar además, que al contrario que con las tareas de Recomendación, con el Filtrado se pretende obtener un conjunto mayor con todos los Miembros del Parlamento que puedan estar interesados en la información de la consulta, por lo tanto el sistema debe ser exhaustivo (mayor *recall*).

Existen diversas técnicas y estrategias para construir un Sistema de Recomendación y Filtrado en función del propósito para el cual esté destinado el mismo. Entre los tipos de sistemas cuyo uso está más extendido caben destacar, entre otros, los siguientes:

- **Basados en contenido**: Probablemente este sea el tipo de Sistema de Recomendación más ampliamente utilizado en la actualidad [110] cuya aplicación se pueden encontrar, aunque formando parte de un sistema híbrido más complejo, en plataformas online altamente reconocidas y utilizadas como Youtube, Amazon, Facebook, Netflix, etc. Un Sistema de Recomendación basado en contenido tiene como principal objetivo recomendar elementos del sistema a los usuarios basándose en sus perfiles. Estos perfiles representan de algún modo las preferencias o intereses de los usuarios y estos vienen definidos a partir de los elementos que el propio usuario ha definido como relevantes por sí mismo. El perfil del usuario, en este tipo de sistema, se puede definir de diversas formas, por ejemplo, en el caso de estudio de esta Tesis Doctoral, los perfiles se forman a partir de las intervenciones que los propios Miembros del Parlamento han llevado a cabo en los debates parlamentarios, asumiendo que estas intervenciones contendrán la información que le es relevante al Miembro del Parlamento. En otros escenarios, las recomendaciones del sistema vienen dadas por elementos que han sido relevantes para el usuario en un momento dado y han pasado a formar parte de su perfil, como por ejemplo, los artículos que recomienda Amazon y que están relacionados con búsquedas o compras anteriores en la misma plataforma.

- **Filtrado colaborativo**: En este tipo de sistemas [119], los propios usuarios comparten sus intereses entre ellos, de manera que se lanza la hipótesis de que usuarios relacionados entre si, de algún modo, tienen los mismos intereses. Estos sistemas operan buscando sobre un grupo generalmente grande de usuarios con el objetivo de detectar patrones sobre gustos e intereses similares en conjuntos de estos usuarios que puedan ser parecidos a los de un usuario en particular. Por tanto, el sistema observa los intereses comunes de este conjunto de usuarios para hacer predicciones a un usuario en particular que comparta estos intereses con ellos. Este tipo de sistemas ha tenido una gran acogida en el ámbito de la publicidad y el marketing, donde empresas perfilan a los potenciales usuarios en función de

las comunidades a las que pertenecen y optimizar de esta forma las campañas publicitarias y hacer recomendaciones más personalizadas a los usuarios.

- **Basados en conocimiento**: Este tipo de sistemas [19] está principalmente enfocado a un tipo de contexto donde no existe un historial de preferencias o intereses previo ni se dispone de un perfil del mismo. Para lidiar con esto, estos sistemas implementan modelos de conocimiento de forma que las recomendaciones no vienen dadas por el perfil del usuario como tal, sino a partir de una consulta explícita al sistema. Un claro ejemplo de este tipo de sistemas se puede encontrar en los portales de búsqueda de casas o coches, donde el sistema no conoce nada sobre el usuario que lo está utilizando, pero el usuario se encarga de definir unas ciertas características específicas de lo que espera encontrar como respuesta a sus necesidades de información. Una de las mayores fortalezas de este tipo de sistemas es que no se produce el problema de inicio frío (*cold start*) que se produce en los casos de sistemas anteriores, es decir, el problema derivado de no tener mucha información sobre el usuario para poder hacer buenas recomendaciones.

- **Sistemas híbridos**: Los Sistemas de Recomendación Híbridos [20] basan su funcionamiento en el uso de una combinación de varias técnicas como las explicadas anteriormente (y alguna más). El objetivo de este tipo de sistemas es el de maximizar los puntos fuertes de las diferentes técnicas para obtener mejores recomendaciones. Ciertamente, salvo en sistemas que estén destinados a un propósito específico, los sistemas más prácticos y utilizados en la actualidad pertenecen a esta familia. Los Sistemas de Recomendación y Filtrado más potentes son, efectivamente, un compendio de las virtudes del Filtrado Colaborativo y al mismo tiempo de los Sistemas Basados en Contenido. De esta forma, las recomendaciones del sistema no vienen únicamente dadas por los propios intereses del usuario sino que también por los intereses del resto de usuarios similares al usuario en cuestión. Netflix, por ejemplo, de forma simplificada basa sus recomendaciones en series o películas que han sido de interés para un usuario y además en los intereses de otros usuarios que a su vez han estado interesados en las mismas series o películas. Diversos estudios donde se comparan de manera empírica el rendimiento de Sistemas Híbridos con Sistemas Colaborativos o Basados en Contenido puros han demostrado que con el enfoque híbrido se obtienen recomendaciones más precisas que con los enfoques puros.

2.9 Especificaciones del Sistema

Dependiendo del contexto donde el sistema se vaya a aplicar, el diseño de este puede variar notablemente, pero de forma general, existe una estructura común para todos los Sistemas de Recuperación de Información la cual también sigue en mayor o menor medida el sistema que en esta Tesis Doctoral se describe. En primer lugar se tiene una fuente de información, en este caso de estudio las iniciativas parlamentarias, con las que se alimenta al sistema. El sistema

estructura dicha información en función de las necesidades de representación y los requerimientos de los usuarios, en el sistema que se discute en esta Tesis Doctoral se extraen las intervenciones de los Miembros del Parlamento y se construyen una serie perfiles para representar sus intereses políticos. Una vez se tiene la representación de la información dentro el sistema, los usuarios pueden lanzar consultas con el propósito de obtener información relevante en relación a lo requerido y, aunque el sistema que aquí se presenta puede recibir cualquier tipo de consulta, por cuestiones de evaluación de los experimentos se ha considerado el extracto de las iniciativas como consulta para medir el rendimiento del enfoque de Recomendación y, por otro lado, el contenido de todas las intervenciones de cada iniciativa como consulta para evaluar el enfoque de Filtrado.

Una vez el sistema recibe y procesa una consulta, este devuelve como resultado una lista ordenada de forma decreciente con subperfiles de los Miembros del Parlamento, es decir un ranking con los Miembros del Parlamento cuyos intereses son más parecidos a la consulta que se ha realizado, y un valor asociado que determina el grado de similitud (score). Cabe destacar que, tanto en una de las representaciones base, como en las aproximaciones de representación que se plantean es esta Tesis Doctoral, un perfil está compuesto por varios subperfiles independientes, por tanto, puede ocurrir que en un ranking aparezcan diversas instancias de un mismo Miembro del Parlamento haciendo referencia a diferentes subperfiles. En ese caso se deben aplicar ciertos criterios de agregación como pueden ser quedarse con el subperfil cuyo score es mayor, reestructurar el ranking con la suma de los scores de un mismo Miembro del Parlamento, etc. En las secciones que hagan referencia a la evaluación experimental de los siguientes capítulos se explicará con más detalle el método de agregación que se ha tenido en cuenta.

Por último, definir que la colección de documentos con la que se va a alimentar el Sistema de Recomendación y Filtrado y, en definitiva, con la que se va a llevar a cabo toda la experimentación de esta Tesis Doctoral, se corresponde con la conjunto de iniciativas parlamentarias de la octava legislatura del Parlamento Regional de Andalucía. Esta colección está constituida por un conjunto de 5.258 documentos XML donde se recogen cada una de las iniciativas parlamentarias, lo que supone un total de 12.633 intervenciones distintas, solo considerando los 132 Miembros del Parlamento que han intervenido al menos en 10 ocasiones.

2.10 Perfiles de los Miembros del Parlamento

Hasta el momento se ha referido en algunas ocasiones la idea de perfil para representar al MP de cara al Sistema de Recomendación y Filtrado. De manera convencional, un perfil se define como un conjunto de rasgos o características que representan a algo o alguien y, en el contexto de la Recomendación y Filtrado, la clave reside en determinar qué información puede ser útil para representar a los elementos del sistema, en este caso los MPs.

Tal y como se ha establecido en la sección anterior, la fuente de información de donde se extraen las características que definen a un MP son las transcripciones literales de sus

intervenciones en el parlamento. Cabe esperar que en estas intervenciones se puedan encontrar los intereses políticos de un MP puesto que el mismo va a incluir asuntos referentes a su labor parlamentaria en ellas. Por tanto, una vez se dispone de la fuente de información, la cuestión pasa a ser la forma en la que esta información se representa en el Sistema de Recomendación y Filtrado para que, no solo funcione correctamente, sino obtener el mayor rendimiento posible. Hasta el momento, uno de los aspectos más relevantes de la representación de los perfiles está claro; los perfiles deben ser documentos de texto, no solo porque es la forma en la viene representada la información originalmente sino porque el objetivo principal del sistema es que los usuarios introduzcan y reciban información en este formato.

En publicaciones previas a esta Tesis Doctoral, como una primera aproximación al problema, se definen los perfiles de los distintos MP de dos formas opuestas y ciertamente extremas entre sí. La primera de ellas consiste en definir el perfil del MP como un conjunto de subperfiles, es decir un conjunto de documentos independientes, donde en cada uno de ellos se recoja cada una de sus intervenciones de forma individual. La segunda aproximación consiste en reunir todas las intervenciones parlamentarias del MP dando lugar a que el perfil esté solo constituido por un único documento. Ambas aproximaciones tienen sus ventajas y desventajas y, en cuestiones de evaluación, son prácticamente opuestas entre sí, por tanto, esto llevó a uno de los principales planteamientos de esta Tesis Doctoral, que es el de encontrar una representación alternativa de los perfiles de los MPs con el objetivo de maximizar las ventajas de las aproximaciones anteriores.

Tal y cómo se ha explicado en la Sección 2.4, una vez se tiene el conjunto de términos de la colección de intervenciones con sus raíces semánticas, se necesita que estos términos queden representados de forma que los algoritmos y técnicas que se van a aplicar para construir los perfiles puedan procesarlos. Para este propósito se han utilizado, o bien los documentos como tal para indexarlos, o bien la representación clásica *Document Term Matrix* (dtm) en el caso de utilizar algoritmos de Aprendizaje Automático, la cual se caracteriza por definir la colecciones de documentos como una matriz donde cada una de las filas se corresponde con cada uno de los documentos de la colección y cada una de las columnas se corresponde con un término de la colección. De esta forma, cada una de las celdas de la matriz tiene un valor entero que se corresponde con el número de apariciones de un término en un documento y, en función del algoritmo que se vaya a utilizar, este valor podrá ser normalizado de algún modo.

CAPÍTULO 3

CAPÍTULO

COMPARATIVA DE APROXIMACIONES BASADAS EN APRENDIZAJE AUTOMÁTICO Y RECUPERACIÓN DE INFORMACIÓN PARA EL FILTRADO DE DOCUMENTOS

En este capítulo se plantea el problema de construir un Sistema de Recomendación y Filtrado basado en contenido en un contexto parlamentario con el objetivo de que, llegado un nuevo documento para ser recomendado, el sistema deba ser capaz de discernir qué Miembros del Parlamento podrían valorar positivamente la recepción de este documento en función de sus intereses o labor política. Para tal propósito se proponen, con el objetivo de realizar una comparativa, dos aproximaciones distintas para abordar el problema; la primera de ellas se corresponde con la implementación de un método basado en Aprendizaje Automático donde los documentos se clasifican de forma automática y, por otro lado, un método basado en la aplicación de técnicas de Recuperación de Información, el cual basa su funcionamiento en el emparejamiento del documento con la representación de los perfiles de los Miembros del Parlamento.

3.1 Introducción

En el ámbito político en general y en lo que corresponde a un Parlamento en particular, los servidores públicos necesitan estar al corriente, practicamente en tiempo real, de la realidad del país, región o territorio donde desempeñan su actividad política. Sin embargo, no toda la información que se genera en el contexto político, la cual puede llegar a ser ingente, tiene por qué ser relevante de forma directa si nos basamos en los intereses políticos de los diferentes Miembros de un Parlamento. Por ejemplo, un Miembro del Parlamento cuya labor política esté vinculada a asuntos relacionados con la sanidad puede estar especialmente interesado en recibir información sobre los temas que giran en torno a ese campo y, por otro lado, el hecho de ser desconocedor de la información generada en referencia a otros aspectos políticos como agricultura o educación no afecte al desempeño de su trabajo. Sin embargo, cabe destacar que en el presente, la cantidad de información que se genera a cada instante y la cual está disponible a través de las Tecnologías de Información y Comunicación adquiere un volumen enorme, lo cual no solo ha suscitado la necesidad de gestionarla de forma eficiente sino que también es necesario abordar la difícil tarea de decidir qué información es interesante y qué información no lo es en función de los intereses de los usuarios. Tal y como Shamin y Neuhold propusieron en [138], en el contexto del Parlamento Europeo, *"Members of the Parliament need to be selective in their information input"*.

Consideremos que se debe distribuir un flujo constante de documentos entre un conjunto de Miembros del Parlamento y estos documentos pueden ser, por ejemplo, noticias de prensa, iniciativas parlamentarias o intervenciones de otros Miembros del Parlamento entre otros. A partir de esto, se pretende construir un sistema automático que sea capaz de, llegado un nuevo documento, seleccionar qué Miembros del Parlamento pueden considerarlo como información relevante, tomando como criterio para determinar la relevancia el propio contenido textual del documento y los intereses políticos y preferencias de cada uno de los Miembros del Parlamento.

El objetivo principal de esta Tesis Doctoral es el de realizar un análisis y una posterior comparativa de las virtudes de aplicar técnicas basadas en Aprendizaje Automático o basadas en los métodos clásicos de Recuperación de Información en el ámbito parlamentario que se considera en este caso de estudio. De esta forma, se proponen dos aproximaciones basales para el diseño de un sistema de información. La primera de las aproximaciones que se han considerado hace uso de un Sistema de Recuperación de Información con el objetivo de explorar las características de la colección de documentos de entrenamiento y, por otro lado, la segunda aproximación utiliza la colección con el propósito de generar un conjunto de clasificadores binarios, uno por cada Miembro de Parlamento. En referencia a la colección de documentos de la que se hace uso para construir ambas aproximaciones, se obtiene de las transcripciones literales de los discursos de los MPs en los debates parlamentarios y se asume que, a partir de esta información y la forma en la que está representada, se pueden extraer los intereses y preferencias en el contexto político de los propios Miembros del Parlamento. Con el objetivo de comparar ambas aproximaciones de forma experimental se ha utilizado una colección de textos que se corresponden con las intervenciones

parlamentarias de los MPs del Parlamento de Andalucía en su octava legislatura.

3.2 Aproximaciones para la recomendación

El escenario que se considera en este caso de estudio es el siguiente: se tiene un conjunto de Miembros del Parlamento $\mathscr{MP} = \{MP_1, \ldots, MP_n\}$. Se tienen además un conjunto de documentos que llegan al parlamento con el objetivo de ser distribuidos entre los MPs en función a sus intereses y preferencias políticas. Dicho esto, se pretende construir un sistema que de forma automática y llegado un documento nuevo, sea capaz de seleccionar aquellos Miembros del Parlamento a los que la información de dicho documento les sea relevante y puedan estar interesados en recibirlo. Asociados a cada Miembro del Parlamento MP_i hay un conjunto de documentos $\mathscr{D}_i = \{d_{i1}, \ldots, d_{im_i}\}$ donde d_{ij} representa las transcripciones literales de cada una de sus intervenciones en los debates parlamentarios. Y finalmente, la colección completa de documentos se representa como $\mathscr{D} = \cup_{i=1}^{n} \mathscr{D}_i$, y se corresponde con la colección de documentos de entrenamiento que va a ser utilizada con el propósito de construir las aproximaciones basadas en Recuperación de Información y Aprendizaje Automático.

3.2.1 La aproximación basada en Aprendizaje Automático

La idea sobre la que se aborda el diseño de esta aproximación es relativamente sencilla; hacer uso de las transcripciones literales de los discursos de los MPs en sus intervenciones en el parlamento, $\mathscr{D}$, como conjunto de entrenamiento a la hora de construir un clasificador binario (relevante/irrelevante) para cada uno de los Miembros de Parlamento. A continuación, llegado un nuevo documento que debe ser filtrado o recomendado, se hace uso de los clasificadores entrenados para discernir de forma automática qué Miembros del Parlamento deben ser receptores de la información que figura en el documento en caso de que el clasificador devuelva como positiva la clase relevante o, de forma alternativa, asumiendo que los clasificadores devuelven un valor numérico, considerado como el grado de pertenencia a la clase positiva, comprendido en el intervalo [0,1] y de esta forma recomendar el documento en cuestión a los Miembros del Parlamento cuyo grado de pertenencia a la clase positiva sea mayor que un umbral fijado.

Con el objetivo de construir un clasificador binario estándar para cada Miembro del Parlamento es necesario disponer de un conjunto de entrenamiento que en este caso de estudio se corresponde con una colección de documentos. A su vez, el conjunto de entrenamiento tiene que estar dividido de forma disjunta en un subconjunto de instancias positivas (documentos relevantes) y otro subconjunto de instancias negativas (documentos irrelevantes). En este capítulo se considera, como primera forma de acercamiento, que las propias intervenciones o discursos de un Miembro del Parlamento se corresponden con las instancias positivas y el resto de intervenciones del resto de MPs se consideran el subconjunto de entrenamiento negativo. De este modo, desde el punto de vista de la nomenclatura, para cada Miembro del Parlamento MP_i, el subconjunto de

instancias positivas es $\mathscr{D}_i$ y, por lo tanto, el subconjunto de instancias de entrenamiento negativas de cada MP_i se corresponde con $\mathscr{D} \setminus \mathscr{D}_i$.

3.2.2 La aproximación basada en Recuperación de Información

En el caso de esta aproximación la cual está basada en la aplicación de las técnicas clásicas para la resolución de este tipo de problemas, se van a abordar dos formas distintas de entrenar un Sistema de Recuperación de Información (IRS). Por otro lado, la misión de este IRS va a estar basada en la recuperación de documentos de entrenamiento de los Miembros del Parlamento que sean más similares a los documentos que se quieren filtrar o recomendar llegados nuevos al sistema, los cuáles hacen el papel de consulta. En referencia a las dos formas en las que se ha transformado el conjunto de entrenamiento $\mathscr{D}$ para pasar a ser una colección indexada de documentos, las cuáles han sido propuestas originalmente en [39], son las siguientes:

La Colección de Intervenciones de los Miembros de Parlamento. Los documentos que indexa el IRS son los que se comprenden dentro de $\mathscr{D}$, es decir, tomando cada una de las intervenciones de cada MP como documentos independientes. A partir de esta idea, si lanzamos una consulta al sistema, en forma de un nuevo documento para ser filtrado o recomendado, lo que se obtiene como resultado es un ranking de documentos ordenado de forma decreciente donde cada uno de ellos está asociado con un Miembro de Parlamento. Cabe destacar que en dicho ranking se pueden encontrar MPs duplicados en el sentido en que más de una intervención de dicho MP, al ser consideradas como documentos independientes, puede entrar dentro del ranking al verse evaluadas por el sistema como relevantes. Dicho esto, con el objetivo de que el sistema devuelva un ranking ordenado de MPs donde cada aparición sea única, se eliminan todas las ocurrencias del MP en el ranking salvo la primera aparición, la cual se corresponde a la que tiene el mayor valor de score. En otros capítulos se considerarán formas alternativas de agregar las distintas apariciones de intervenciones de un mismo MP, pero en este punto de la investigación, el criterio de máximo servirá como base para realizar los experimentos. Esta aproximación se va a denominar IR-i.

La Colección de Perfiles de los Miembros del Parlamento. Con el objetivo de solventar los problemas derivados del anterior enfoque, donde nos encontrábamos con instancias duplicadas de un mismo MP las cuáles debían ser eliminadas del ranking que devolvía como salida el IRS, se propone un enfoque alternativo que consiste en la agrupación de todas las intervenciones de cada Miembro del Parlamento de forma que solo se disponga de un único documento por MP, al que llamaremos perfil monolítico, y por lo tanto se dispondrá de igual número de documentos en la colección de entrenamiento que de MPs. De forma más precisa se tiene que por cada conjunto de intervenciones de un mismo MP, $\mathscr{D}_i$ se construye como un único documento $d_i = \cup_{j=1}^{m_i} d_{ij}$ para posteriormente usar como la colección de documentos que se va a indexar por el IRS $\cup_{i=1}^{n} d_i$. De esta forma, el ranking que se obtiene como salida tras lanzar una consulta al sistema es directamente una lista de perfiles de Miembros del Parlamento únicos ordenada de forma

decreciente. En este caso, IR-p será el nombre que recibirá esta aproximación.

En los dos casos que se han considerado como enfoques (ML, IR), el sistema devuelve como resultado una lista ordenada de Miembros del Parlamento en orden decreciente en función de la similitud que exista entre el documento que se ha usado como consulta y el conjunto de documentos indexados, con independencia de como la consulta esté representada. Sin embargo, el IRS no tiene en cuenta la longitud de la consulta, y como consecuencia el valor de los scores del ranking de salida puede ser mayor o menor en función del número de términos de los que está compuesta la consulta, y estos valores no están normalizados. A pesar de que los scores no estén normalizados, en principio, los rankings de Miembros de Parlamento obtenidos son completamente válidos puesto que lo importante es el orden en el que aparecen listados. Pero de forma particular, en este sistema, uno de los propósitos que se plantean es el de encontrar un umbral común que pueda ser utilizado para recomendar aquellos documentos de un MP (ya sea en forma de intervención independiente o perfil completo) cuyo valor de score sea mayor que el valor fijado en el mencionado umbral con independencia de la longitud en términos de las consultas que se lancen contra el IRS. Para este propósito, sí es necesario que los scores estén normalizados, ya que en caso contrario habría que fijar un umbral distinto para cada consulta en función de su tamaño. En referencia al valor del umbral y cómo establecer su valor, se procede normalizando el valor de los scores, quedando cada valor del ranking dividido entre el valor de score de la cabeza del ranking y, por lo tanto el uso de los scores normalizados toma un nuevo significado, pasando a considerarse una medida de porcentaje de similitud con respecto al Miembro del Parlamento que queda a la cabeza en el ranking, es decir, el MP a la cabeza del ranking siempre tendrá el valor de score 1 y los MPs en las posiciones consecutivas del ranking tendrán un valor de score en proporción al primer MP.

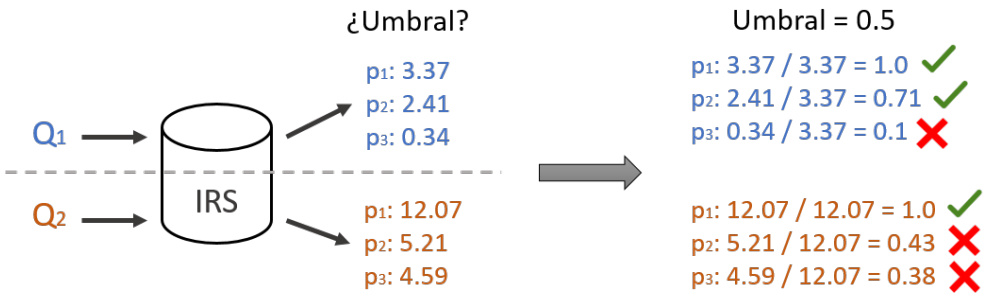

Figura 3.1: Proceso de normalización de los scores.

3.3 Evaluación Experimental

Para la evaluación de las distintas aproximaciones que se plantean en este trabajo se va utilizar la colección de iniciativas parlamentarias de la octava legislatura del Parlamento de Andalucía,

la cual consta de un total de 5.258 documentos, los cuáles están formados con estructura XML [42].

Cada una de las iniciativas parlamentarias contiene, entre otra serie de conceptos, las transcripciones literales de todos los discursos o intervenciones de los Miembros de Parlamento que alguna vez hayan intervenido en los debates parlamentarios, las cuáles ascienden a un total de 12,633 intervenciones distintas, solo considerando aquellos 132 MPs cuyas intervenciones hayan superado al menos las 10 veces, con el objetivo de evitar trabajar con Miembros del Parlamento que tengan una baja actividad y puedan contaminar con ruido los resultados de la experimentación. Todas las intervenciones de los MPs han sufrido una sucesión de procesamientos entre los que se comprenden la eliminación de stopwords (palabras carentes de significado semántico) y se ha llevado a cabo un proceso de stemmización (proceso de obtención de la raíz semántica) de los términos.

En referencia a la metodología de evaluación, se ha utilizado el método de $repeated\ holdout$ [76]. De forma más concreta se ha particionado de forma aleatoria el conjunto de iniciativas en dos subconjuntos disjuntos de entrenamiento, 80% del total de iniciativas, y de test, el 20% restante, realizando este proceso repetidamente (5 veces para este caso) con el propósito de obtener una medida promedio de los resultados de las distintas particiones.

Dicho esto, con el objetivo de construir la colección de documentos de entrenamiento $\mathscr{D}$, se extraen todas las intervenciones de todos los Miembros de Parlamento que estén contenidas en las iniciativas parlamentarias del subconjunto de entrenamiento. Una vez este proceso se ha llevado a cabo, se construye un clasificador binario por cada Miembro del Parlamento en el caso de la aproximación basada en Aprendizaje Automático (explicada en la Sección 3.2.1), o bien un IRS con los documentos de la colección (ambas formas descritas en la Sección 3.2.2) siguiendo la aproximación basada en Recuperación de Información. Tal cómo se ha referido, con el objetivo de entrenar un clasificador binario para cada MP_i a partir de sus instancias positivas $\mathscr{D}_i$ y sus instancias negativas $\mathscr{D} \setminus \mathscr{D}_i$, se ha hecho uso del algoritmo Support Vector Machine [35], el cual se considera en la literatura como la técnica más utilizada a la hora de ser aplicada en clasificación automática de documentos (se ha usado la implementación de SVM disponible en R). Desde la perspectiva de la Recuperación de Información se ha utilizado el clásico modelo de Recuperación de Información BM25 (que se corresponde con el motor de búsqueda utilizado por la librería Lucene), el cual es, de la misma forma, considerado como una técnica de referencia en la literatura al respecto de la recuperación de documentos [7].

En lo referente a las iniciativas pertenecientes al subconjunto de test, estas se van a utilizar como documentos destinados a ser filtrados o recomendados (solo usando las transcripciones de los discursos o intervenciones que están contenidas en el texto del documento). Para definir de alguna forma un criterio por el cual una iniciativa pueda resultar relevante o irrelevante para un MP en cuestión, sólo las iniciativas en las que haya participado de forma activa un MP, es decir, interviniendo en ellas, van a ser consideradas como relevantes. Este criterio es muy

conservador en el sentido en que un MP puede estar interesado en una iniciativa en la que no haya intervenido, ya sea porque se tratan temas relativos a la comisión donde participa, o bien por ser interpelado por algún otro Miembro del Parlamento, o algún otro caso similar.

Las medidas usadas para evaluar el rendimiento de las aproximaciones que se plantean para la construcción de un sistema de filtrado o recomendación son las clásicas en el campo de estudio de la clasificación textual.

precision: Cociente entre el número de documentos recuperados que son relevantes y el total de documentos recuperados. Es decir, la probabilidad de que un documento recuperado por el sistema sea relevante.

recall: Cociente entre el número de documentos verdaderamente relevantes recuperados y el total de documentos relevantes en la colección. Es decir, la probabilidad de que un documento verdaderamente relevante sea recuperado por el sistema.

F-measure: Esta medida es un estadístico que se emplea en Recuperación de Información para obtener un valor único ponderado de las medidas de *precision* y *recall*.

Se tiene en cuenta que para nuestro sistema, una instancia que sea verdadera positiva TP_i es aquella en la que el MP ha participado de forma activa y el valor de score es mayor que el umbral fijado, un falso positivo FP_i ocurre cuando el MP no ha participado en la iniciativa pero aún así su valor de score es mayor que el umbral fijado y un falso negativo FN_i se produce cuando un MP si ha participado en la iniciativa pero su valor de score es menor que el umbral fijado (véase Tabla 3.1). En nuestro caso de estudio, la *precision* se calcula llevando a cabo el cociente entre el número de instancias que han sido correctamente clasificadas como verdaderas positivas para un MP_i y el número de iniciativas consideradas por el sistema como relevantes para un MP_i, es decir, $TP_i + FP_i$, $p_i = TP_i/(TP_i + FP_i)$. En el caso del *recall*, este se calcula con el cociente del número de instancias clasificadas como verdaderas positivas TP_i y el número de iniciativas de test que son verdaderamente relevantes para el Miembro del Parlamento MP_i, por lo tanto, $TP_i + FN_i$, $r_i = TP_i/(TP_i + FN_i)$. Con estas dos medidas se puede proceder al cálculo de la *F-measure*, en concreto *F1*, como la media armónica de la *precision* y el *recall*, $F_i = 2p_i r_i/(p_i + r_i)$. Finalmente, en lo que a las medidas de evaluación se refiere y con el objetivo de tener una visión más a nivel global del comportamiento del sistema, se propone el cálculo de las medidas macro-averaged (M) y micro-averaged (m) [144].

	Verdadero relevante	Verdadero irrelevante
score ≥ umbral	TP_i	FP_i
score < umbral	FN_i	TN_i

Tabla 3.1: Relaciones entre TP_i, FP_i y FN_i con la verdadera relevancia de los documentos a ser recomendados y su valor de score.

$$Mp = \frac{1}{n}\sum_{i=1}^{n} p_i \qquad Mr = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad MF = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{3.1}$$

$$mp = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)} \qquad mr = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)} \qquad mF = \frac{2mp\,mr}{mp + mr} \tag{3.2}$$

Todas las medidas que se proponen para evaluar la calidad del sistema dependen en un grado muy alto de la elección del valor del umbral que se utiliza con el objetivo de filtrar o recomendar documentos solo a aquellos MPs cuyo valor de score sea superior a dicho umbral. Aunque el valor de umbral natural para la ejecución de los experimentos podría ser 0.5, con la intención de observar el comportamiento del sistema, se va a llevar a cabo el proceso de experimentación con diferentes valores para el umbral, utilizando valores comprendidos en el rango desde 0.1 hasta 0.9. Cabe destacar que los valores de score obtenidos por el sistema en su variante basada en Aprendizaje Automático y su variante basada en Recuperación de Información tienen distintos significados. Mientras que en el primer caso representan probabilidad de pertenencia a la clase positiva, en la aproximación de Recuperación de Información representan la similaridad con respecto al mejor resultado, y de la misma forma, los valores que se fijan para el umbral son probabilidad de pertenencia a la clase positiva para la aproximación basada en Aprendizaje Automático y similaridad con respecto al mejor resultado en la aproximación de Recuperación de Información.

Por otro lado, con el propósito de evaluar el sistema desde el punto de vista de la calidad del ranking obtenido como resultado, al margen de los valores que pueda presentar el umbral, se plantea una medida de evaluación alternativa: la medida Normalized Discounted Cumulative Gain (NDCG) [63]. Esta métrica se utiliza de forma muy asidua en la literatura referente al ámbito de la Recuperación de Información puesto que los usuarios del sistema tienden, de forma general, a prestar atención solo a aquellos resultados que quedan en la parte alta del ranking. De esta forma se añade un factor de descuento progresivo con respecto a como desciende la posición del elemento en la lista, lo cual provoca que la relevancia del documento con respecto a su posición en el ranking se vea afectada negativamente. El valor de esta métrica para una lista ordenada de forma decreciente de Miembros de Parlamento se calcula de la siguiente forma:

$$NDCG@k = \frac{1}{N}\sum_{i=1}^{k} \frac{2^{rel(d_i)} - 1}{\log(i+1)}, \tag{3.3}$$

donde k se corresponde con el número de elementos de la lista evaluados (10 en nuestros experimentos); i es el valor de la posición en el ranking para el Miembro del Parlamento que está siendo evaluado; el MP en la posicion i es d_i; $rel(d_i)$ es el valor de relevancia de d_i (puede tomar el valor 0 o 1 en nuestro caso); el factor de normalización N es la DCG para un ranking

óptimo, donde todos los resultados considerados como relevantes se encuentran posicionados de forma consecutiva en las primeras posiciones del ranking. Con este tipo de normalización, los valores de la medida oscilan entre los valores del intervalo [0,1], haciendo posible el cálculo de las medias entre documentos distintos. La medida NDCG se calcula de forma independiente sobre los rankings que se obtienen tras lanzar cada uno de los documentos de test sobre el sistema a modo de consultas y finalmente realizando un promedio entre todos los resultados.

3.3.1 Resultados

Los resultados del proceso de experimentación para las medidas macro (M) y micro (m) sobre *precision*, *recall* y *F-measure*, están representados en las Figuras 3.2, 3.3, 3.4 respectivamente.

A la vista de los resultados, se puede observar de forma clara que, por lo general, cuando el umbral se fija en los valores más bajos del rango, el sistema adopta como relevantes a un mayor número de documentos, lo cual provoca un claro y lógico incremento del número de falsos positivos, lo cual tiene como resultado que la métrica que evalúa la *precision* del sistema se vea afectada negativamente y por otro lado, se produce un decremento del número de falsos negativos, lo cual se traduce en un considerable incremento del valor del *recall*. Sin embargo, cuando se fija el umbral en valores altos, se produce la situación opuesta, es decir, se reduce el número de falsos positivos y por lo tanto la *precision* aumenta pero en detrimento de elevar el número de falsos negativos lo cual conlleva la pérdida en la medida de *recall* del sistema. Cabe destacar que, de forma anómala, este comportamiento general se ve truncado en la aproximación basada en Aprendizaje Automático en lo que respecta a la macro *precision*, la cual tiende a decrecer a medida que el valor del umbral es más alto. Esta casuística se debe a que el sistema, en esta aproximación, actúa de forma negligente con aquellos MPs que tienen un número bajo de intervenciones, lo cual tiene como consecuencia un conjunto de entrenamiento para dichos MPs muy pobre y de mala calidad y por lo tanto, se produzca un decremento de documentos clasificados como verdaderos positivos más notable con respecto a la disminución del número de falsos positivos a medida que el umbral va aumentando su valor (destacando que en las métricas macro todos los Miembros del Parlamento son considerados de la misma forma independientemente del número de veces que hayan intervenido en los debates parlamentarios). En la siguiente sección se podrá visualizar con más detalle esta situación y cómo está relacionada con respecto al número de intervenciones de los MPs.

Aproximación	ML	IR-i	IR-p
Umbral	0.1	0.8	0.9
mF	0.2978	0.2896	0.2829
MF	0.2475	0.2423	0.2513
NDCG@10	0.6263	0.6246	0.6776

Tabla 3.2: Mejores valores obtenidos por ML, IR-i y IR-p para macro y micro *F-measure* y NDCG@10.

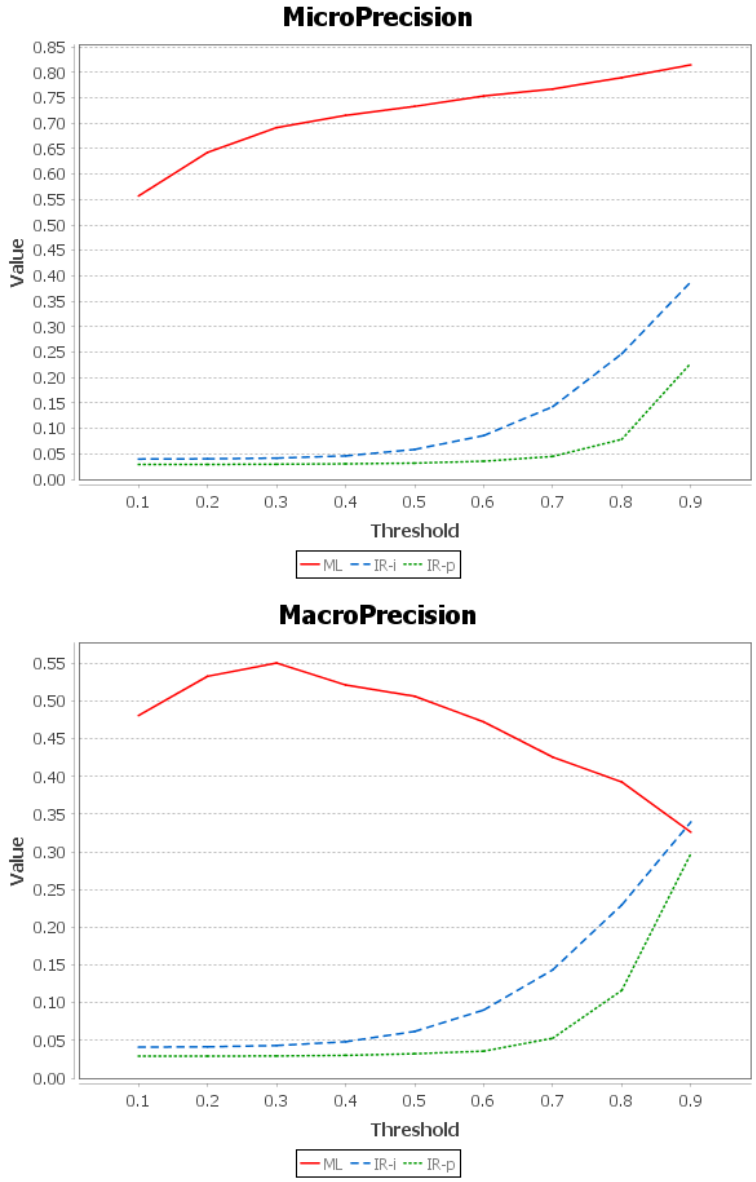

Figura 3.2: Micro y macro *precision* para ML, IR-i y IR-p usando diferentes umbrales.

Sin embargo, comparando las dos aproximaciones, el comportamiento de ambas a efectos prácticos es notablemente distinto. La aproximación cuyo enfoque esta basado en técnicas de Aprendizaje Automático destaca por obtener mejores valores en lo que a *precision* se refiere, mucho mejores a los que se obtienen en la aproximación basada en Recuperación de Información. No obstante, en contraposición a la métrica de *precision*, los valores de *recall* en la aproximación basada en Aprendizaje Automático son bastante malos y sin embargo, una vez más ocurre lo contrario en la aproximación basada en Recuperación de Información, la cual tiene unos valores para la métrica de *recall* considerablemente mejores. En referencia a los dos distintos enfoques de la aproximación basada en Recuperación de Información (IR-i y IR-p) el comportamiento es

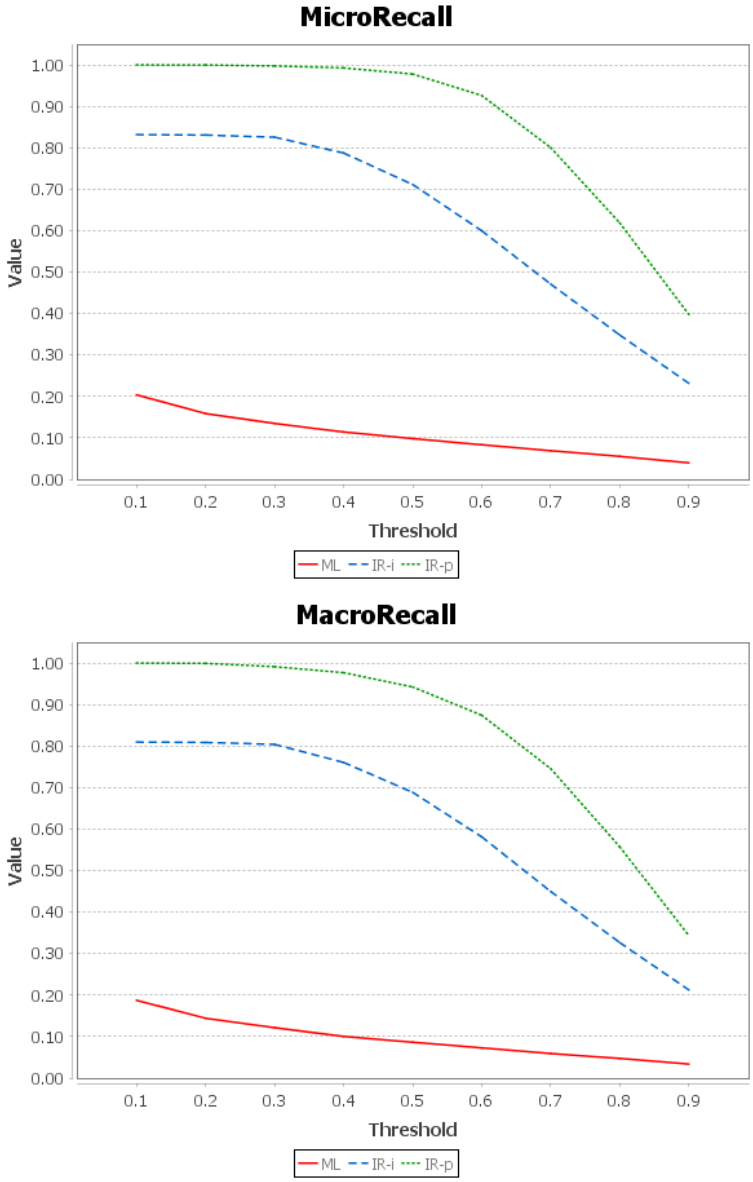

Figura 3.3: Micro y macro *recall* para ML, IR-i y IR-p usando diferentes umbrales.

muy similar y, aunque sería difícil discernir cual de ellas se comporta de mejor manera desde el punto de vista del rendimiento, se puede remarcar que el enfoque IR-p obtiene valores más extremos con respecto al enfoque IR-i (mejor comportamiento en *recall* y peor en lo referente a la *precision*).

La *F-measure*, la cual se utiliza para obtener un balance y una visión más global que la que aportan las medidas de *precision* y *recall* de forma individual, detecta de forma clara que la aproximación basada en Aprendizaje Automático se comporta mejor con valores de umbral bajos y lo contrario ocurre con la aproximación basada en Recuperación de Información, la cual obtiene mejores resultados cuanto más alto es el valor del umbral fijado. No obstante, con la

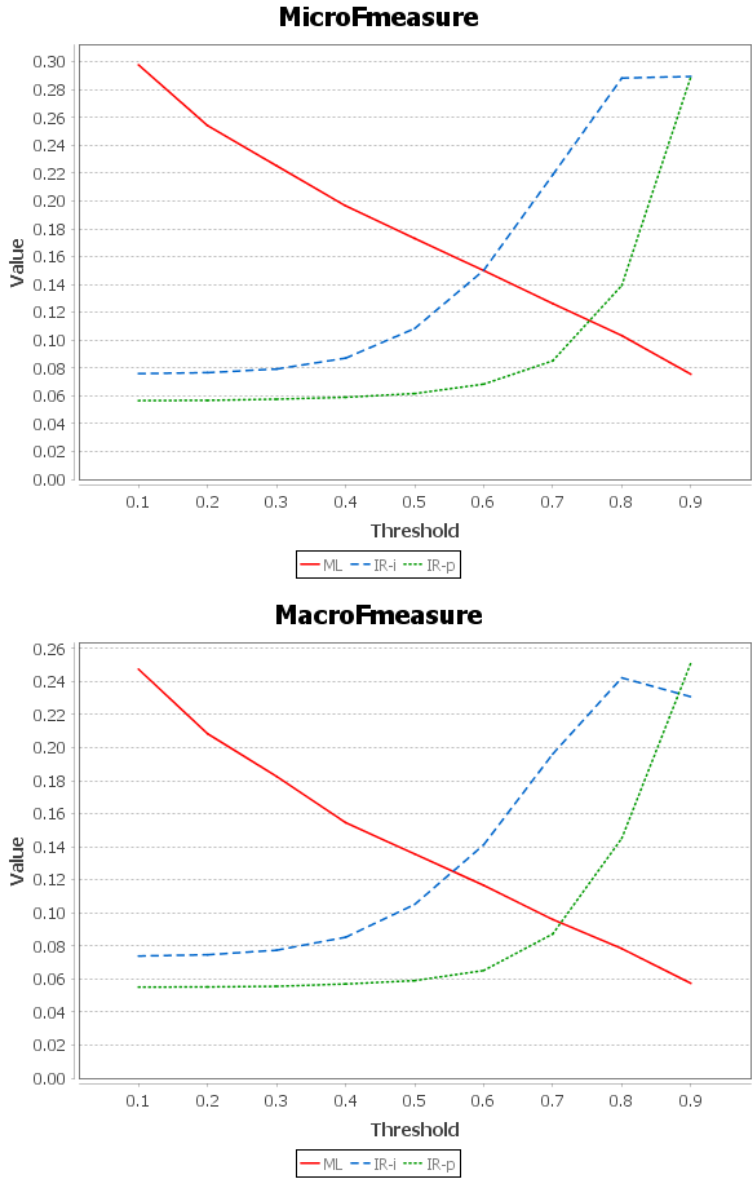

Figura 3.4: Micro y macro *F-measure* para ML, IR-i y IR-p usando diferentes umbrales.

aportación de los resultados de estos experimentos, en primera instancia no se puede asignar un ganador de forma unívoca. En la Tabla 3.2 se pueden apreciar los mejores valores obtenidos para la *F-measure* tanto desde el punto de vista macro-averaged (MF) y micro-averaged (mF) como los respectivos valores de las medidas de calidad del ranking NDCG@10.

En lo referente a los valores obtenidos en las medidas MF y mF, el comportamiento es bastante semejante para todas las aproximaciones que se han propuesto en este trabajo. La aproximación basada en Aprendizaje Automático obtiene una cierta, aunque estrecha ventaja en mF y por otro lado, desde el punto de vista de la MF, la aproximación IR-p obtiene una mejoría aunque no de forma notable. De hecho, un test estadístico basado en la t de student usando

los resultados de las 5 particiones aleatorias y con un nivel de confianza del 99%, no aporta diferencias estadísticamente significativas entre todos estos métodos. En cuanto a la métrica NDCG@10, la aplicación de un t-test indica que la aproximación IR-p es significativamente mejor que las otras dos, las cuáles tampoco representan diferencia significativa alguna entre sí.

3.3.2 Resultados cuando se varía el número de intervenciones

Tal y como se ha referido al comienzo de la Sección 2.3, para este proceso de experimentación solo se han estudiado aquellos MPs que hayan participado en los debates parlamentarios un mínimo de 10 veces. Esto representa un conjunto muy dispar de MPs desde el punto de vista de su actividad en el parlamento, es decir, el conjunto de documentos de entrenamiento está compuesto tanto por las intervenciones de MPs que han participado poco más de una decena de veces, como por aquellos MPs más activos en el desempeño de su labor, los cuáles cuentan en su haber con cientos de intervenciones dentro de una misma legislatura. Esto claramente supone un problema en la medida en la que, a la hora de evaluar, los MPs están siendo considerados de igual forma y, para ratificar como afecta esta problemática, se ha procedido a realizar un análisis del rendimiento del sistema pero esta vez variando el mínimo de intervenciones que un MP debe cumplir para formar parte del estudio, concretamente más de 25, 75 y 150 (véase Figura 3.5).

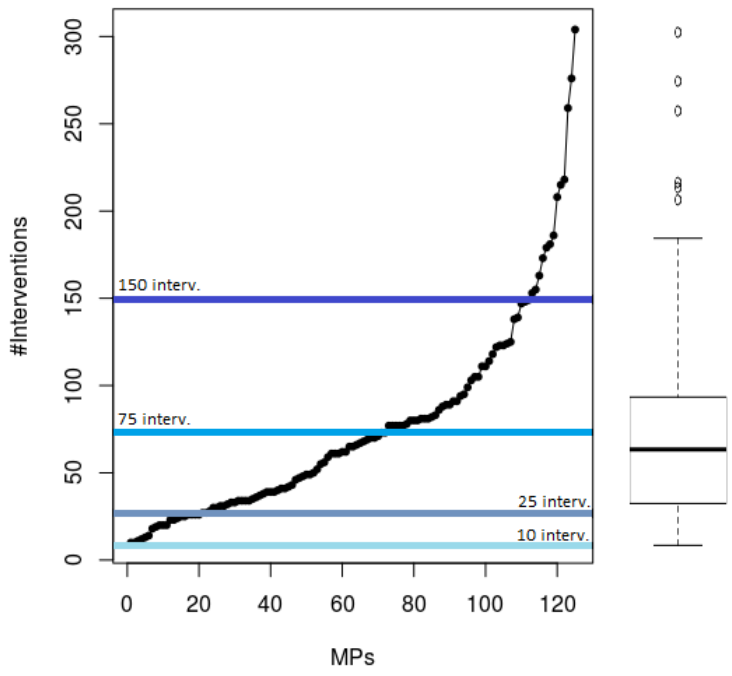

Figura 3.5: MPs con número de intervenciones mayor que 10, 25, 75 y 150.

El objetivo de estudiar el comportamiento del sistema bajo estas nuevas reglas es el de evaluar si, efectivamente construir un conjunto de entrenamiento con aquellos MPs con un mayor número de intervenciones, se traduce por tanto en un mejor rendimiento del sistema. Los resultados de esta experimentación para macro y micro de las aproximaciones ML, IR-i, IR-p se pueden observar en las Figuras 3.6, 3.7 y 3.8, respectivamente. Tal y como se puede observar en la figuras, las tendencias con respecto a los experimentos anteriores se mantienen de forma prácticamente inmutable: en la aproximación basada en Aprendizaje Automático la *F-measure* tanto en micro como en macro decrece en la misma medida que el valor del umbral va aumentando, mientras que lo opuesto ocurre en ambos enfoques de la aproximación basada en Recuperación de Información. Además, cabe destacar que efectivamente los resultados obtienen una destacable mejoría cuando el número mínimo de intervenciones crece. Por lo tanto, esto indica que con un conjunto de entrenamiento más sano y de mejor calidad, ambas aproximaciones podrían alcanzar potencialmente mejores resultados si se dispusiese de un mayor número de intervenciones por MP en el caso de aquellos que tienen una menor actividad parlamentaria.

En la Tabla 3.3, se muestran los mejores valores obtenidos para los diferentes valores mínimos de intervenciones requeridos para las *F-measure* en sus variantes macro y micro y para NCDG@10. En lo que se refiere a las *F-measure*s, un t-test muestra que no existen diferencias estadísticamente significativas entre las aproximaciones ML y IR-i en ningún caso y, sin embargo las aproximaciones ML y IR-i son considerablemente mejores que el enfoque de Recuperación de Información IR-p para la medida mF con número mínimo de intervenciones igual a 75 y 150. Con respecto a la medida de calidad del ranking NDCG@10, de nuevo se produce que las diferencias entre las aproximaciones ML y IR-i no son significativas desde el punto de vista estadístico pero, en este caso, IR-p es significativamente mejor que ML y IR-i sea cual sea el valor mínimo de intervenciones requeridas.

	mF			MF			NDCG@10		
Approach	ML	IR-i	IR-p	ML	IR-i	IR-p	ML	IR-i	IR-p
10	0.2978	0.2896	0.2829	0.2475	0.2423	0.2513	0.6263	0.6246	0.6776
25	0.3037	0.2971	0.2939	0.2658	0.2661	0.2829	0.6267	0.6242	0.6806
75	0.3568	0.3509	0.3085	0.3355	0.3288	0.3368	0.6132	0.6192	0.7086
150	0.4408	0.4282	0.3120	0.4039	0.3948	0.3532	0.5622	0.5744	0.6782

Tabla 3.3: Mejores valores obtenidos por ML, IR-i y IR-p para micro y macro *F-measure* y NDCG@10, usando distintos valores mínimos de iniciativas.

3.4 Conclusiones

En este capítulo se han propuesto, con el objetivo de ser estudiadas y posteriormente comparadas, dos aproximaciones distintas para construir un sistema que sea capaz de filtrar o recomendar de forma automática documentos de diversa índole a los Miembros del Parlamento. La primera de

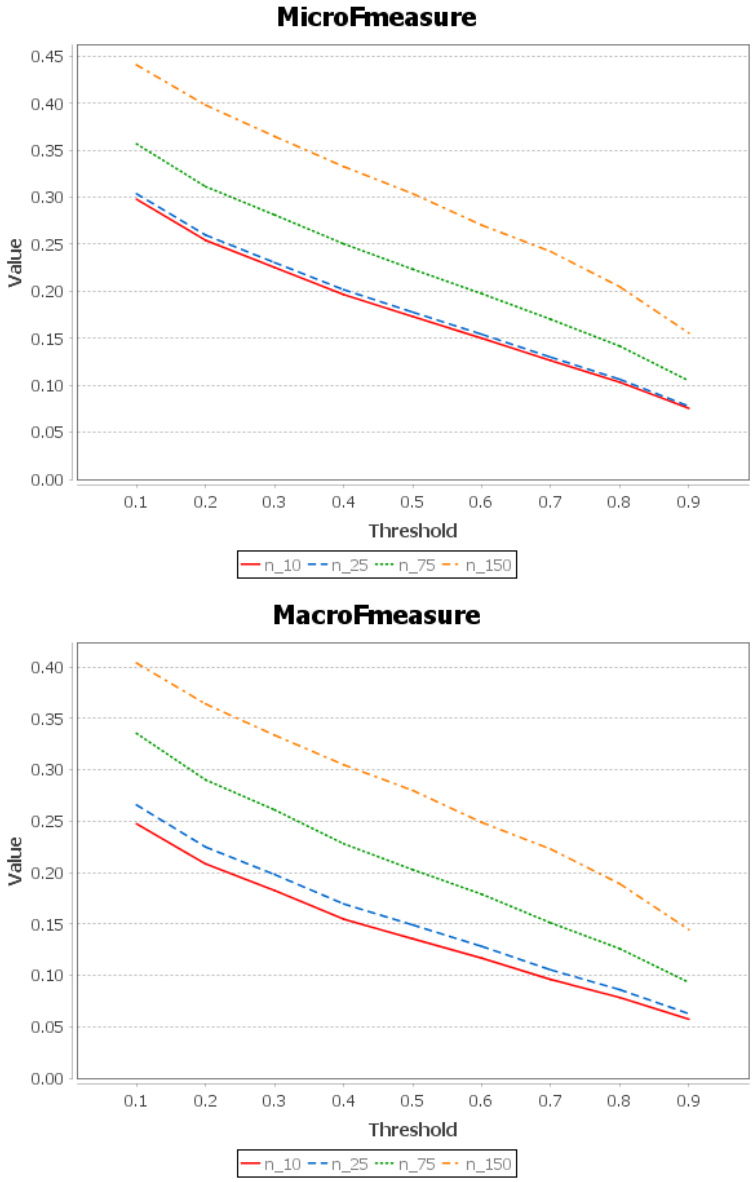

Figura 3.6: Micro y Macro *F-measure* para ML, usando diferentes umbrales y variando el valor mínimo de intervenciones.

estas aproximaciones se ha basado en técnicas de Aprendizaje Automático para la clasificación automática de documentos, mientras la segunda aproximación está basada en los métodos clásicos de Recuperación de Información. Ambas aproximaciones parten de una colección de documentos de entrenamiento, la cual está compuesta de las intervenciones de los MPs en los debates parlamentarios, de donde se puede asumir que hay un volumen de información suficiente como para poder determinar los intereses y preferencias políticas de los Miembros del Parlamento. Mientras que en la aproximación basada en Aprendizaje Automático se usa la colección como conjunto de entrenamiento para construir un clasificador binario para cada MP, en

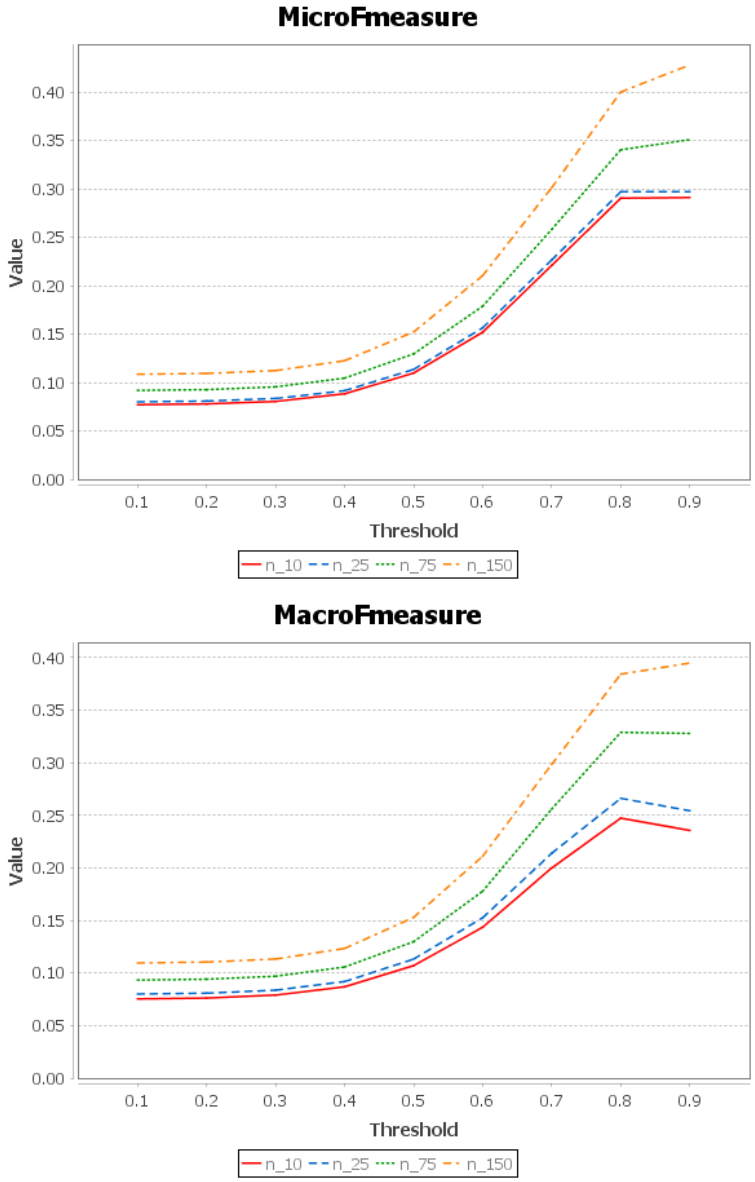

Figura 3.7: Micro y Macro *F-measure* para IR-i, usando diferentes umbrales y variando el valor mínimo de intervenciones.

la aproximación basada en Recuperación de Información esta colección de documentos se utiliza para ser indexada en un Sistema de Recuperación de Información para posteriormente recuperar los Miembros del Parlamento que son más afines a los documentos recibidos como consulta para ser filtrados o recomendados. En ambos casos, la salida que se recibe por parte del sistema está representada en forma de un ranking de MPs ordenados de forma decreciente, la cual representa, o bien un porcentaje de pertenencia a la clase positiva, en el caso de la aproximación basada en Aprendizaje Automático, o bien el grado de similaridad de los MPs con respecto a la consulta realizada. Finalmente, estableciendo un valor para el umbral de relevancia, el sistema filtra o

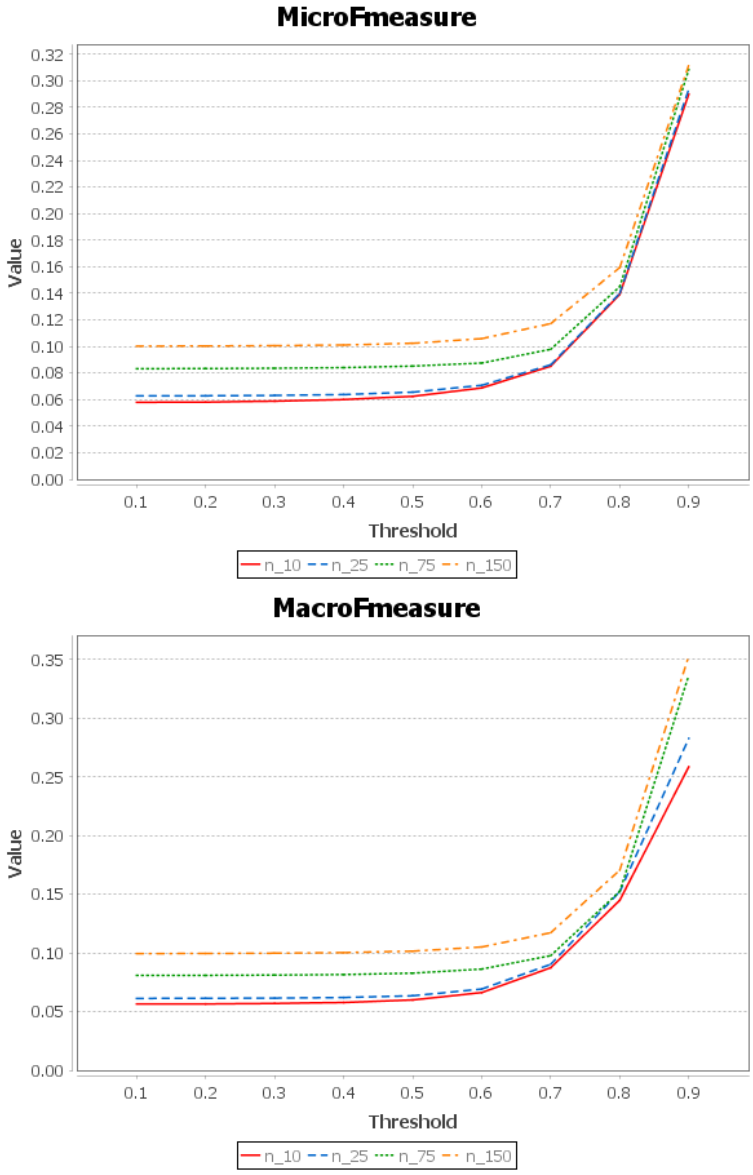

Figura 3.8: Micro y Macro *F-measure* para IR-p, usando diferentes umbrales y variando el valor mínimo de intervenciones.

recomienda el documento utilizado como consulta a aquellos Miembros de Parlamento cuyo valor de score en el ranking esté por encima del umbral fijado.

Las dos aproximaciones que han sido objetivo de este estudio, en lo que al punto de vista de las métricas se refiere, se comportan de forma bastante distinta en términos de *precision* y *recall*, y su rendimiento está fuertemente ligado a diferentes valores a la hora de fijar el umbral de relevancia. Por otro lado, en lo que a la *F-measure* respecta, tanto para macro-averaged como para micro-averaged y en NDCG@10, aunque en menor grado, los mejores resultados obtenidos con ambas aproximaciones son realmente parecidos hasta el punto en el que no se podría establecer

con propiedad cual de las dos aproximaciones es mejor que la otra.

Cabe destacar, en otro orden de ideas, que una de las posibles debilidades que se han producido en la aproximación basada en Aprendizaje Automático ha sido el hecho de considerar como subconjunto de entrenamiento negativo para un MP todas aquellas intervenciones de los MPs que no eran las suyas propias. Esto en cierto modo se podría considerar cuestionable, en el sentido en que las intervenciones del resto de Miembros de Parlamento que traten la misma temática que las propias de un MP deberían ser igualmente consideradas como relevantes para dicho MP. Por ejemplo, un MP que trabaje en una comisión sobre sanidad pública, claramente puede no estar interesado en asuntos como la educación o la agricultura, pero ¿y sobre asuntos económicos? A priori, la respuesta es no, pero cabe la posibilidad de que en esa información sobre economía se recoja una partida de presupuestos para centros hospitalarios la cual no debería ser descartada como relevante. Dicho esto, y con el objetivo de lidiar con el problema derivado de la pobre consideración del conjunto de documentos de entrenamiento de un MP, se propone como medida para encontrar una mejor aproximación al problema desde el enfoque basado en el Aprendizaje Automático, la aplicación de las llamadas técnicas de Positive Unlabeled Learning (PUL) [159], las cuáles basan su funcionamiento en solo asumir la existencia de un subconjunto como instancias positivas en los datos de entrenamiento y otro subconjunto, generalmente mayor, de instancias sin etiquetar, lo cual hace que se disponga de un conjunto de datos de entrenamiento sin instancias negativas con el objetivo de que estas sean determinadas por PUL.

Por otro lado, en este capítulo se plantean dos formas distintas de representar los perfiles de un usuario del sistema, en este caso Miembros del Parlamento, y ambas son los casos más extremos y opuestos de hacerlo. Mientras que en una de ellas se plantea que el perfil esté compuesto por tantos documentos como intervenciones tenga el MP, en la otra aproximación se plantea reunir todas las intervenciones de forma consecutiva en un único documento para representar el perfil de forma monolítica. Ambas aproximaciones tienen sus ventajas e inconvenientes, por tanto para intentar maximizar las virtudes de estas dos formas extremas de representación de perfiles, se propone una nueva representación intermedia basada en la aplicación de técnicas de clustering sobre las intervenciones para generar subperfiles compuestos de intervenciones similares o que traten las mismas temáticas.

3.5 Trabajos relacionados

El dominio sobre el que se plantea este capítulo recae en el contexto de los Sistemas de Filtrado o Recomendación Basados en Contenido [57, 110], los cuáles sugieren elementos del sistema a los usuarios, que hacen uso de él, acordes a sus intereses o preferencias, las cuáles suelen estar representadas por un perfil o un modelo de algún tipo, teniendo en cuenta además las propias características de los elementos que conforman el sistema (el contenido textual en nuestro caso de estudio). Existen un gran número de trabajos y planteamientos enfocados a tratar el

problema del filtrado y recomendación en diversos ámbitos y aplicaciones, como por ejemplo los descritos en los trabajos [15, 91, 92]. No obstante no existe constancia de un sistema como el que se plantea en esta tesis el cual esta destinado a ser utilizado en un contexto político y más específicamente en el contexto parlamentario a excepción de las aproximaciones propuestas por nuestro equipo de investigación en trabajos previos [39–41, 118]. En lo referente a los Sistemas de Recomendación Basados en Contenido se pueden construir tanto con enfoques basados en Recuperación de Información tal y como una de las aproximaciones de este trabajo sugiere, los cuáles generan recomendaciones heurísticas [7, 12, 44, 90], como desde el punto de vista del enfoque del Aprendizaje Automático como también se plantea en este trabajo, los cuáles son abordados de forma general con algoritmos de clasificación supervisada con el objetivo de aprender modelos de usuarios [13, 32, 65, 73, 109, 143].

CAPÍTULO 4

CAPÍTULO

POSITIVE UNLABELED LEARNING PARA LA CONSTRUCCIÓN DE SISTEMAS DE RECOMENDACIÓN EN EL ÁMBITO PARLAMENTARIO

Al igual que en el capítulo anterior, el objetivo de esta Tesis Doctoral es el de encontrar la forma de aprender cuáles son los intereses o preferencias de los Miembros del Parlamento mediante la extracción de información de su actividad parlamentaria, con el objetivo de desarrollar un sistema de filtrado o recomendación que, dado un flujo de documentos, tenga la capacidad de decidir qué documentos deben recibir los Miembros del Parlamento. Pero en este caso se propone usar técnicas de Positive Unlabeled Learning para afrontar la construcción de este sistema, puesto que este sistema solo dispone de información sobre los documentos relevantes (las propias intervenciones de los Miembros del Parlamento), pero no de los documentos irrelevantes y, de este modo, no existe la posibilidad de construir un clasificador binario estándar entrenado con instancias positivas y negativas. Además, se ha desarrollado un nuevo algoritmo de este tipo con el objetivo de ser comparado con otras técnicas ya existentes y usando como casos base las aproximaciones del Capítulo 3.

4.1 Introducción

Una de las formas anteriormente comentadas para abordar el problema de diseñar un sistema de filtrado y recomendación, es la aproximación basada en Aprendizaje Automático. Este sistema debe tener la capacidad de aprender un modelo que defina los intereses y preferencias de los Miembros del Parlamento, de forma automática, a partir de la información contenida en las intervenciones de los debates parlamentarios. Para extraer dicha información de las intervenciones, se utilizan las transcripciones literales y a continuación se utilizan para entrenar un clasificador binario clásico. Este clasificador recibe como entrada, para cada Miembro del Parlamento, por un lado las intervenciones propias del MP, que son consideradas como instancias positivas y, por otro lado, el resto de intervenciones del resto de los MPs, que son consideradas como instancias de aprendizaje negativo. De esta forma se obtiene así un conjunto de clasificadores binarios, uno por cada Miembro del Parlamento, los cuales tienen el propósito de, llegada una nueva consulta al sistema, determinar qué Miembros del Parlamento podrían estar interesados en la consulta. Es decir, la consulta se lanza contra todos los clasificadores binarios y la consulta puede ser considerada como relevante para un MP, si el clasificador binario asociado a este devuelve la clase positiva. Normalmente, un clasificador binario devuelve el valor 1 si la instancia de test se clasifica como positiva y 0 si la instancia de test se clasifica como negativa pero, en este caso de estudio particular, los clasificadores van a devolver un valor comprendido entre el intervalo [0,1]. Este valor no responde a la pregunta de si la consulta es relevante o no sino en qué grado la consulta le es relevante al MP, es decir, el porcentaje de pertenencia a la clase positiva. De este modo, cada clasificador devuelve un valor numérico con el que se puede construir una lista ordenada en sentido decreciente de Miembros del Parlamento a modo de ranking, pudiendo así filtrar o recomendar el documento usado como consulta a aquellos MPs que queden a la cabeza del ranking.

El problema derivado de esta aproximación se basa en la actuación ingenua de considerar como información no relevante todas las intervenciones de los Miembros del Parlamento ajenas a las propias de un MP en cuestión. De esta forma, a la hora de construir un clasificador binario, se dispone de un conjunto de datos de entrenamiento relativamente poco robusto, donde las instancias positivas son aquellas intervenciones que ha realizado el propio MP y, por otro lado, las instancias negativas o irrelevantes son todas las que han sido de la autoría del resto de Miembros del Parlamento, independientemente de que hayan tratado temáticas similares a las que al MP en cuestión puedan resultar interesantes. Dicho esto, es obvio que las intervenciones en los debates parlamentarios de un MP en cuestión son inequívocamente relevantes para él y, por lo tanto, deben ser correctamente consideradas como instancias positivas a la hora de entrenar un clasificador binario, sin embargo, la afirmación que se hacía en la aproximación basada en Aprendizaje Automático en el Capítulo 3 sobre el subconjunto de entrenamiento negativo deja mucho que desear en lo referente a que, en el caso más que posible de producirse, considerar como irrelevantes documentos que tratan la misma temática que los documentos considerados

como relevantes provoca una confusión en el clasificador, lo cual derivará en un futuro en el pobre funcionamiento del sistema. Por ejemplo, tomando como referencia un Miembro del Parlamento cuya especialidad política está directamente relacionada con el entorno de la educación, es más que probable que pueda encontrar cierto tipo de relevancia en la mayoría de intervenciones del resto de MPs, las cuales con gran seguridad puedan ir incluso dirigidas a la comisión a la que el MP pertenece o directamente hacia su persona, sobre la misma temática. Por lo tanto, sería lógico encontrar una forma efectiva de discernir automáticamente qué instancias del subconjunto de datos, considerado anteriormente como negativo en su totalidad, realmente no lo son.

Esta situación puede ser paliada usando una serie de técnicas denominadas como Positive Unlabeled Learning (PUL) [159], las cuales basan su funcionamiento en la suposición de que existe un subconjunto de datos positivos completamente reconocido y otro subconjunto, el cual es de forma general considerablemente más grande, de instancias que no están etiquetadas por falta de información, lo cual lleva a que no exista un determinado subconjunto de instancias negativas. En este caso de estudio, el subconjunto de datos positivos son, tal y como se ha referido con anterioridad, las propias intervenciones de un MP y, por otro lado, el subconjunto de datos no etiquetado se corresponde con las demás intervenciones del resto de los Miembros del Parlamento. Las técnicas PUL se definen como un caso extremo de aprendizaje semisupervisado [26] las cuales consideran de forma simultánea la existencia de los subconjuntos de datos de entrenamiento positivos, negativos y no etiquetados.

Dado que la naturaleza del problema que se trata en este estudio se ajusta a la perfección al dominio de las técnicas de PUL, se propone la aplicación de dichas técnicas para la construcción de un sistema de recomendación basado en contenido para filtrar o recomendar documentos a los Miembros del Parlamento con el objetivo de mejorar el rendimiento de aproximaciones previas. De forma más específica, la aproximación que se plantea en este capítulo está basada en primer lugar en la determinación de entre el subconjunto de datos de entrenamiento no etiquetados, aquellas instancias que pueden conformar un subconjunto de instancias negativas y, en segundo lugar, utilizar este subconjunto de instancias negativas junto con el que se dispone de instancias positivas para entrenar un clasificador binario para cada Miembro del Parlamento. Para desarrollar la tarea de determinar un conjunto consistente de instancias negativas se han hecho uso de una técnica ya conocida de PUL que se encuentran en la literatura y, por otro lado, también se ha diseñado e implementado un nuevo método de PUL el cual basa su funcionamiento en una adaptación del conocido algoritmo de clustering K-means para este tipo de propósito.

Con el objetivo de validar de forma experimental las propuestas que se plantean en este trabajo, se ha realizado la metodología de estudio, descrita en el Capítulo 3, utilizando también la colección de transcripciones literales de las intervenciones de los MPs en los debates parlamentarios de la octava legislatura del Parlamento de Andalucía.

4.2 Positive Unlabeled Learning en el ámbito parlamentario

La situación ante la que nos encontramos en este caso de estudio está definida como sigue: sea $\mathscr{MP} = \{MP_1, \ldots, MP_n\}$ el conjunto que comprende a todos los Miembros del Parlamento de los que se tiene en disposición alguna intervención que haya sido recogida en la colección de documentos de entrenamiento. Se parte de la suposición de que la institución del parlamento recibe o bien genera una serie de documentos textuales que necesitan ser distribuidos de forma adecuada entre los diferentes Miembros del Parlamento. Sin embargo, teniendo como propósito el facilitar la labor política de los Miembros del Parlamento, todos los MPs no deberían recibir todo el flujo de documentos constante que llega al parlamento [138], lo cual supondría qué como individuos no pudieran afrontar ni gestionar la ingente cantidad de información de la que pudieran disponer. En lugar de eso, cada Miembro del Parlamento debería recibir solo aquellos documentos que hagan referencia de forma directa a los intereses o preferencias políticas del MP en cuestión y que sean de utilidad a la labor que desempeña desde el punto de vista político dentro del parlamento. Por lo tanto, se debe diseñar para tal propósito un sistema que sea capaz de realizar este proceso de filtrado previo de forma automática y para ello, tal como se ha descrito en la Sección 3.2.1, se plantea el diseño y construcción de un sistema con un enfoque basado en el Aprendizaje Automático, más especificamente basado en las técnicas de Positive Unlabeled Learning. La fuente de donde se extrae la información, la cual es pública y presumiblemente confiable, sobre la actividad política de los diferentes Miembros del Parlamento va a ser la contenida en las transcripciones literales de las intervenciones de los MPs en los debates parlamentarios. Por lo tanto, asociado a cada MP_i hay un conjunto de documentos $\mathscr{D}_i = \{d_{i1}, \ldots, d_{im_i}\}$, donde cada d_{ij} representa un documento donde se recoge cada intervención del MP en cuestión MP_i. El conjunto completo de documentos está definido por $\mathscr{D} = \cup_{j=1}^{n} \mathscr{D}_j$ y por lo tanto se va a entrenar un conjunto de n clasificadores de texto binarios a partir de $\mathscr{D}$, uno por cada Miembro del Parlamento, donde el conjunto de instancias positivas de entrenamiento es $\mathscr{D}_i$ y por otro lado, el resto de instancias consideradas como no etiquetadas está definido por $\mathscr{D} \setminus \mathscr{D}_i$.

Dicho esto, la propuesta sobre la utilización de las técnicas PUL con el objetivo de construir un Sistemas de Recomendación y Filtrado de documentos para los Miembros del Parlamento está planteada desde el punto de vista de una estrategia en dos pasos. En primer lugar, la aplicación de la pertinente técnica para la detección de un conjunto consistente de instancias negativas, $\mathscr{N}_i$, a partir del subconjunto de instancias sin etiquetar de cada MP_i del que se dispone y está conformado por las intervenciones del resto de Miembros del Parlamento MP_j. Y, como segundo paso, para cada MP_i, la consecuente construcción de un clasificador binario estándar tomando como instancias positivas las contenidas en el conjunto $\mathscr{D}_i$ y, como instancias de entrenamiento negativas, las que pertenecen al conjunto extraído denominado como $\mathscr{N}_i$, utilizando como clasificador el algoritmo Support Vector Machine [35], el cual está considerado en la literatura como la técnica más utilizada en el ámbito de la clasificación de documentos. En otro orden de cosas, cabe destacar la apreciación de que es bastante probable que el subconjunto

de entrenamiento negativo $\mathcal{N}_i$ tenga un mayor número de instancias que el correspondiente subconjunto de instancias positivas $\mathcal{D}_i$ lo cual tiene como consecuencia que el conjunto de entrenamiento no esté adecuadamente balanceado y a partir de esta consideración se plantea la posibilidad de la aplicación de métodos para lidiar con el problema de clases no balanceadas.

4.2.1 Positive Unlabeled Learning basado en Naive Bayes

Con el objetivo de comparar la aproximación de PUL basada en K-means que se propone en este capítulo y además poder afianzar que el uso de técnicas de PUL, de forma general, sirve para mejorar el rendimiento de los Sistemas de Recomendación y Filtrado, se ha tenido en consideración una aproximación ya existente en el estado del arte.

Esta aproximación se basa en entrenar un clasificador Naive Bayes (Figura 4.1) tomando como instancias de entrenamiento positivas las propias intervenciones de un MP y como instancias negativas el resto de intervenciones del resto de MPs. A continuación, una vez se ha entrenado el clasificador, se vuelven a reclasificar las mismas instancias que previamente se han considerado como entrenamiento negativo para el clasificador. Entonces, las instancias negativas que se vuelven a clasificar como negativas pasan a formar parte del conjunto de instancias verdaderamente negativas del MP y, por otro lado, las instancias negativas que se vuelven a clasificar como positivas se descartan. De esta forma, se extraen del conjunto de entrenamiento del MP, aquellas intervenciones que tratan temas similares a sus propias intervenciones, pero que por no ser suyas, anteriormente eran consideradas como negativas, evitando de esta forma que el clasificador asociado al MP pueda asignar un valor negativo a intervenciones relevantes.

El problema de esta aproximación de PUL es que el conjunto de entrenamiento negativo del clasificador Naive Bayes, es considerablemente mayor que el conjunto instancias positivas. Por este motivo, a la hora de volver a clasificar las instancias negativas, el sobreajuste provoca que la gran mayoría de instancias se vuelvan a clasificar como negativas, dejando de esta forma el conjunto de entrenamiento para el MP prácticamente intacto.

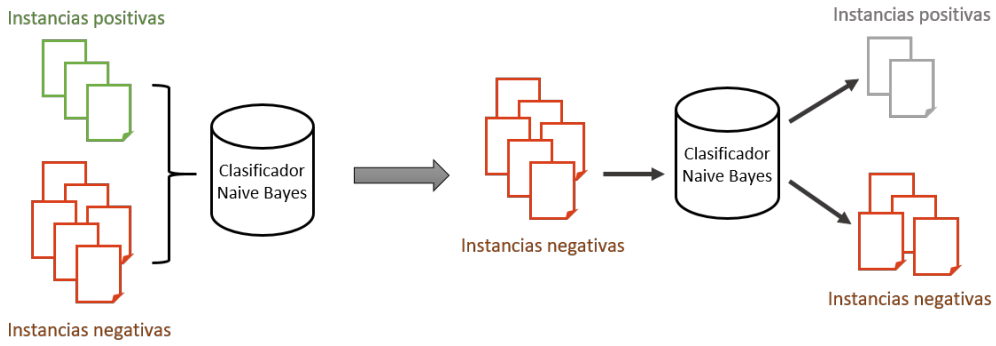

Figura 4.1: Descripción de Positive Unlabeled Learning con Naive Bayes.

4.2.2 Positive Unlabeled Learning basado en K-means

El algoritmo de agrupamiento clásico K-means es un método no supervisado e iterativo, el cual basa su funcionamiento en establecer de forma aleatoria un centroide por cada uno de los K clusters y de esta forma se asigna cada una de las instancias al correspondiente cluster cuyo centroide esté más cercano a la instancia en cuestión. A continuación el algoritmo vuelve a computar los centroides de cada cluster en base al punto medio de las instancias que han sido asignadas a cada uno de ellos para, de nuevo, volver a asignar las instancias más cercanas a los nuevos centroides de sus respectivos clusters. Este proceso se repite, de manera general, hasta que el criterio de convergencia se cumple, o bien no existen variaciones en la asignación de instancias en repeticiones consecutivas. En este caso particular, donde se modifica el funcionamiento canónico del algoritmo K-means con el propósito de ser utilizado como técnica de PUL, el valor que define el número de clusters se establece en K=2 y la función de similitud entre dos documentos se establece a través de la clásica medida coseno [7]. La modificación que se propone para el algoritmo K-means consiste en que sabiendo cuales son las instancias positivas del conjunto de entrenamiento, estas quedan obligadas a permanecer en el cluster (positivo) que se les ha asignado al inicio de la ejecución sin posibilidad de abandonarlo en las sucesivas iteraciones, por otro lado, las instancias no etiquetadas, las cuales han sido asignadas inicialmente al otro cluster (negativo) pueden moverse libremente entre ambos cluster en función de la similitud al centroide más cercano. Para inicializar el proceso, al contrario que en la versión clásica, el centroide del cluster positivo se calcula a partir del punto medio de todas las instancias positivas y el centroide del cluster negativo se establece en el punto medio obtenido de todas las instancias no etiquetadas. Al final del proceso de ejecución del algoritmo, las instancias no etiquetadas que han sido asignadas al cluster positivo son descartadas puesto que representan documentos que son similares a los considerados como relevantes pero no son de la autoría del MP en cuestión y, por otro lado, las instancias no etiquetadas que han permanecido en el cluster negativo pasan a ser consideradas como un subconjunto consistente de instancias negativas a la hora de entrenar el clasificador binario (véase Figura 4.2).

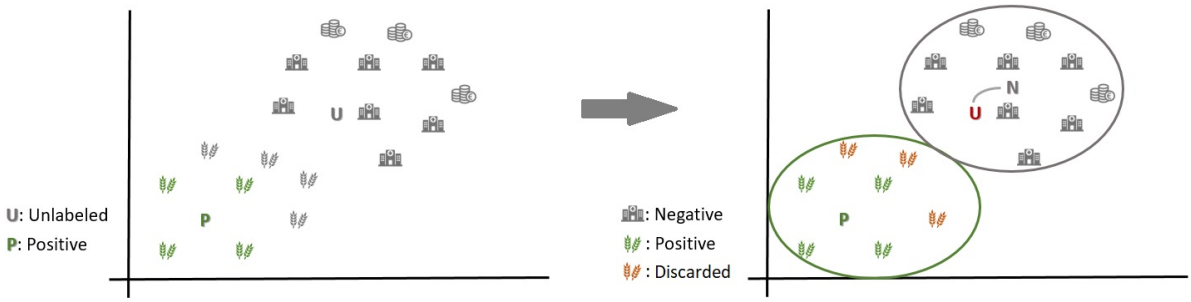

Figura 4.2: Descripción de Positive Unlabeled Learning con K-means.

4.3 Evaluación experimental con Positive Unlabeled Learning

Para evaluar de forma experimental las propuestas que se plantean en este capítulo se va a utilizar el mismo proceso de experimentación propuesto en el Capítulo 3, mas concretamente en la Sección 3.3.

Una vez definidos los criterios que se van a utilizar para evaluar el rendimiento del sistema se procede a la definición de las distintas aproximaciones que se plantean para ser comparadas entre sí. La primera que se ha considerado, la cual se establece como caso base (*b*as) para un primer acercamiento al problema consiste en el entrenamiento de los clasificadores binarios SVM sin la previa aplicación de las técnicas de PUL, de esta forma, tal y como se describe en el Capítulo 3, las instancias positivas son $\mathscr{D}_i$ y el conjunto completo de instancias no etiquetadas se considera como el subconjunto de instancias negativas $\mathscr{D} \setminus \mathscr{D}_i$. Como segunda propuesta, con el objetivo de ser comparada con la que se propone en este trabajo, se plantea la aplicación de una de las técnicas de PUL más utilizadas y referidas en la literatura la cual basa su funcionamiento en clasificadores bayesianos (*p*ul-nb) donde una vez las instancias negativas son detectadas por este método, se procede al entrenamiento de los clasificadores SVM. Finalmente, como tercera aproximación al problema se incluye en el proceso de experimentación el método basado en la modificación del algoritmo K-means (*p*ul-km), propuesto en la Sección 4.2.2, como primer paso de la técnica PUL. De este modo, la comparación entre las propuestas *b*as y *p*ul-km se llevará a cabo con el objetivo de visualizar la aportación positiva de las técnicas de PUL en el contexto del filtrado y recomendación de documentos. Por otro lado, llevar a cabo una comparativa entre las propuestas *p*ul-nb y *p*ul-km servirá para dar una idea del potencial de la nueva técnica de PUL que se plantea en este capítulo.

Tal y como se ha mencionado en la Sección 4.2, para lidiar con el problema de los conjuntos de entrenamiento mal balanceados que se producen en gran medida en este caso de estudio, se van a incluir como aproximaciones alternativas unas versiones de *b*as, *p*ul-nb y *p*ul-km, las cuales recibirán la nomenclatura de *b*as-b, *p*ul-nb-b y *p*ul-km-b respectivamente donde, como paso previo al entrenamiento de los clasificadores binarios SVM, se aplica un método para tratar de evitar el problema de los conjuntos de entrenamiento no balanceados. De forma específica, se ha aplicado la técnica de balanceado Synthetic Minority Over-sampling Technique (SMOTE) [27], la cual se define esencialmente como un algoritmo estadístico para la creación de nuevas instancias a partir de los casos existentes de la clase minoritaria. El algoritmo que compone SMOTE funciona tomando instancias de la clase con menos observaciones, la clase positiva en este caso de estudio, y un conjunto de sus k vecinos más cercanos, de este modo, se genera una nueva instancia sintética estableciendo un punto de forma aleatoria en los segmentos definidos por el punto que define la instancia en cuestión y los puntos de sus k vecinos más cercanos. Para llevar a cabo la implementación de estas aproximaciones al problema se han usado las versiones de SVM, NB y SMOTE disponibles en R (paquetes *c*aret, *e*1071 y *D*MwR). Los pasos correspondientes para el preprocesamiento de los datos (se han eliminado las palabras vacías y

se ha realizado un stemming de los términos de todas las iniciativas) se ha llevado también a cabo con el uso de paquete de R (*t*m y *s*nowBallC). La implementación de la versión modificada del algoritmo K-means se ha desarrollado en el lenguaje de programación Java.

Con el objetivo de obtener una lista ordenada de Miembros del Parlamento a modo de ranking, la versión que se ha utilizado del algoritmo de clasificación SVM permite devolver un valor numérico comprendido dentro del intervalo [0,1], la cual representa la probabilidad de que un documento, que se lanza como consulta al sistema con el objetivo de ser filtrado o recomendado d, sea relevante para el Miembro del Parlamento MP_i, $pr_i(d)$. De este modo se puede usar dicha probabilidad con el propósito de considerar un documento d como relevante si $pr_i(d) \geq 1 - pr_i(d)$, es decir $pr_i(d) \geq 0.5$. Aunque de forma genérica, se puede establecer un valor de umbral t ($0 \leq t \leq 1$) y considerar que d es relevante para MP_i si, y solo si, $pr_i(d) \geq t$. En este sentido, los valores de TP_i, FP_i y FN_i que se utilizan para realizar los cálculos de *precision* y *recall* se obtienen de acuerdo a la contingencia propuesta en la Tabla 4.1. El proceso de experimentación se ha replicado para múltiples valores del umbral t, los cuales quedan comprendidos entre los valores desde 0.1 hasta 0.9.

	Verdadero relevante para MP_i	Verdadero irrelevante para MP_i
$pr_i(d) \geq t$	TP_i	FP_i
$pr_i(d) < t$	FN_i	

Tabla 4.1: Tabla de contingencia para el umbral t.

4.3.1 Resultados con conjuntos de datos no balanceados

Los resultados obtenidos a partir del proceso experimental llevado a cabo desde el punto de vista global micro y macro para *precision*, *recall* y *F-measure* estableciendo diferentes valores de umbral de relevancia se muestran en las Figuras 4.3, 4.4, 4.5, respectivamente.

En primer lugar, las gráficas de resultados que se pueden apreciar en las Figuras 4.3 y 4.4 permiten establecer ciertas tendencias generales para las tres aproximaciones: a medida que el valor de umbral de relevancia se establece en valores más altos, se puede observar en la *precision* una predisposición a aumentar de forma proporcional, sin embargo, en la medida de *recall* se obtienen peores resultados. No obstante, este es el comportamiento esperado en el sentido en que cuando el valor del umbral de relevancia aumenta los clasificadores actúan de forma más selectiva a la hora de determinar la relevancia de un documento, lo cual tiene como consecuencia que el número de falsos positivos sea menor y por lo tanto aumente la *precision* y, en contraposición aumente el número de falsos negativos lo cual conlleva a la obtención de valores de la medida de *recall* más bajos. Sin embargo, existe una excepción a este comportamiento común entre las aproximaciones y este se muestra en la medida de macro *precision* en la propuesta basal *b*as: esta medida alcanza un máximo en un valor de umbral bajo para luego

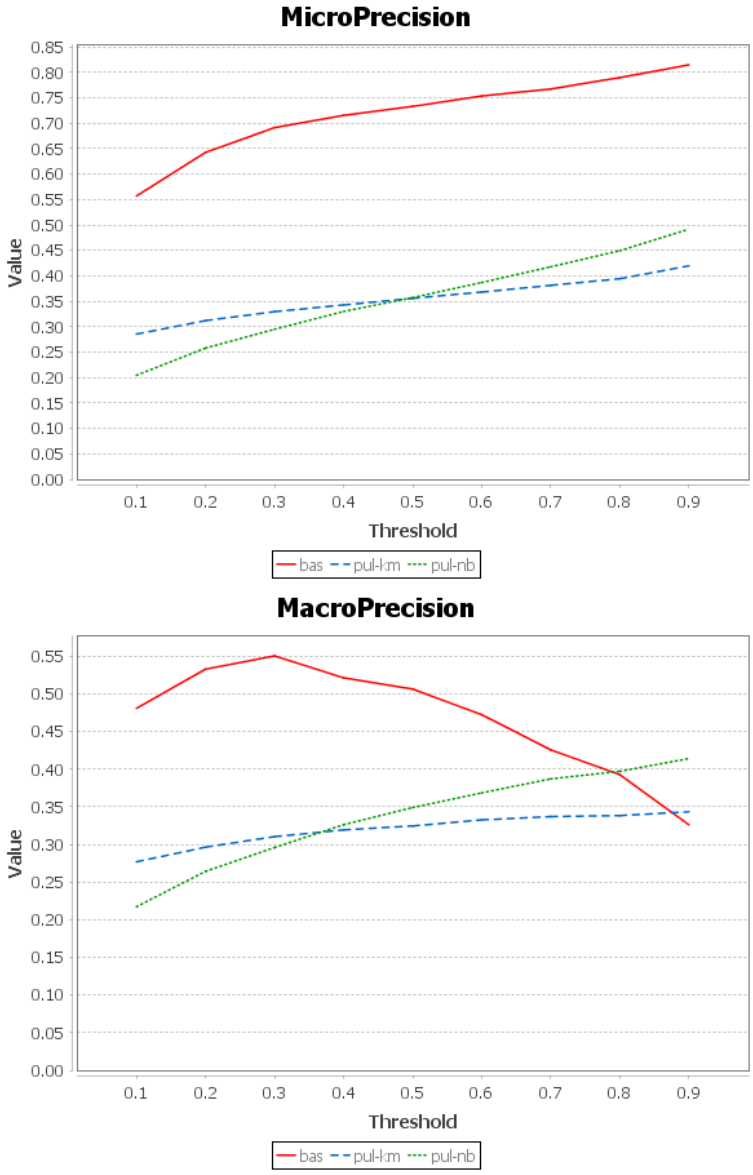

Figura 4.3: Micro y Macro *precision* para *b*as, *p*ul-km y *p*ul-nb utilizando diferentes umbrales.

decrecer bruscamente a medida que el valor de umbral de relevancia aumenta. La explicación lógica para este comportamiento podría ser la pobre actuación de los clasificadores binarios en esta aproximación ante conjuntos de datos de entrenamiento con un número muy bajo de instancias positivas, es decir, los clasificadores asociados a los Miembros de Parlamento que han intervenido escasas veces en los debates, lo cual tiene como consecuencia directa que el conjunto de entrenamiento esté extremadamente desbalanceado. En estos casos, lo que se produce es que el número de verdaderos positivos disminuye a medida que el valor de umbral de relevancia aumenta de forma más brusca con respecto a como aumenta el número de falsos negativos. Cabe destacar que este comportamiento solo se produce en la macro *precision* y no en la micro *precision*,

ya que en el primero caso todos los Miembros del Parlamento contribuyen de igual forma al valor de esta medida independientemente del número de intervenciones que este tenga y esto no ocurre en la medida global de micro *precision*, la cual evalúa la importancia de la contribución de cada Miembro de Parlamento a la medida en función del número de sus intervenciones.

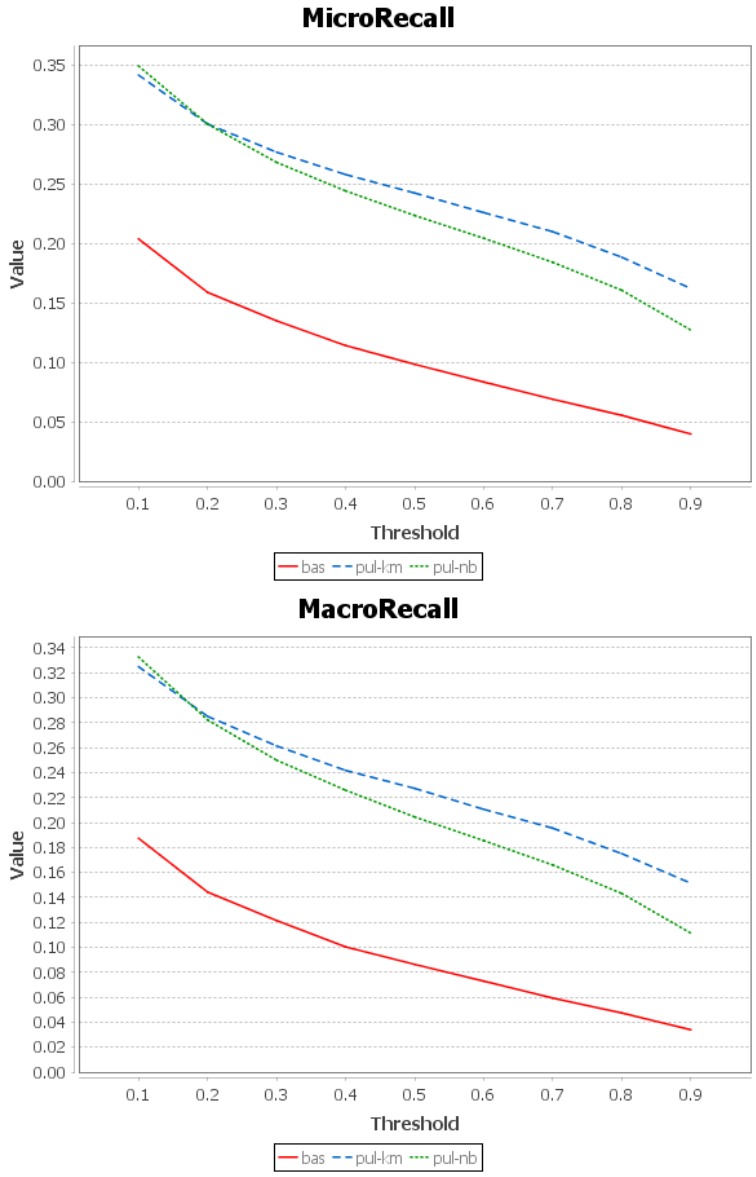

Figura 4.4: Micro y macro *recall* para *b*as, *p*ul-km y *p*ul-nb usando distintos valores de umbral.

En estas figuras también se muestra que la aproximación base *b*as tiene una tendencia distinta en referencia a las dos aproximaciones donde se aplican las técnicas de PUL, *p*ul-nb y *p*ul-km: de forma general, la aproximación *b*as se comporta de mejor manera desde el punto de vista de la medida de *precision* y, en contrapartida las aproximaciones *p*ul-nb y *p*ul-km actúan mejor con respecto a la medida de recall. Dadas las características del proceso de evaluación

que se ha llevado a cabo en esta experimentación, cabe destacar que la medida de *recall* debe ser tomada en mayor consideración que la medida de *precision* y la razón de esto viene definida porque el número de falsos negativos, el cual afecta de forma directa al *recall*, representa un error real, es decir, un Miembro del Parlamento que ha participado en la iniciativa que se ha utilizado como consulta y, sin embargo, el sistema no ha sido capaz de recomendársela a dicho MP. Por otro lado, el número de falsos positivos, los cuales afectan de forma directa a la *precision*, definen a un Miembro del Parlamento que no ha participado de forma explícita en la iniciativa pero el sistema la ha considerado como relevante para el MP y esto ocurre cuando un MP puede considerar como relevante la iniciativa puesto que se tratan temas en ella que son afines a sus intereses políticos. De esta forma, un *recall* bajo representa una señal objetiva de un bajo rendimiento del sistema, mientras que un mal valor de *precision* no debe ser apreciado necesariamente de la misma forma puesto que esto se debe a que los criterios para determinar la relevancia que se han establecido para un documento son demasiado estrictos.

Aproximación	bas	pul-km	pul-nb
		Micro-F	
Valor	0.2978	**0**.3105	0.2802
Umbral	0.1	0.1	0.3
		Macro-F	
Valor	0.2475	**0**.2644	0.2454
Umbral	0.1	0.1	0.2

Tabla 4.2: Mejor micro y macro *F-measure* obtenida por *b*as, *p*ul-km y *p*ul-nb.

De forma específica, en la Figura 4.5 se pueden observar los resultados para las medidas de micro y macro *F-measure*, las cuales representan el balance entre las medidas de *precision* y *recall* y, por lo tanto son una medida apropiada para evaluar el rendimiento del sistema desde un punto de vista global. En primera instancia, se puede observar que los mejores resultados para estas medidas se obtienen de forma sistemática cuando el valor del umbral de relevancia se establece en valores bajos. Además, se puede apreciar de forma evidente que la aproximación que se plantea como novedosa en este trabajo obtiene un mejor rendimiento con respecto al resto de aproximaciones.

Por otro lado, la Tabla 4.2 indica los mejores resultados obtenidos en referencia a la medida *F-measure* para cada una de las distintas aproximaciones así como los correspondientes valores del umbral de relevancia donde se han alcanzado esos resultados. Se han llevado a cabo, comparando entre sí todas las aproximaciones que se han planteado, una serie de test estadísticos t-test usando los resultados de las cinco particiones aleatorias y un nivel de confianza del 95% con el objetivo de encontrar ciertas significancias estadísticas entre los resultados. Dicho esto, el test estadístico muestra que *p*ul-km es siempre mejor que el resto de aproximaciones *p*ul-nb y *b*as de forma significativa. Desde el punto de vista de las medidas globales micro, el caso base que se plantea en este trabajo *b*as es significativamente mejor que *p*ul-nb mientras que por otro lado no

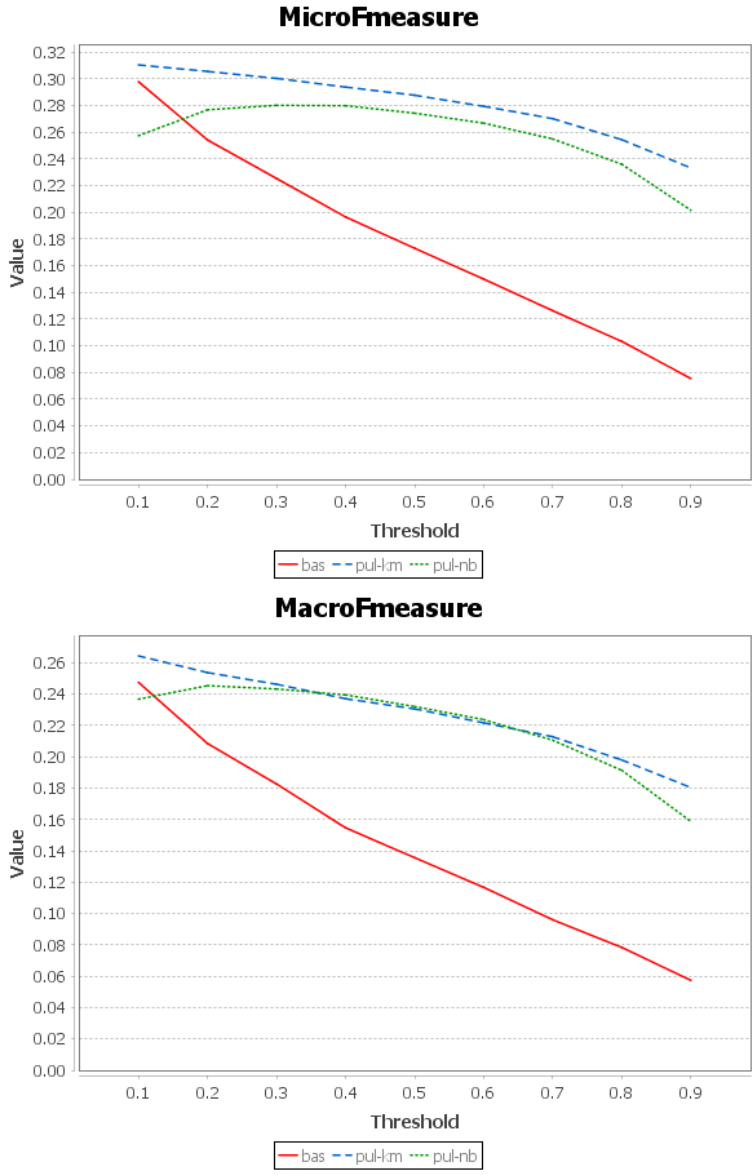

Figura 4.5: Micro y macro *F-measure* para *b*as, *p*ul-km y *p*ul-nb usando distintos valores de umbral.

se observan diferencias estadísticas significativas entre estas dos aproximaciones al nivel global macro.

4.3.2 Resultados con conjuntos de datos balanceados

En esta sección se pretende analizar los resultados de las versiones de las aproximaciones previas pero con la particularidad de que el conjunto de datos va a ser balanceado como paso previo al entrenamiento de los clasificadores binarios. Desde el punto de vista de la nomenclatura, estas aproximaciones se definen como *b*as-b, *p*ul-km-b y *p*ul-nb-b. Los resultados relativos a las medidas

de carácter global micro y macro para *F-measure* se muestran en la Figura 4.8. Las Figuras 4.6 y 4.7 que hacen referencia a la *precision* y el *recall* respectivamente, muestran un comportamiento similar al descrito en la sección anterior, tendencia a crecer en lo que respecta a la *precision* y por el contrario, tendencia a decrecer en la medida de *recall* en relación al incremento del valor de umbral de relevancia, no obstante, cabe destacar que en este caso las líneas que definen los valores de las medidas a través de los distintos valores del umbral de relevancia son más cercanas. Además, otro punto interesante a tener en cuenta es que el comportamiento anómalo de la macro *precision* en la aproximación *b*as ya no se produce, dando así validez a la hipótesis donde se planteaban las causas por las que este fenómeno ocurría.

Tomando como referencia la Figura 4.8 se pueden apreciar diversos sucesos que son interesantes para ser tenidos en cuenta. En primer lugar, los valores del umbral de relevancia donde se alcanzaban los mejores resultados en las respectivas aproximaciones previas han variado, en este caso en particular, dichos valores se establecen en torno al valor medio 0.5, el cual se considera como el valor natural por encima del cual se le podría asignar el título de relevante a un documento determinado. Esto indica, efectivamente, que en este punto los clasificadores están mejor calibrados y por lo tanto no se necesitan relajar los criterios de relevancia para obtener unos resultados que sean competentes. Otro hecho destacable es que una vez más, la versión con el conjunto de entrenamiento balanceado de la aproximación donde se utiliza el algoritmo K-means modificado como técnica de PUL, *p*ul-km-b, sigue siendo el mejor enfoque para la construcción de este sistema, aunque las diferencias con respecto al caso base con el conjunto de entrenamiento balanceado *b*as-b son relativamente menores que las que existían en las aproximaciones homólogas sin aplicar técnicas de balanceo del conjunto de datos de entrenamiento. Por último, remarcar que aplicar técnicas de balanceo, SMOTE en este caso, al conjunto de datos de entrenamiento producido por la técnica de PUL basada en redes bayesianas, *p*ul-nb, no es una buena idea en el sentido en que se obtienen resultados significativamente peores con respecto a *b*as-b y *p*ul-km-b. La Tabla 4.3 representa de forma homóloga los resultados de la Tabla 4.2 pero para el caso de los experimentos con el conjunto de datos de entrenamiento balanceado. Usando las Figuras 4.8 y 4.5 y las Tablas 4.3 y 4.2 para realizar una comparativa, se puede observar de forma clara que aplicando SMOTE para balancear los conjuntos de entrenamiento de los clasificadores se obtienen mejores resultados desde el punto de vista de la medida macro *F-measure*, a excepción de la aproximación *p*ul-nb-b, pero como contrapunto, los mejores valores alcanzados con la medida micro *F-measure* se ven afectados negativamente de forma sistemática. La razón de este comportamiento puede venir derivada ciertamente a causa de que balancear los clasificadores de los Miembros del Parlamento con menos intervenciones afecta positivamente en relación a los resultados, sin embargo esto puede hacer que los clasificadores de aquellos Miembros de Parlamento con un gran número de intervenciones, los cuales suponían un mayor peso en la medida micro *F-measure*, se puede ver afectados de forma negativa. En este caso, los test estadísticos t-test avalan que no existen diferencias estadísticas significativas

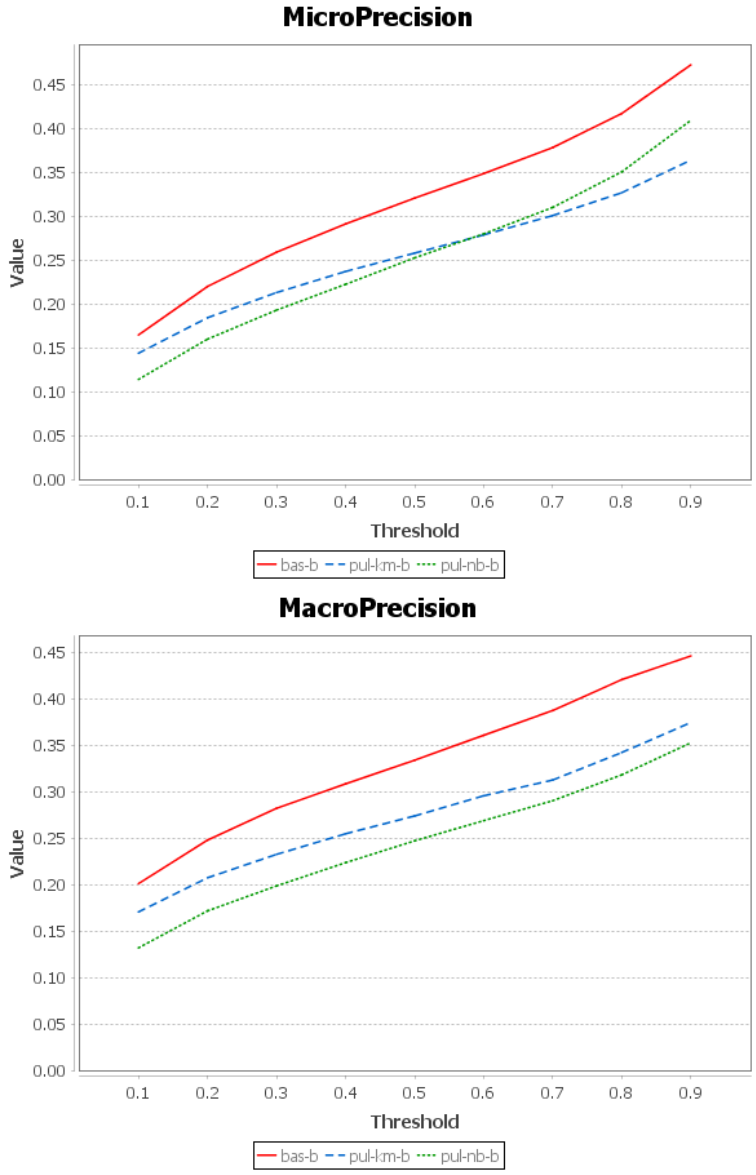

Figura 4.6: Micro y macro *precision* para *b*as-b, *p*ul-km-b y *p*ul-nb-b usando distintos valores de umbral.

entre las aproximaciones *p*ul-km-b y *b*as-b aunque ambas son significativamente mejores que la aproximación de PUL con enfoque en redes bayesianas *p*ul-nb-b.

4.3.3 Resultados cuando se varía el número de intervenciones

En todos los experimentos que se han llevado a cabo previamente en este capítulo, se ha construido un clasificador binario independiente para cada uno de los Miembros de Parlamento que hubiesen participado en las iniciativas de los debates parlamentarios un mínimo de 10 veces. Esto constituye un conjunto bastante heterogéneo de Miembros de Parlamento, en efecto, hay un

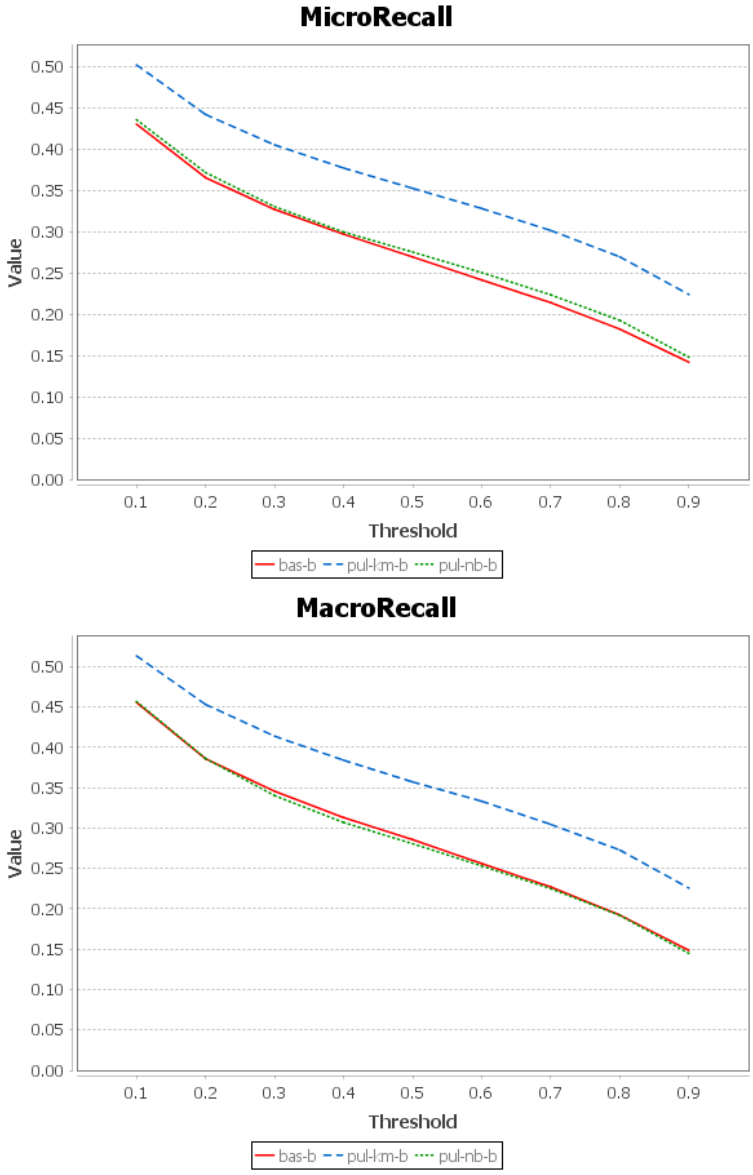

Figura 4.7: Micro y macro *recall* para *b*as-b, *p*ul-km-b y *p*ul-nb-b usando distintos valores de umbral.

conjunto relativamente amplio de MPs que participan en cientos de iniciativas parlamentarias a lo largo de la legislatura y, por otro lado, existe un grupo aún mayor de Miembros del Parlamento cuya labor es más pasiva en lo referente a su participación en los debates, en los cuales apenas están presentes más de una decena de veces. El objetivo principal de esta sección es el de visualizar una vez más el comportamiento de las aproximaciones que se plantean cuando se impone un valor superior al número mínimo de intervenciones que un Miembro del Parlamento debe tener para formar parte del estudio (Figura 3.5).

Dicho esto, al igual que se describe en el Capítulo 3, se ha llevado a cabo el mismo proceso

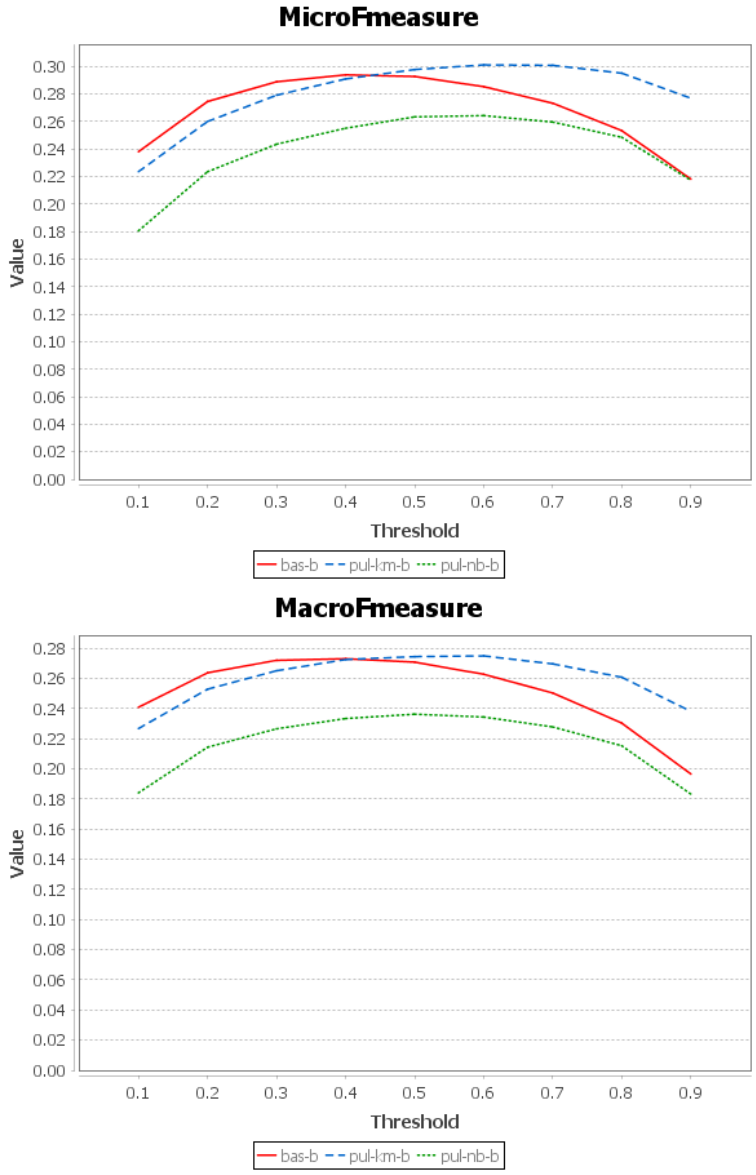

Figura 4.8: Micro y macro *F-measure* para *b*as-b, *p*ul-km-b y *p*ul-nb-b usando distintos valores de umbral.

de experimentación pero solo teniendo en cuenta aquellos Miembros del Parlamento que hayan participado al menos 25, 75 y 150 veces en las iniciativas parlamentarias. La hipótesis sobre la que se sustentan estos experimentos es la de que los resultados van a ir mejorando de forma progresiva en la medida en que un número mayor de iniciativas requeridas va a excluir a aquellos Miembros del Parlamento cuyos clasificadores asociados sean menos precisos como consecuencia de ser entrenados con conjuntos de datos de mala calidad. En la Tabla 4.4, quedan representados los valores donde se alcanzan las mejores medidas de *F-measure* para las tres aproximaciones en los casos donde se ha aplicado la técnica de balanceo de datos y en los casos donde no. Además,

Aproximación	bas-b	pul-km-b	pul-nb-b
		Micro-F	
Valor	0.2940	**0**.3012	0.2643
Umbral	0.4	0.6	0.6
		Macro-F	
Valor	0.2732	**0**.2751	0.2364
Umbral	0.4	0.6	0.5

Tabla 4.3: Mejor micro and macro *F-measure* obtenida por *b*as-b, *p*ul-km-b y *p*ul-nb-b.

Aproximación	bas	bas-b	pul-km	pul-km-b	pul-nb	pul-nb-b
			Micro-F			
mF10	0.2978	0.2940	**0**.3105	0.3012	0.2802	0.2643
mF25	0.3037	0.3038	**0**.3175	0.3084	0.2859	0.2705
mF75	0.3558	0.3597	**0**.3768	0.3647	0.3437	0.3072
mF150	0.4408	0.3987	**0**.4446	0.4171	0.4183	0.3584
			Macro-F			
MF10	0.2475	0.2732	0.2644	**0**.2751	0.2454	0.2364
MF25	0.2658	0.2920	0.2863	**0**.2941	0.2630	0.2511
MF75	0.3355	0.3563	**0**.3694	0.3629	0.3361	0.2887
MF150	0.4039	0.3761	**0**.4236	0.3984	0.3976	0.3407

Tabla 4.4: Mejor micro y macro *F-measure* obtenida por *b*as, *b*as-b, *p*ul-km, *p*ul-km-b, *p*ul-nb y *p*ul-nb-b con diferentes valores mínimos de intervenciones.

puesto que es la aproximación donde se obtienen los mejores resultados, en la Figura 4.9, se pueden observar los valores para las medidas de carácter global micro y macro *F-measure* obtenidas en la aproximación *p*ul-km usando valores distintos para establecer el umbral de relevancia.

Tal y como se puede apreciar en la Tabla 4.4, en efecto los resultados obtenidos en todas las aproximaciones mejoran de forma significativa en función de como aumenta el número mínimo de intervenciones necesarias para formar parte del estudio y en la Figura 4.9 este hecho se confirma de forma visual y quedando así comprobada de forma empírica la hipótesis que se planteaba anteriormente con respecto a cómo los MPs con un número bajo de intervenciones afectaban negativamente al sistema. Además se puede observar también que los méritos relativos a cada una de las aproximaciones permanecen de forma proporcional: *p*ul-km sigue siendo la mejor aproximación seguida del caso base *b*as dejando una vez más a la aproximación *p*ul-nb en el peor puesto. También cabe destacar que, desde el punto de vista de la medida micro *F-measure*, balancear el conjunto de datos de entrenamiento es contraproducente y de forma más concisa, cuando se aumenta el número mínimo de intervenciones a valores altos (75 y 150) aplicar SMOTE para balancear pasa a ser igual de ineficiente en lo que se refiere a la medida global macro *F-measure*.

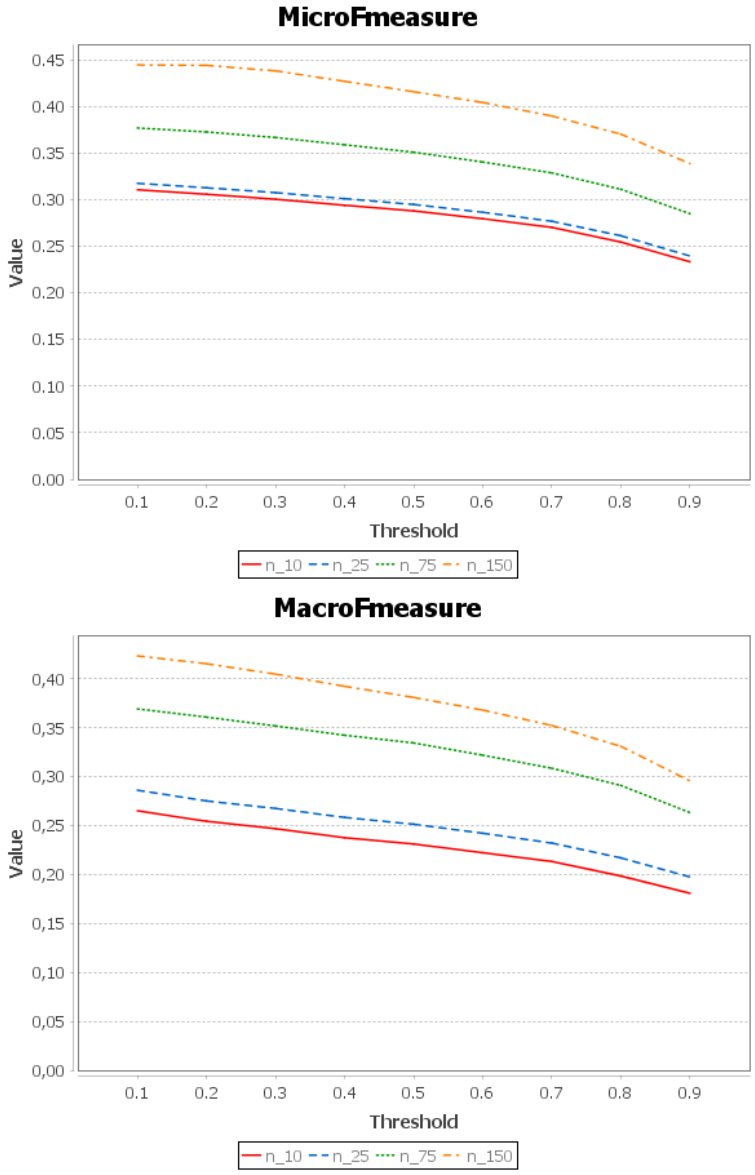

Figura 4.9: Micro y macro *F-measure* para *p*ul-km usando distintos valores de umbral, para un mínimo de 10, 25, 75 y 150 intervenciones.

4.3.4 Comparativa con aproximaciones basadas en Recuperación de Información

En esta sección, se plantea realizar una comparativa entre la aproximación que se propone en este caipítulo, *p*ul-km, y las dos aproximaciones basadas en Sistemas de Recuperación de Información descritas en el Capítulo 3 [39]. Estas aproximaciones utilizan los documentos en $\mathscr{D}$ como entrada para construir un Sistema de Recuperación de Información (IRS) y, en ambos casos, los documentos que necesitan ser filtrados o recomendados se lanzan como consulta contra el IRS,

el cual devuelve una lista ordenada de forma decreciente de los Miembros del Parlamento que son más afines a dicha consulta.

	Micro-F			Macro-F		
Aproximación	pul-km	ir-i	ir-p	pul-km	ir-i	ir-p
10	**0**.3105	0.2896	0.2892	**0**.2644	0.2423	0.2513
25	**0**.3175	0.2971	0.2939	**0**.2863	0.2661	0.2829
75	**0**.3768	0.3509	0.3085	**0**.3694	0.3288	0.3368
150	**0**.4446	0.4282	0.3120	**0**.4236	0.3948	0.3530

Tabla 4.5: Mejor micro y macro *F-measure* obtenida por *p*ul-km, *ir*-i y *ir*-p, con distintos valores mínimos de intervenciones.

En la primera de las aproximaciones basadas en la aplicación de técnicas de Recuperación de Información el IRS indexa todas las intervenciones de los Miembros del Parlamento del conjunto de entrenamiento como documentos independientes, es decir, todos los documentos en $\mathscr{D}$. Esta aproximación se denomina *ir*-i. Por el contrario, en la aproximación a la que se ha denominado *ir*-p se construye, previamente a ser indexado por el IRS, un perfil de los Miembros de Parlamento el cual está compuesto de los textos de todas las intervenciones del MP en cuestión recogidas en un único documento de mayor tamaño, es decir todos los documentos en $\mathscr{D}_i$ conforman un único documento $d_i = \cup_{j=1}^{m_i} d_{ij}$. En ambos casos, como cada uno de los documentos, independientemente de cómo esté representado, está ligado de forma unívoca a un Miembro del Parlamento, se puede interpretar el ranking de documentos más afines a la consulta, como un ranking de MPs. No obstante, cabe destacar que en el ranking obtenido en la aproximación *ir*-i se pueden encontrar MPs replicados, es decir, más de una de sus intervenciones se ha considerado relevante y aparece en el ranking. Puesto que el ranking debe presentarse como una lista de MPs únicos, se procede a eliminar todas las ocurrencias de un mismo MP en la lista exceptuando la primera aparición, la cual se corresponde con el máximo valor de score.

Puesto que se utiliza un único valor de umbral de relevancia para filtrar o recomendar un documento a los Miembros de Parlamento cuyo valor de score sea mayor que el umbral fijado y, como el valor de los scores que devuelve el IRS está afectado de forma directa por el número de términos que componen la consulta, se debe remarcar la necesidad de normalizar los valores de score dividiendo estos por el valor máximo de score del ranking. De esta forma se procura que el rango en el que se muevan los scores sea independiente de las características de la consulta.

En la Tabla 4.5 se pueden observar los mejores valores alcanzados por las medidas de *F-measure* para las dos aproximaciones basadas en Recuperación de Información en una comparativa con las mejores medidas obtenidas en la aproximación *p*ul-km.

Tal y como se puede apreciar, es evidente que la aproximación *p*ul-km funciona considerablemente mejor que cualquiera de las aproximaciones clásicas basadas en la aplicación de técnicas de Recuperación de Información. De hecho, el test estadístico t-test demuestra que existen diferencias estadísticas significativas entre la aproximación *p*ul-km y las aproximaciones

*i*r-i y *i*r-p en todos los casos, a excepción de en la medida de carácter global macro *F-measure*, en la aproximación *i*r-i con 150 como número mínimo de intervenciones, y en la aproximación *i*r-p con 25 intervenciones, donde no se encuentran diferencias significativas.

4.4 Conclusiones sobre el uso de Positive Unlabeled Learning

En este capítulo se ha propuesto una aproximación para construir un sistema que debe ser capaz, de forma automática, de filtrar o recomendar documentos a los Miembros del Parlamento diseñado mediante técnicas de Aprendizaje Automático y, más específicamente, con técnicas de clasificación automática de documentos. La fuente de la que ha obtenido la colección de documentos que se han usado para construir los clasificadores binarios que componen este sistema tiene como origen las intervenciones de los MPs en lo debates parlamentarios bajo la suposición de que en estas intervenciones se encuentra información de forma implícita sobre los intereses y preferencias políticas de los MPs. No obstante, las intervenciones de los Miembros del Parlamento solo proporcionan información sobre qué cosas son consideradas como relevantes para un MP en cuestión, pero no aportan información alguna sobre los diversos ámbitos que no son de su incumbencia. Por esta razón, la aproximación que aquí se plantea utiliza técnicas de Positive Unlabeled Learning puesto que el funcionamiento de los clasificadores binarios clásicos, los cuales deben ser entrenados con un subconjunto de instancias negativas y otro de instancias positivas, se ve truncado al no disponer de un conjunto de ejemplos clasificados como negativos a ojos del clasificador. En este contexto, también se propone una novedosa técnica PUL, *p*ul-km, que en primer lugar obtiene un conjunto consistente de instancias negativas a partir del conjunto de instancias no etiquetadas, que se corresponden con las intervenciones de otros MPs, y a continuación utiliza el conjunto de instancias positivas y el conjunto consistente de instancias negativas para entrenar un clasificador binario de la forma tradicional (SVM en este caso). El método que aquí se propone con el objetivo de obtener un conjunto consistente de instancias negativas está enfocado tomando como base una modificación del algoritmo de agrupamiento clásico K-means. Además, se han tenido en consideración los problemas derivados del no balanceo del número de instancias de las clases del conjunto de datos de entrenamiento con un algoritmo de balanceo que añade instancias sintéticas a la clase minoritaria, denominado por sus siglas SMOTE.

Tomando como base las intervenciones de los Miembros del Parlamento en el Parlamento de Andalucía, el proceso experimental que se ha llevado a cabo compara la aproximación que en este capítulo se plantea, *p*ul-km, con otras aproximaciones distintas: un caso base donde se toma en consideración que todas las intervenciones ajenas a un MP en cuestión, las no etiquetadas, se consideran como subconjunto de entrenamiento negativo; y llevar a cabo el mismo proceso pero con la aplicación de una técnica de PUL alternativa ya existente basada en Naive Bayes; y, finalmente, dos aproximaciones basadas en el enfoque de las técnicas clásicas de Recuperación de

Información las cuales indexan la colección de documentos y recuperan aquellos Miembros del Parlamento que tienen una mayor afinidad con el documento que debe ser filtrado o recomendado. En todos los experimentos que se han llevado a cabo, la aproximación de PUL basada en la modificación del algoritmo K-means obtiene mejores resultados que el resto de sus oponentes y, de forma general con importantes y significativas diferencias desde el punto de vista estadístico, por lo que se puede afirmar de forma concisa que la aproximación *p*ul-km parece comportarse de mejor manera a la hora de lidiar con los problemas del filtrado y la recomendación en el ámbito de este caso de estudio. De hecho, la novedosa técnica de PUL que se plantea en este trabajo deja fuera de combate a la técnica más referida en la literatura, *p*ul-nb lo cual es una potente evidencia de que *p*ul-km podría tener un mayor potencial de utilidad en problemas de otro ámbito donde es necesario recurrir a la aplicación de métodos de PUL.

A la vista de los resultados obtenidos a partir de utilizar o no utilizar el método SMOTE para lidiar con el problema de los conjuntos de datos de entrenamiento con clases no balanceadas, se ha podido observar que la tendencia al usarlos hace que empeoren los resultados obtenidos con la medida micro *F-measure* y, sin embargo los valores de la medida macro *F-measure* se ven beneficiados, salvo en la aproximación *p*ul-nb, donde la medida macro *F-measure* también se ve afectada negativamente. Esto se debe a que probablemente la aplicación de métodos de balanceo de clases solo sea recomendable para el conjunto de datos de entrenamiento de aquellos MPs con un número menor de intervenciones. Además, en lo referente a los problemas que se derivan del proceso de evaluación que se ha seguido hasta el momento, se plantea la cuestión de cómo afectaría encontrar un valor de umbral de relevancia individualizado para cada uno de los Miembros del Parlamento para llevar a cabo la evaluación del ranking.

4.5 Selección de umbrales de relevancia para mejorar los Sistemas de Recomendación

En los experimentos llevados a cabo anteriormente se ha detectado que el correcto funcionamiento de los clasificadores binarios que se han entrenado está estrechamente relacionado con el número de intervenciones de las que se disponen del MP asociado a dicho clasificador. Sin embargo, la diferencia entre la cantidad de veces que unos y otros Miembros del Parlamento participan en los debates parlamentarios es bastante significativa y, por lo tanto, esto provoca que los clasificadores y las probabilidades de pertenencia a las clase positiva que estos devuelven se comporten de forma distinta. Por ejemplo, un MP que apenas haya intervenido una decena de veces, a la hora de asignarle un documento como relevante, un valor probabilidad de pertenencia a la clase positiva del documento relativamente bajo derivado de la baja cantidad de información de la que dispone el clasificador podría ser suficiente para asignar el valor de relevancia positiva al documento y, sin embargo, el clasificador asociado a un MP con cientos de intervenciones, siempre va a ser más preciso con probabilidades de pertenencia a la clase positiva más altas. De este modo,

con el criterio de evaluación que se ha seguido hasta el momento donde se establecía un valor común de umbral de relevancia para todos los clasificadores de los Miembros de Parlamento, se estaba penalizando en cierto modo a MPs con pocas intervenciones en el caso de establecer el umbral de relevancia a valores altos y de la misma forma para los MPs con mayor número de intervenciones en valores del umbral de relevancia bajos. Esto tiene como consecuencia que algunos clasificadores de los MPs no sean lo suficientemente selectivos, sobrecargando al Miembro del Parlamento con más información de la que pueda necesitar y, por el contrario, se pueden generar ciertos clasificadores demasiado restrictivos que pueden producir que algunos Miembros del Parlamento pierdan información que probablemente les pudiera resultar de utilidad. A la vista de esta casuística, en esta sección se relatan los métodos planteados para realizar distintas aproximaciones con el objetivo de establecer de forma individualizada el valor de umbral de relevancia para los distintos MPs.

4.5.1 Aproximaciones para determinar los umbrales de relevancia

Sea d un documento nuevo que llega al sistema con el objetivo de ser filtrado o recomendado a los pertinentes Miembros del Parlamento en función de sus intereses o preferencias políticas. Como resultado de lanzar dicho documento contra el sistema a modo de consulta se obtiene un conjunto n (uno por cada MP) de valores numéricos $p_i(d)$, $i = 1\ldots,n$ que oscilan en el intervalo [0,1] los cuales representan la probabilidad de pertenencia a la clase positiva, es decir relevante, del documento d para el Miembro del Parlamento MP_i.

A continuación, con el objetivo de tomar una decisión con respecto a si el documento d debe ser enviado o descartado como relevante para el Miembro del Parlamento MP_i es necesario calibrar en primera instancia qué valores de probabilidad de pertenencia a la clase positiva son los más adecuados para cada uno de los distintos Miembros del Parlamento para establecer a partir de los cuales la relevancia de un documento. Probablemente el lugar donde establecer el valor del umbral de forma mas simple y natural sería en 0.5, puesto que de esta forma se comparan equitativamente la probabilidad de que el documento sea relevante $p_i(d)$ con respecto a la probabilidad de que no lo sea $1 - p_i(d)$. Por lo tanto, si $p_i(d) \geq 1 - p_i(d)$, y esto solo ocurre cuando $p_i(d) \geq 0.5$, el documento d pasa a ser considerado como relevante para MP_i. Dicho esto y como, para este problema, se han considerado todos los Miembros del Parlamento de igual forma, el razonamiento obvio sería establecer el valor del umbral de relevancia en 0.5 para todos los MPs.

Generalizando sobre esta estrategia, se puede seleccionar un valor de umbral de relevancia t, con $0 \leq t \leq 1$, y asumir que d es relevante para un Miembro del Parlamento MP_i si, y solo si $p_i(d) \geq t$. Esto puede cobrar cierta lógica en el sentido en que, por alguna razón, los clasificadores tienen tendencia a generar una masa de probabilidad sesgada a los extremos del intervalo [0,1], bien con valores de probabilidad de pertenencia a la clase positiva muy altos, o bien con valores muy bajos, no dándose prácticamente en ningún caso que la masa de probabilidad esté centrada

con respecto al intervalo [0,1]. Avanzando un paso más en el proceso de investigación, se ha podido detectar que el comportamiento de los clasificadores binarios de los diferentes Miembros del Parlamento es completamente distinto, es decir, los clasificadores se ven afectados de forma directa con el número de intervenciones de los respectivos MPs en los debates parlamentarios lo que genera conjuntos de entrenamiento completamente dispares. Otra razón a tener en cuenta con respecto a la diferencia de comportamiento de los clasificadores puede ser que existen Miembros del Parlamento que tienen un abanico de intereses políticos muy amplio con respecto a otros, lo cual se traduce en MPs con una participación en múltiples comisiones en las cuales se tratan temáticas muy diferentes y, por lo tanto, se generan discursos más diversos, heterogéneos y poco precisos. Para estos casos, entre otros, es interesante establecer un valor del umbral de relevancia t_i individualizado que dependa de forma unívoca de las propias características de un Miembro del Parlamento en cuestión y, de este modo considerar un documento d como relevante para MP_i solamente si $p_i(d) \geq t_i$. Sería obvio pensar que la cuestión que se plantea a partir de esto es el cómo o qué aproximación sería la más adecuada para seleccionar los valores del umbral de relevancia t_i más apropiados para cada Miembro del Parlamento MP_i de forma independiente. Además cabe destacar que establecer este valor de umbral individualizado también dependerá seguramente de que el conjunto de entrenamiento del MP en cuestión esté balanceado o no y, en consecuencia, el criterio para determinar si balancear o no el conjunto de entrenamiento de un MP viene definido de la misma forma por ciertas características de los propios Miembros del Parlamento.

4.5.1.1 Conjunto de validación para estimar los umbrales de relevancia

De forma general, la forma más estandarizada de establecer los parámetros de configuración de un algoritmo de clasificación y más específicamente, en este caso de estudio, de encontrar el valor adecuado para el umbral de relevancia es hacer uso de un conjunto de validación extraído de forma aleatoria del subconjunto de datos de entrenamiento [135]. De esta forma, en esta aproximación, el conjunto de entrenamiento disponible para cada Miembro del Parlamento se ve dividido en un nuevo subconjunto de entrenamiento algo más pequeño y un subconjunto de validación, ambos disjuntos entre sí. El nuevo subconjunto de entrenamiento se utiliza para que el clasificador binario aprenda y a continuación se lanzan las instancias de validación contra el clasificador con el objetivo de obtener el valor de probabilidad de pertenencia a la clase positiva $p(d)$ de cada una de ellas. Asumiendo que existe alguna forma de evaluar el rendimiento del clasificador con las instancias de validación, la cual será especificada de forma concreta en la siguiente sección, se pueden aplicar distintos valores de umbral de relevancia para comprobar en qué valor, t, el clasificador obtiene mejores resultados. Como paso final de esta aproximación, se vuelve a entrenar el clasificador del Miembro del Parlamento con el conjunto de entrenamiento completo y estableciendo como valor de umbral de relevancia individualizado aquél obtenido anteriormente t.

Esta aproximación se basa en la suposición lógica de que el rendimiento del clasificador debe ser similar tanto con el conjunto de validación como con el conjunto de test. Una posible problemática de esta suposición es que, al no entrenar el clasificador binario con el conjunto de entrenamiento completo sino con un subconjunto más pequeño, el comportamiento puede cambiar con respecto al clasificador que finalmente va a definir al MP puesto que en algunos casos, la cantidad de instancias remanentes puede no ser lo suficientemente grande. Por otro lado, otro posible problema puede venir derivado de que el número de instancias de validación sea también muy escaso y por lo tanto no se pueda extraer una conclusión fiable para establecer el valor del umbral de relevancia. Dicho esto, la siguiente aproximación intenta lidiar con el problema de los MPs con un conjunto de entrenamiento pobre en lo que a número de intervenciones se refiere.

4.5.1.2 Conjunto propio de entrenamiento para estimar los umbrales de relevancia

En lugar de utilizar un subconjunto de validación, el cual puede acabar siendo muy reducido, en la propuesta de esta aproximación se plantea usar el conjunto completo de instancias de entrenamiento tanto, como conjunto de validación para inducir la estimación del mejor valor para el umbral de relevancia, como para entrenar el clasificador final asociado al MP. En este sentido se utiliza el conjunto completo de instancias de entrenamiento para que el clasificador aprenda y a continuación se lanza de nuevo el mismo conjunto de entrenamiento para obtener las probabilidades de pertenencia a la clase positiva $p(d)$ para cada una de las instancias y, de este modo, probando con distintos valores de umbral de relevancia establecer cual de ellos es el que el obtiene un mejor rendimiento para el clasificador.

Esta aproximación se propone con el objetivo de solventar los problemas derivados de usar un subconjunto de validación ya que, por un lado, el número total de instancias que se utilizan para obtener el mejor valor para el umbral de relevancia es considerablemente mayor y, por otro lado, se usa exactamente el mismo clasificador para obtener el valor del umbral de relevancia que el que se va a utilizar finalmente. No obstante, esta aproximación genera un nuevo y obvio problema relacionado con el riesgo de sobreajuste en el sentido en que se están clasificando las mismas instancias con las que se ha entrenado el clasificador. Y aunque no está claro si este sobreajuste puede tener influencia a la hora de encontrar el correcto valor para el umbral de relevancia, sí cabe esperar que por lo general las probabilidades de pertenencia a la clase positiva de las instancias de entrenamiento sean más altas y de la misma forma esto afecte al umbral.

4.5.1.3 Relacionar los umbrales de relevancia con características de los MPs

Volviendo la vista a una sección previa de este mismo capítulo, donde se obtenía el mejor valor posible para el umbral de relevancia a partir de los mejores resultados de las medidas de rendimiento en el conjunto de test, usando así información privilegiada, se puede observar que los umbrales obtenidos cuando no se balancea el conjunto de entrenamiento de los Miembros del Parlamento son generalmente bajos. En contrapartida, cuando se balancea el conjunto de

entrenamiento de los MPs, los mejores valores alcanzables para el umbral de relevancia están situados en este caso en torno a los valores centrales del intervalo [0,1]. Este comportamiento particular lleva a tener en consideración la posibilidad de que la clave sobre la que tomar la decisión de balancear o no balancear el conjunto de entrenamiento de un Miembro del Parlamento puede venir dada por las propias características del perfil del MP, puesto que, el simple hecho de alterar de alguna forma el conjunto de entrenamiento, como en este caso es aplicando alguna técnica de balanceo, el valor del umbral de relevancia cambia, lo cual afecta en muchos casos de forma positiva.

A la vista de esta hipótesis, se han extraído del perfil de los Miembros del Parlamento ciertas características con el objetivo de encontrar cuales de ellas podrían estar correladas de alguna forma con el correspondiente valor del umbral de relevancia balanceando o no balanceando el conjunto de entrenamiento. Se han tenido bajo consideración múltiples características de los perfiles de los Miembros del Parlamento pero finalmente se han preservado aquellas con un índice de Gini mejor con respecto al valor del umbral de relevancia para llevar a cabo el proceso de experimentación. Las características evaluadas han sido: el número de intervenciones $interventions$ de los MPs, el número total de términos únicos distintos en todas las intervenciones antes $terms$ y después $NP-terms$ de procesar el texto, el promedio de términos por intervención antes $meanTermTnterv$ y después $NP-meanTermInterv$ de procesar el texto y finalmente la densidad léxica $lexicalDensity$ [145], que representa el cociente entre el número de unidades léxicas (nombres, verbos, adjetivos, adverbios) y el número total de términos.

Dicho esto, se pretende realizar el estudio de las correlaciones que se pueden obtener entre cada una de las características obtenidas y el valor del umbral de relevancia. Una correlación alta, ya sea de forma positiva o negativa, entre una característica determinada del perfil del MP y el valor del umbral puede ser considerada como determinante a la hora de establecer el valor de umbral de relevancia más apropiado para un Miembro del Parlamento. Además, destacar que también se han construido algunos modelos de predicción utilizando las características del perfil del MP juntas.

4.6 Evaluación experimental con umbrales de relevancia

En esta sección, tanto la colección con la que se van a llevar a cabo los correspondientes experimentos, como el procesamiento previo de esta, es completamente idéntico al de procesos de experimentación anteriores. Matizando que se van a entrenar los 132 clasificadores asociados a cada uno de los MPs pero en versiones con el conjunto de datos balanceado y no balanceado. A continuación, se lanzan las intervenciones de test contra todos los clasificadores para obtener el porcentaje de pertenencia a la clase positiva $p_i(d)$ de la consulta para cada uno de los MPs.

Una vez se han calculado los valores de $p_i(d)$ para cada iniciativa d del conjunto de test y para cada MP_i (realmente para cada MP i se calculan dos valores, $p_i^b(d)$ y $p_i^n(d)$, para los

clasificadores balanceados y no balanceados asociados a un MP_i, respectivamente), se comparan estos valores con un umbral fijado t_i (en realidad son dos valores de umbral para cada MP i, t_i^n y t_i^b), con el objetivo de determinar qué documento d es relevante a un MP_i. De este modo se pueden calcular, para cada MP_i, el número de Verdaderos Positivos (TP_i), Falsos Positivos (FP_i) y Falsos Negativos (FN_i), con el objetivo de calcular las mismas medidas de evaluación que venimos usando hasta ahora: *precision* (p_i), *recall* (r_i) y *F-measure* (F_i), tanto en su versión macro como en su versión micro.

El proceso de evaluación se lleva a cabo una vez se ha seleccionado el valor de umbral t_i para cada uno de los clasificadores. Y el resultado de la experimentación aporta información sobre la calidad del método de selección de umbrales que se haya utilizado, a partir del rendimiento que se obtenga en el sistema con el método evaluado. Los valores posibles para el umbral de relevancia t_i van desde el mínimo 0.1 hasta el máximo 0.9, independientemente de la aproximación que se utilice para estimar el valor.

Para los experimentos donde se usa el propio conjunto de entrenamiento para estimar los mejores valores de umbral, se lleva a cabo el procedimiento descrito pero con la particularidad de que se utilizan los documentos de entrenamiento en lugar de los de test para calcular la medida F_i, para cada uno de los valores de umbral posibles, y considerando solo el umbral con el mejor valor de *F-measure*. Por otro lado, en los experimentos donde se utiliza un conjunto de validación para obtener los mejores valores de umbral, primero se divide de forma aleatoria cada conjunto de entrenamiento para extraer un nuevo conjunto de entrenamiento (80% de las instancias del conjunto de entrenamiento original) y un conjunto de validación (20% del conjunto de entrenamiento original). Entonces, se construye otro conjunto de clasificadores con el nuevo subconjunto de entrenamiento con los que se evalúan las instancias del subconjunto de validación. De esta forma, se calcula la medida F_i, para cada uno de los valores de umbral posibles, y se considera solo el umbral con el mejor valor de *F-measure*.

Como última aproximación, además de usar las versiones de los clasificadores con el conjunto de datos balanceado y no balanceado de forma independiente, se ha implementado un método combinado, es decir, para cada MP_i se evalúan (usando tanto el conjunto de validación como el conjunto de entrenamiento al completo) ambos clasificadores, obteniendo de esta forma el mejor umbral para cada uno de ellos, t_i^n y t_i^b, y por último se selecciona la versión del clasificador cuya evaluación con el umbral t_i obtenga el mejor valor de *F-measure*.

4.6.1 Resultados

Los resultados del proceso de experimentación llevado a cabo en este punto de la investigación se sitúan en la Tabla 4.6. Como complemento a los resultados obtenidos en los experimentos donde se usaba un conjunto de validación o el conjunto completo de entrenamiento, también se muestran los resultados de la aproximación basal la cual establecía de forma fija el valor del umbral de relevancia a 0.5 para todos los clasificadores. Destacar también que se representan en

la misma tabla los resultados de las aproximaciones balanceadas, no balanceadas y la versión combinada de los clasificadores.

Tabla 4.6: Valores para la macro y micro *F-measure* obtenidos en todos los experimentos.

Caso Base con Umbral Estático (0.5)		
	Macro F-measure	**micro F-measure**
No Balanceado	0.2343	0.2967
Balanceado	0.2722	0.2944
Umbral Variable (Validación)		
	Macro F-measure	**micro F-measure**
No Balanceado	0.2275	0.2767
Balanceado	0.2612	0.2709
Combinado Balanceado/No Balanceado	0.2436	0.2844
Umbral Variable (Entrenamiento)		
	Macro F-measure	**micro F-measure**
No Balanceado	0.2556	0.3129
Balanceado	0.2385	0.2912
Combinado Balanceado/No Balanceado	0.2541	0.3079
Solución Ideal Alcanzable (Test)		
	Macro F-measure	**micro F-measure**
No Balanceado	0.2807	0.3322
Balanceado	0.3193	0.3273
Combinado Balanceado/No Balanceado	0.3220	0.3435

En lo referente a los resultados del caso base, se puede observar que los valores obtenidos en términos de micro *F-measure* son bastante similares para las aproximaciones donde se balancea el conjunto de entrenamiento de los clasificadores y donde no se balancea, la cual toma una ligera ventaja con respecto a la primera, sin embargo, la aproximación balanceada es claramente mejor desde el punto de vista de la medida de carácter global macro *F-measure*. Esto parece indicar que el hecho de balancear el conjunto de entrenamiento de forma particular mejora el rendimiento de los clasificadores asociados a aquellos MPs con un menor número de intervenciones, los cuales son precisamente aquellos que disponen de un conjunto de entrenamiento más desbalanceado. Este tipo de Miembros del Parlamento, desde el punto de vista de la medida macro *F-measure*, tienen la misma importancia que el resto de MPs con mayor número de intervenciones, cosa que no ocurre en el caso de la medida micro *F-measure*, en la cual adquieren un papel más pasivo.

Fijando la vista ahora en los resultados obtenidos con la aproximación donde se usaba el subconjunto de validación, cabe decir que son desalentadores y que esta aproximación no es buena para este problema puesto qué se obtienen resultados de forma sistemática mucho peores que los obtenidos en el caso base (entre el 3% y el 8% de perdidas). Por lo tanto, a pesar de ser una práctica común en la literatura la de usar un subconjunto de validación con el objetivo de estimar los parámetros de un clasificador, en nuestro caso de estudio esta aproximación se comporta de peor forma. Este comportamiento responde a las sospechas con las que se anticipaba

esta aproximación en lo referente al pequeño número de intervenciones que conformaban los subconjuntos de validación de muchos de los Miembros del Parlamento, donde por ejemplo, para un MP con 20 intervenciones totales en su conjunto de entrenamiento solo 3 de ellas formaban parte del subconjunto de validación, las cuales no son suficientes para capturar las características que definen el perfil de un MP.

Con la intención de atajar el problema producido por el pequeño número de intervenciones en los conjuntos de validación de algunos MPs, se ha llevado a cabo el mismo procedimiento de experimentación pero esta vez con el conjunto de entrenamiento al completo. Lo que se espera de esta aproximación es que cuanto mayor sea el número de documentos mejor se ajustará el valor del umbral de relevancia. A la vista de los resultados de la Tabla 4.6 se puede corroborar que esta suposición es mayormente cierta en la aproximación no balanceada, donde ambas medidas de carácter global, macro y micro *F-measure*, aumentan sus valores con respecto a los resultados del caso base (un 9% y un 5% respectivamente). Cabe destacar que la mejora de rendimiento en la medida micro *F-measure* en esta aproximación es considerable, puesto que obtiene el valor más alto de todas las medidas homólogas en el resto de clasificadores. Esto ocurre porque en esta aproximación los valores del umbral de relevancia de los MPs con mayor número de intervenciones, los cuales son considerados de mayor importancia para esta medida, están mejor estimados. Sin embargo, la medida macro *F-measure* para esta aproximación no es buena en términos generales probablemente a causa del hecho de que, en esta medida en concreto, se está dando la misma importancia a todos los Miembros del Parlamento independientemente de la calidad de su conjunto de entrenamiento en lo que a número de intervenciones se refiere y por lo tanto, los valores de umbral obtenidos a partir de un conjunto de entrenamiento poco robusto no están bien estimados. Por otro lado, la aproximación combinada de balanceo o no balanceo en este caso no mejora en ningún caso los resultados obtenidos por la aproximación donde no se aplica balanceo del conjunto de entrenamiento. Y finalmente, una vez más la aproximación balanceada obtiene peores resultados para esta aproximación que los que arroja el caso base. Para tener una mejor perspectiva de los resultados obtenidos usando los subconjuntos de validación o bien el conjunto de entrenamiento al completo, en la Tabla 4.6 se muestran además los resultados ideales que se podrían alcanzar con estas medidas en las aproximaciones balanceada, no balanceada y combinada. Estos valores se calculan a partir de seleccionar los mejores umbrales (y en el último caso, decidiendo además por cada MP si se balancea o no) basándose en los resultados obtenidos en el conjunto de test. Estos resultados muestran que la decisión de balancear solo el conjunto de Miembros del Parlamento que lo necesiten está bien planteada, al menos en teoría, si se pudiera encontrar de forma unívoca aquellos Miembros del Parlamento cuyo conjunto de entrenamiento debe ser balanceado o no balanceado, de hecho, se obtienen mejoras del 1% y el 15% para macro *F-measure* y del 5% y 3% para micro *F-measure*, en referencia a las aproximaciones balanceada y no balanceada respectivamente. Además se puede observar que, desde el enfoque de la medida micro *F-measure*, siempre es preferible no balancear los conjuntos de datos de entrenamiento y,

sin embargo, lo contrario ocurre de forma general en la medida macro *F-measure* cuyo valor sale perjudicado si no se balancean los conjuntos de entrenamiento de los Miembros del Parlamento. Para la medida micro *F-measure*, los mejores resultados que se han alcanzado han sido usando el conjunto de entrenamiento al completo para estimar los valores del umbral de relevancia y además no aplicando ningún método de balanceo de clases, alcanzando así hasta un 91% del rendimiento ideal. Para la macro *F-measure*, lo mejor a lo que se ha podido aspirar para esta medida ha sido establecer un umbral fijo para todos los MPs y balancear los conjuntos de entrenamiento de estos, alcanzando así el 85% del rendimiento ideal.

Con el objetivo de garantizar la exhaustividad del proceso de experimentación, en la Tabla 4.7 se representan los valores para las medidas de *precision* y *recall* desde el punto de vista global micro y macro correspondientes a las medidas de *F-measure* que se muestran en la Tabla 4.6. Sobre estos resultados se puede observar que, cuando se usa el conjunto de validación para estimar el valor del umbral de relevancia, los clasificadores que han sido entrenados con un conjunto de datos no balanceado obtienen de forma general un valor de *precision* relativamente alto pero al precio de un *recall* bajo. Por otro lado, cuando se utiliza el conjunto de entrenamiento completo para estimar los valores del umbral de relevancia, se obtiene una *precision* algo más baja pero en contrapartida se obtiene un *recall* considerablemente mejor. Cabe destacar que desde el punto de vista de la aplicación de filtrado, el *recall* se puede considerar con mayor importancia que la *precision* porque en este tipo de problema una instancia clasificada como falso positivo no es un error como tal, sino un MP que puede estar interesado en el documento pero en el que no ha participado de forma activa. Como norma general, el comportamiento de los clasificadores balanceados es más errático tanto cuando se usa el subconjunto de validación como cuando se usa el conjunto de entrenamiento completo para estimar los valores de los umbrales de relevancia: en el primer caso, la medida de *recall* es considerablemente mayor que la medida de *precision*, mientras que en el segundo caso ocurre lo contrario. Esto parece indicar que los valores de umbral seleccionados en estos casos son completamente diferentes, valores muy bajos en el caso de usar el subconjunto de validación y, por el contrario, muy altos en el caso de usar el conjunto completo de entrenamiento.

Con respecto a la aproximación donde se pretende relacionar de algún modo los valores de umbral de relevancia obtenidos con ciertas características de los respectivos Miembros del Parlamento, los coeficientes de correlación obtenidos para cada una de las características y el mejor valor del umbral de relevancia, tanto aplicando balanceo como no aplicándolo, se presentan en la Tabla 4.8. Incluso cuando el índice de Gini sugería que las características del perfil del MP seleccionadas eran las más importantes de entre todas las que se consideraron en primera instancia, la correlación obtenida entre todas ellas y el valor de umbral es considerablemente baja. La conclusión en este caso está clara, o bien la extracción de características de los perfiles no tiene validez para estimar el umbral, o bien ninguna de las características tomadas en consideración es buena para tal propósito.

Tabla 4.7: Valores de la macro y micro *precision* y *recall* obtenidos en todos los experimentos.

Caso Base con Umbral Estático (0.5)				
	Macro Precision	**micro Precision**	**Macro Recall**	**micro Recall**
No Balanceado	0.3117	0.3593	0.2434	0.2527
Balanceado	0.2689	0.2500	0.3810	0.3580
Umbral Variable (Validación)				
	Macro Precision	**micro Precision**	**Macro Recall**	**micro Recall**
No Balanceado	0.3258	0.3914	0.1886	0.2046
Balanceado	0.2679	0.2203	0.3792	0.3514
Combinado Balanceado/No Balanceado	0.3575	0.3754	0.2312	0.2289
Umbral Variable (Entrenamiento)				
	Macro Precision	**micro Precision**	**Macro Recall**	**micro Recall**
No Balanceado	0.3216	0.3411	0.2762	0.2891
Balanceado	0.3307	0.3523	0.2446	0.2481
Combinado Balanceado/No Balanceado	0.3278	0.3391	0.2740	0.2811
Solución Ideal Alcanzable (Test)				
	Macro Precision	**micro Precision**	**Macro Recall**	**micro Recall**
No Balanceado	0.3512	0.3692	0.2929	0.3020
Balanceado	0.3743	0.3728	0.3071	0.2917
Combinado Balanceado/No Balanceado	0.3814	0.3772	0.2984	0.3153

Además se ha llevado a cabo un experimento con el objetivo de combinar las características que definen al perfil y comprobar si esta combinación es capaz de predecir de algún modo los mejores valores para el umbral de relevancia. Para realizar esto se ha entrenado un modelo de regresión lineal con los datos asociados a estas características para los Miembros del Parlamento, usando un 80% de los datos para el subconjunto de entrenamiento y el 20% restante para el subconjunto de test. A continuación se comprobó para cada uno de los MPs en el conjunto de test las diferencias entre su mejor valor de umbral real y el predicho a través de este método, obteniendo de esta forma resultados bastante pobres.

Tabla 4.8: Correlaciones entre las diferentes características de los MPs y el mejor umbral, para los casos balanceados y no balanceados.

	Umbral No Balanceado	**Umbral Balanceado**
interventions	-0.1331	-0.2214
terms	-0.0553	-0.1941
meanTermInterv	-0.0283	-0.1475
NP-Terms	-0.0709	-0.2057
NP-MeanTermInterv	0.0556	-0.0469
lexicalDensity	0.1348	0.0064

4.7 Conclusiones sobre la selección de umbrales

En este capítulo se han considerado diversas alternativas para lidiar con el problema de encontrar el mejor valor del umbral de relevancia para utilizarlo de forma combinada con un conjunto de clasificadores de texto binarios. El objetivo principal que aquí se plantea es el de calibrar la salida

numérica generada por cada uno de los clasificadores, dado un documento para ser filtrado o recomendado, para decidir si el documento se puede considerar relevante o por el contrario no lo es. En este caso de estudio, los clasificadores han sido entrenados a partir de las intervenciones en los debates de los Miembros de Parlamento, y el objetivo de dichos clasificadores es el de filtrar o recomendar nuevos documentos a los pertinentes MPs de acuerdo a sus interese políticos. Y tomando en consideración que el conjunto de entrenamiento de un clasificador asociado a un Miembro del Parlamento puede estar desbalanceado, también se ha tenido en cuenta la posibilidad de balancear los conjuntos de datos de entrenamiento.

La primera de las aproximaciones tomaba como caso base establecer un umbral estático con valor 0.5 para todos los MPs, mientras que el resto de propuestas establecían un umbral de relevancia individualizado para cada Miembro del Parlamento, tanto usando un subconjunto de validación extraído del conjunto de entrenamiento, como usando el conjunto de entrenamiento al completo. Además se ha propuesto de forma alternativa, relacionar el umbral ideal de un MP con ciertas características de su propio perfil basándose en la forma de su discurso. Tras llevar a cabo el proceso experimental de evaluación de las diferentes aproximaciones, se pueden extraer ciertas conclusiones al respecto. En primer lugar, aunque la extracción de un subconjunto de validación sobre el conjunto de entrenamiento, con el objetivo de configurar los parámetros de un clasificador, es una técnica comúnmente utilizada en la literatura, en este caso de estudio, los resultados obtenidos no han sido buenos en absoluto, de hecho han arrojado resultados peores que los de la aproximación basal. En segundo lugar, la utilización de las mismas instancias tanto para entrenar el modelo como para validarlo y obtener el mejor valor de umbral de relevancia, incluso teniendo en consideración el riesgo de sobreajuste, funciona razonablemente bien en este caso de estudio, mejorando de forma considerable los resultados obtenido por el caso base. En tercer lugar, balancear los conjuntos de datos de entrenamiento de los MPs antes de construir los clasificadores binarios no es, en general, muy útil, aunque tienden a beneficiar a la medida de carácter global macro *F-measure*, probablemente porque balancear beneficia a aquellos MPs con un número menor de intervenciones, también perjudica al mismo tiempo los clasificadores del resto de Miembros del Parlamento. Sin embargo, en lo que a la medida micro *F-measure* se refiere, balancear el conjunto de datos de entrenamiento provoca que se obtengan resultados notablemente peores. En cuarto y último lugar, cualquier intento de relacionar ciertas características del perfil de un MP con su correspondiente umbral de relevancia para encontrar cierta dependencia han resultado ser un completo fracaso. Se han intentado encontrar algunas características, como el número de intervenciones de cada MP o el número de términos únicos en dichas intervenciones, pero sin embargo, los valores de correlación encontrados han sido ínfimos. Además, cuando de forma alternativa se propuso construir un modelo de regresión lineal utilizando para ello los datos de las características obtenidas para intentar predecir el valor del umbral, los resultados dejaron mucho que desear al respecto. Por lo tanto, a la vista de los resultados, se puede concluir que la mejor aproximación de las que en esta etapa de la

investigación se han planteado es la de estimar los valores del umbral de relevancia de forma individualizada a partir del conjunto completo de datos de entrenamiento de cada Miembro del Parlamento sin aplicar ninguna técnica de balanceo de clases previamente.

4.8 Trabajos relacionados

Las principales propuestas que se plantean en este trabajo son, en primer lugar, la aplicación de técnicas basadas en Aprendizaje Automático para abordar el problema de diseñar y construir un Sistema de Recomendación Basado en Contenido para documentos en el ámbito parlamentario tal y como se plantea en los artículos publicados [39, 41, 118] donde se muestran distintos planteamientos del problema propuesto pero desde el punto de vista de los métodos basados en Recuperación de Información más que desde el enfoque del Aprendizaje Automático. En segundo lugar, se propone la utilización de técnicas de Positive Unlabeled Learning aplicadas a la clasificación automática de documentos, donde en la literatura no aparecen muestras previas de este caso de estudio donde se vean relacionados estos conceptos [43, 47, 82, 83, 152]. Y en último lugar, se propone un nuevo diseño de una técnica de Positive Unlabeled Learning basada en una modificación del algoritmo clásico de agrupamiento K-means.

Existen múltiples estudios enfocados al problema de filtrado y recomendación de documentos en diferentes dominios y aplicaciones los cuales están incluidos entre otros en los trabajos [15, 57, 92]. Existen diversas formas sobre las que se puede construir un Sistema de Recomendación Basado en contenido, pero de entre todas ellas, los métodos más comunes en la literatura son los basado en Recuperación de Información [7, 12, 44, 90, 91] y la aplicación de algoritmos de Aprendizaje Automático para el aprendizaje de modelos de usuario [13, 32, 65, 73, 109, 143, 155]. No obstante, las aplicaciones en el contexto parlamentario son más reducidas [39, 41, 118], y en todos los casos estudiados en la literatura solo se han aplicado métodos basados en Recuperación de Información.

En de Campos et al. [39] se considera una primera aproximación al problema donde se recogen todos los discursos de los Miembros del Parlamento con el objetivo de generar una colección documental con información de los MPs en lugar de construir perfiles elaborados de Miembros de Parlamento para, posteriormente utilizar un Sistema de Recuperación de Información con el propósito de encontrar los Miembros del Parlamento más afines a los documentos que se pretenden filtrar o recomendar. Esta aproximación se intenta mejorar por los mismos autores en [41], donde se generan perfiles de términos (palabras) para los diferentes Miembros del Parlamento a partir de extraerlos directamente de los discursos a través de diversos métodos. Un aproximación distinta en la que se plantea en el trabajo [118], donde los perfiles de los MPs se construyen a partir de una serie de palabras clave las cuales han sido asignadas a la intervenciones de forma manual por un grupo de documentalistas con la ayuda de un tesauro, en lugar de utilizar los propios términos recogidos en las intervenciones.

Por otro lado, existen de forma destacable tres propuestas distintas para métodos de Positive Unlabeled Learning tal y como se describe en Zhang and Zhuo [159]. La primera clase que se recoge hace uso de una estrategia en dos pasos, donde el primer paso supone intentar identificar un subconjunto consistente de instancias negativas a partir de un conjunto de instancias no etiquetadas, y el segundo paso consiste en el entrenamiento de un algoritmo clasificador de forma tradicional con un subconjunto positivo de datos y el subconjunto negativo extraído en el primer paso [82, 83, 152]. La segunda clase de enfoque está basada en el modelo de aprendizaje probabilístico a partir de una consulta. Por ejemplo, en [43] una alteración del funcionamiento clásico de Naive Bayes para la clasificación de textos obtiene una estimación de las probabilidades condicionales de los términos dada una clase positiva de la forma usual que aplica esta técnica y, por otro lado, las probabilidades condicionales dada una clase negativa a partir de la estimación de la distribución de probabilidad a priori de la clase positiva, En [22] se plantea otro tipo de clasificador basado en redes bayesianas el cual también se adapta para ser utilizado como técnica de PUL. Finalmente, la tercera clase de métodos trata a los datos no etiquetados como un conjunto de instancias negativas donde se encuentran ejemplos considerados como ruido tras la aplicación de una regresión logística [77] o bien mediante el uso parcial del algoritmo Support Vector Machine [82], por ejemplo. Las técnicas de PUL están siendo utilizadas también en casos donde se trata con flujos de datos constantes [79] la cual es un área completamente activa de investigación [48, 59, 112]. En este trabajo se ha tomado el enfoque de la primera clase referida para las distintas forma de aplicar técnicas de PUL, la cual es la más utilizada y la más similar al enfoque que se ha planteado en este trabajo. En [82], los autores utilizan un clasificador bayesiano y las instancias positivas como respectivo subconjunto de entrenamiento positivo y los datos no etiquetados como subconjunto de entrenamiento negativo y a continuación, el clasificador bayesiano resultante se utiliza con el propósito de volver a clasificar el conjunto de instancias no etiquetadas como positivas o negativas, tomando así como subconjunto consistente de instancias negativas aquellas que en la clasificación han sido consideradas negativas por el clasificador bayesiano. Una aproximación similar se lleva a cabo en [82], donde el clasificador bayesiano se sustituye por un método de clasificación de documentos de Rocchio, el cual utiliza los pesos de tf-idf y el criterio de similitud coseno. Otra propuesta que se recoge en la técnica Spy [83] la cual selecciona de forma aleatoria un subconjunto de instancias positivas con el objetivo de añadirlo al conjunto de datos no etiquetados y, a continuación se aplica un algoritmo de *expectation-maximization* para entrenar un clasificador bayesiano a partir de este conjunto de datos y este se utiliza con el propósito de obtener un umbral el cual tiene la función de identificar un subconjunto consistente de instancias negativas. El método PEBL [152] trata de identificar algunas características (términos en este caso), que se denominan características positivas, las cuales se encuentran con más frecuencia entre los términos relativos a las instancias positivas que en los documentos que generan el resto de instancias no etiquetadas, por lo tanto, aquellos documentos que no contengan características positivas pasan a formar parte del subconjunto de

instancias consideradas como negativas. Existen también algunas propuestas (por ejemplo [47]) en las que se intenta obtener ambos subconjuntos de datos de entrenamiento, positivo y negativo, a partir de considerar todo el conjunto de datos de entrenamiento como instancias sin etiquetar.

CAPÍTULO 5

CAPÍTULO

CONSTRUCCIÓN AUTOMÁTICA DE PERFILES MULTIFACÉTICOS USANDO CLUSTERING Y APLICACIONES EN FILTRADO Y RECOMENDACIÓN DE EXPERTOS

En la era de la información en la vivimos hoy en día, no solo estamos interesados en acceder al contenido del que disponemos como vídeos, documentos, música, etc. sino también en acceder a expertos profesionales, personas o celebridades, ya sea con objetivos profesionales o lúdicos. Los sistemas de acceso a la información necesitan tener la capacidad de extraer y explotar múltiples fuentes de información (comúnmente en formato textual) sobre las características de los individuos, y representarlos de forma que pueden estar definidos por un perfil que recoja información sobre ellos. En este capítulo, se aborda el problema de la recomendación de expertos basada en perfiles y el filtrado de documentos desde la perspectiva del Aprendizaje Automático aplicando para ello técnicas de agrupamiento de fuentes textuales de expertos para construir perfiles más concretos y capturar así mejor las temáticas ocultas en las que un experto está interesado. De esta forma, los expertos pasan a estar representados por lo que se ha denominado perfiles multifacéticos. En definitiva, la experimentación realizada en este capítulo tiene como principal objetivo validar esta técnica para mejorar el rendimiento de los problemas de búsqueda de expertos y filtrado de documentos.

5.1 Introducción

El contenido disponible en internet a día de hoy es increíblemente vasto y con una cantidad de información abrumadora tanto por la cantidad como por la variedad, siendo por lo tanto una interesante tarea a investigar el hecho de cómo ayudar a los usuarios con el problema particular de acceder y gestionar a esta información de forma eficiente y práctica. Por ejemplo, se puede buscar un doctor que trate un área específica de la medicina, un electricista para reparar una avería eléctrica en una casa o un político con la pretensión de dialogar sobre un problema concreto y encontrar una solución al respecto. A este tipo concreto de búsqueda de información se le denomina como búsqueda de expertos [10] por el cual los usuarios buscan expertos en un área específica. Para que esta tarea pueda llevarse a cabo de manera satisfactoria, es necesario en primer lugar que los expertos estén representados de alguna forma en el Sistema de Recuperación de Información. La forma más precisa y especializada es la de considerar a los expertos como un perfil compuesto por los términos más representativos que definen a su área de ocupación. Estos perfiles se construyen de forma que sean tomados en consideración todos los documentos que mejor representen al experto en sí, por ejemplo, para un científico, estos documentos representativos podrían ser los artículos o congresos publicados; para un escritor, los libros que haya escrito; para un programador, los códigos fuente que haya implementado; para un abogado, los casos jurídicos en los que trabaje; y para un político, las intervenciones en los debates de un parlamento.

Con todos estos documentos, un sistema podría de forma automática construir unos perfiles de expertos a partir de seleccionar los mejores términos que definen los campos de estudio donde los propios expertos son profesionales. Por lo tanto, el sistema de Búsqueda de Expertos de utilizar estos documentos para poder encontrar coincidencias con los requisitos de información de un usuario. Existen principalmente dos problemas destacables en lo referente a la búsqueda de personas relevantes en los casos donde se haga uso de la representación de estas en forma de perfiles.

En el primero de ellos, dado un conjunto de expertos o profesionales, el problema consiste en recuperar aquellos que se ajustan mejor a las necesidades de información de un usuario, que por lo general están representadas en forma de una consulta corta de texto. En este caso, solo aquellos expertos que se sitúen en la parte alta del ranking serán los seleccionados para ser recomendados. A esto se le considera comúnmente como un problema propio del área de estudio de la búsqueda de expertos o, hablando en términos generales, un problema de recomendación basado en contenido [91].

El segundo problema consiste en que, cuando llega un documento por primera vez al sistema a modo de consulta de texto larga, el objetivo radica en este caso en la decisión sobre qué expertos deberían recibir el documento y que les sea de incumbencia. De esta forma el problema se convierte al enfoque del filtrado de documentos [57] donde se lidia con el problema de encontrar a todo experto para el que pueda existir relevancia en el documento independientemente del

ranking.

A pesar de esto, ambos problema deben ser tratados como " dos caras de una misma moneda" [12] y abordarlos con aproximaciones similares y, en este capítulo se plantean las principales diferencias que existen entre ambos enfoques en términos de cómo ambos se formulan y de cómo se interpretan. En este capítulo, se debe considerar que el campo profesional de un experto no se limita solo a un área en concreto: un científico, por ejemplo, aunque pueda estar especializado concretamente en Recuperación de Información, también puede publicar sobre distintas líneas de investigación como pueden ser Sistemas de Recomendación, Modelos de Recuperación, Personalización, etc. o un político que pertenezca a tres comisiones parlamentarias distintas a la vez, por ejemplo, economía, agricultura y medio ambiente, y cuyas intervenciones estén relacionadas con esas áreas. En estos casos, si se construye un perfil único del experto, la diversidad de las temáticas en las que este pueda estar interesado se mezclan y se diluyen hasta el punto de poder llegar a no ser detectables, y esto puede dar como resultado la extracción de temáticas más generales en detrimento de las temáticas más especializadas, derivando esto en que el perfil no refleja adecuadamente los intereses del experto y esto podría significar que estos no sean recuperados cuando se les busque por una temática especifica. Una solución a este problema podría ser por lo tanto, considerar que un perfil de experto no puede tener una forma monolítica sino una estructura multifacética que reúna otros perfiles o subperfiles donde cada uno de ellos recoja de forma presumiblemente unívoca las diferentes temáticas. De esta forma, en el caso de ejemplo del político, este podría quedar representado como un conjunto de tres perfiles distintos en los que se recojan de forma independiente todas las áreas donde desempeña su labor parlamentaria.

Siguiendo esta línea de hipótesis, en esta Tesis Doctoral se plantea realizar un estudio para encontrar personas relevantes en un contexto parlamentario. En una aproximación inicial, los perfiles de los Miembros del Parlamento se construían a partir de sus discursos parlamentarios los cuales se podían utilizar para encontrar MPs relevantes [41]. En primera instancia, los perfiles de los MPs se creaban considerando todas las intervenciones para construir un perfil monolítico y más adelante, puesto que muchos de los discursos de los Miembros del Parlamento pertenecían a comisiones parlamentarias especializadas, en [40] la composición de los perfiles pasó a ser concebida como que un MP tenía tantos subperfiles como número de comisiones en las que participaba y en cada uno de ellos se recogían solo y exclusivamente las intervenciones propias de cada comisión. En el estudio que se recoge en [40] se demuestra que este método de construir los perfiles de los expertos es mucho más interesante para un problema de recomendación por su interpretabilidad.

Dicho esto, la novedad que se presenta en este caso de estudio es ir un paso más lejos en la representación de los perfiles de los Miembros del Parlamento aplicando técnicas de Aprendizaje Automático y, de forma más específica, técnicas de clustering con el objetivo de poder detectar de forma automática las diferentes temáticas en las que un experto pueda estar interesado y de esta manera construir unos subperfiles más específicos en base a esto. La detección automática de

temáticas (grupos) puede ser particularmente útil cuando no existe una asociación de manera explícita entre los documentos o, si existe, no sería la mejor forma de representación de los perfiles desde el punto de vista del filtrado o recomendación puesto que, las temáticas que deberían estar separadas, si existe una asociación impuesta, quedarían agrupadas juntas en el mismo subperfil. Por ejemplo, si se considera, desde el punto de vista político, una comisión parlamentaria que se crea con carácter político para cubrir simultáneamente y de forma asociada las áreas de agricultura, ganadería y pesca, las intervenciones de los Miembros del Parlamento que participen en ella quedarán recogidas en el mismo subperfil incluso cuando es obvio que se representan temáticas distintas desde el punto de vista objetivo. Además, las estructuras de las comisiones parlamentarias no son estáticas y sufren mutaciones en el tiempo y, por lo tanto, agrupar las intervenciones de un MP en torno a cómo se construyen las comisiones va a generar una distribución de temáticas que va a ser dependiente de cómo se organicen las comisiones políticamente. Finalmente, otro de los problemas que se plantean es el de inicio frío al comienzo de cada legislatura, donde las comisiones no se han creado y por lo tanto no existe una asociación explícita en las intervenciones que se generan en ese periodo de tiempo inicial.

Con el objetivo de lidiar con estos problemas a la hora de construir perfiles de expertos más precisos, en este capítulo se muestra como las técnicas no supervisadas de clustering contribuyen a descubrir temáticas ocultas en documentos y así construir perfiles compuestos para representar los intereses de un usuario. Los resultados de la experimentación llevada a cabo para argumentar tal propósito muestran además como las técnicas de clustering pueden ser aplicadas con éxito en los problemas de filtrado y recomendación para construir perfiles multifacéticos, donde cada subperfil se obtiene a partir de documentos que son relevantes para el usuario que se agrupan de manera conjunta. Estos dos problemas, el filtrado y la recomendación, se puede resolver a la vez desde una perspectiva unificada en ambos contextos, dada una consulta, el resultado del ranking obtenido por el sistema donde se listan los usuarios expertos puede ser utilizado tanto para recomendarles un documento específico como para recomendar al experto en sí. También se ha investigado sobre dos aproximaciones distintas de aplicar clustering al conjunto de documentos: un enfoque global, donde se lleva a cabo el proceso de clustering sobre todas las intervenciones de todos expertos y, por otro lado, una aproximación local, donde el clustering solo se realiza sobre las intervenciones propias de un único experto.

5.2 Contexto del estudio

Dado que el contexto de este estudio combina la construcción y uso de perfiles de usuario para el acceso a la información y, por otro lado, la aplicación de métodos de clustering para conseguir perfiles más precisos y organizados, en esta sección se va a llevar a cabo la presentación de una serie de conceptos relacionados con la temática que se aborda y la combinación de ambas.

5.2.1 Perfilado de usuarios

Un perfil se puede definir como la representación de un modelo de usuario en el cual se almacena la información básica del usuario en cuestión, ya sea su nombre, género, localización, etc. o bien información más compleja como sus áreas de conocimiento, intereses personales, habilidades, etc. El proceso de aprendizaje de un perfil se conoce por el nombre de perfilado de usuarios y consiste básicamente de la recolección de información explícita donde el usuario expresa abiertamente sus intereses o preferencias [50] o, por otro lado recolectando información de forma implícita, donde en este caso es el sistema el que se encarga de detectar de forma automática los elementos de información que pueden ser de relevancia al usuario a partir de analizar sus datos.

Este caso de estudio se centra de forma específica en perfiles que expresan principalmente intereses de los usuarios y por lo tanto, es necesario un método adecuado para representar dichos intereses de forma eficiente y efectiva. Gauch et al. en su artículo [50] consideraban que los perfiles se pueden representar generalmente a partir de palabras clave, redes semánticas o conceptos. Además, las técnicas inteligentes basadas en Aprendizaje Automático y Minería de Datos también son comúnmente utilizadas en la representación de perfiles de usuarios [134]. Prestando atención de forma específica a los perfiles basados en palabras clave, estos se componen de una lista de términos relevantes que se extraen de las fuentes de información utilizadas para construirlos, como pueden ser documentos textuales, páginas web, descripciones de elementos de cualquier tipo, etc. Estas palabras clave o términos descriptivos del perfil del usuario tienen asociado un peso con el objetivo de reflejar la importancia de ese término para el usuario, por ejemplo, usando un esquema de pesos utilizando la medida TF-IDF [91]. Además, los intereses de un usuario se pueden modelar en cierto modo como conceptos abstractos en lugar de palabras clave y, de esta forma, representaciones más elaboradas de perfiles de usuario combinan para construirlos elementos diferentes, por ejemplo temáticas y palabras clave. Aunque los perfiles basados en conocimiento se pueden obtener como una representación legible desde el punto de vista humano sobre los intereses de un usuario, por lo general , no son muy útiles en lo que a ser aplicados a problemas de filtrado y recomendación se refiere, particularmente cuando estos perfiles se construyen a partir de documentos que contienen información en formato de discursos o discusiones orales.

Los perfiles se pueden considerar como herramientas básicas para la adaptación de las características que definen a un usuario en múltiples áreas de las ciencias de la computación [49] y de forma más específica [134], para indicar los diferentes dominios relacionados con el acceso a la información. Teniendo en cuenta el contexto de este estudio en lo relativo al perfilado de usuarios, este se aplica además dentro del dominio de la Recuperación de Información [51], los Sistemas de Recomendación [15] y la Búsqueda de Expertos [81].

5.2.2 Clustering de textos

Desde un punto de vista general, el objetivo principal del análisis de clusters es el de pretender encontrar una estructura común sobre las instancias de un conjunto de datos no etiquetados con el propósito de dividirlas en grupos disjuntos con características similares [72]. Cuando los datos sobre los que se aplica alguna técnica de clustering son representaciones textuales de información, este proceso recibe el nombre de Clustering de Documentos. La primera vez que esta técnica de Aprendizaje Automático fue utilizada en la historia de la Recuperación de Información fue hace 40 años, con la pretensión de mejorar la eficiencia de los procesos de recuperación, originando así los Modelos de Recuperación basados en Clusters [62]. Una vez se agrupan y los documentos relacionados se sitúan juntos en un mismo grupo, dada una consulta, esta se lanza contra los documentos que conforman cada uno de los cluster y, a continuación el sistema devuelve los documentos del cluster que se ajustan mejor a la consulta realizada. Fundamentalmente, la suposición bajo la que se aplica este modelo de recuperación basado en clusters es la hipótesis de grupos; "los documentos asociados más cercanos tienden a ser relevantes para una misma consulta" [120].

En la Figura 5.1 se muestra el proceso seguido de forma general con respecto a la aplicación de técnicas de clustering en el campo de la Recuperación de Información. Dada una colección documental, donde se va a llevar a cabo el proceso de clustering, el primer paso es el de pre-procesamiento, el cual consiste en la eliminación de términos carentes de significado semántico y stemming para eliminar los sufijos de los términos y dejar las palabras en su raíz semántica. En el siguiente paso se dispone a una reducción de las características del conjunto de datos puesto que en este estudio se aborda un problema de alta dimensionalidad [156]. A continuación, en el siguiente paso del proceso se calcula la matriz de documentos/términos, donde las filas de la matriz ser corresponden con las instancias del problema, es decir los documentos, y las columnas responden a cada uno de los términos de los que están compuestos los documentos. De esta forma, en cada celda de esta matriz se representará un peso con la importancia que tiene un término específico dentro de un documento particular. Esta matriz, la cual suele ser significativamente dispersa en lo referente a que la mayoría de sus entradas tienen el valor 0, será la que sirva como conjunto de datos de entrada para el algoritmo de clustering pertinente además de un valor entero que representa el número de grupos que se quieren extraer y, como salida, el algoritmo devolverá el conjunto de datos de entrada separado de forma disjunta entre el número de grupos especificado. Estos grupos obtenidos se puede utilizar en diversas tareas dentro del ámbito de la Recuperación de Información [141], por ejemplo, recuperación de documentos, acceso y organización de documentos, resumen de textos, etc.

De todas las técnicas existentes de clustering [124, 137] hay dos familias principales que caben destacarse. La primera de estas familias responde al nombre de clustering basado en conectividad o más comúnmente conocido como clustering jerárquico [61, 72, 126, 161]. Este tipo de técnicas de detección de grupos generan un árbol de distancias llamado dendrograma,

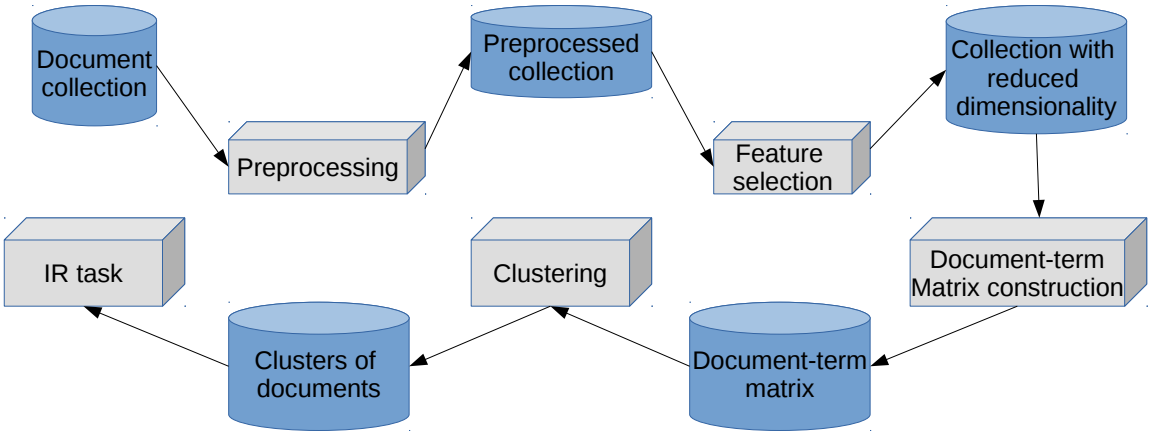

Figura 5.1: Pasos en el proceso de clustering de textos.

el cual representa a grandes rasgos que elementos pertenecientes a una misma rama son más similares entre sí que con respecto a los elementos alojados en ramas distintas. Esta primera familia de algoritmos está a su vez dividida en dos diferentes categorías en función de cómo se construye en dendrograma: la aproximación aglomerativa [129], donde cada elemento pertenece en primera instancia a un cluster independiente y se van fusionando entre sí en función del grado de similitud que exista entre ellos de forma recursiva. tal y como lo hace el algoritmo de anidado aglomerativo (AGNES, [72]), y por otro lado, la aproximación divisiva [64], que al contrario de la anterior todos los elementos del conjunto de datos comienzan formando parte de un único cluster el cual se va dividiendo en subgrupos en función de cómo de distintos sean los elementos y, como representante de esta aproximación está el algoritmo de análisis de clustering divisivo (DIANA, [72]).

En la segunda familia se recogen los algoritmos de clustering basados en centroides. En este tipo de algoritmos los diferentes clusters se agrupan en torno a un punto medio, el cual no tiene porque ser necesariamente una instancia del conjunto de datos, y cada elemento se asigna al cluster cuyo punto medio le sea más cercano [8, 94]. Centrando el estudio en dos métodos diferentes para poder comparar el comportamiento de los datos en distintas aproximaciones se propone utilizar el algoritmo clásico de clustering K-means [89, 153] el cual basa su funcionamiento en dividir n instancias en k grupos diferentes de forma aleatoria e ir asignando de forma iterativa cada instancia al grupo cuya media le sea más cercana y recalculando en sucesivas iteraciones la posición del centro del grupo. Por otro lado, el algoritmo PAM (partitioning around medoids) [72, 103] funciona de una forma muy similar al anterior pero con la particularidad de que los puntos en torno a los que se agrupan los elementos no son un punto medio sino un propio elemento representante que constituye la mediana de todos los datos del grupo.

Como valor añadido al estudio y al margen de las técnicas clásicas de detección no super-

visada de grupos, en la literatura se han encontrado otras técnicas de diferente índole que, no estando consideradas exactamente como algoritmos de clustering, pueden intentar capturar la semántica subyacente de los datos y por lo tanto ser adaptados de tal forma para ser aplicados a este problema. Como primer ejemplo, dentro del contexto de las colecciones de documentos textuales, Latent Dirichlet Allocation (LDA) [14, 18], es un algoritmo que se usa principalmente en el contexto del procesamiento del lenguaje natural (NPL). LDA es un modelo jerárquico Bayesiano de tres niveles el cual es capaz de encontrar las temáticas latentes dentro de una colección de documentos y asigna una distribución de probabilidad de estas temáticas dentro de cada documento y a su vez una distribución de probabilidad de los términos de la colección en cada temática. Otro ejemplo alternativo a las técnicas clásicas de clustering son los Mapas Auto Organizativos (SOM), los cuales son una herramienta efectiva para la visualización e interpretación de espacios de alta dimensionalidad reduciendo el espacio de los datos a un mapa de generalmente dos dimensiones. SOM implementa una red neuronal artificial que se entrena con un conjunto de datos no supervisados con el objetivo de condensar toda la información del conjunto de entrenamiento de forma que, la estructura topológica de los datos y las relaciones entre ellos se preservan de manera intacta, creando así una especie de abstracción del espacio de entrada [74].

Por otro lado, en el contexto del análisis de clustering de datos, uno de los mayores retos que se presentan es el de establecer el número de clusters a priori y la forma de calcularlo. Existen diversas formas para estimar dicho parámetro de manera que se ajuste mejor al conjunto de datos de entrada y, en muchos casos un simple estudio de los datos puede servir para vislumbrar de forma natural el valor del número de clusters que se necesita extraer. Sin embargo, tal y como ocurre en este caso de estudio no se dispone de ninguna pista ni objetiva ni subjetiva que pueda conducir a una correcta estimación de ese parámetro. De esta forma, buscando en la literatura una alternativa genérica, se encontraron diversas formas válidas de estimar este parámetro en base a las características de los datos. La primera, la cual es específica para colecciones documentales, se basa en considerar n como el número total de documentos, m como el número total de términos distintos y t como el número de entradas distintas de cero en la respectiva matriz de documentos/términos y, por lo tanto, el número de clusters vendría definido como $k = mn/t$ [23]. Finalmente, otra aproximación alternativa de determinar el valor de este parámetro es a partir de calcularlo con el método general y efectivo $\sqrt{n/2}$ [72].

Con respecto a la evaluación de la calidad de los procesos de clustering que se han llevado a cabo, se han utilizado medidas clásicas para maximizar la similaridad de los documentos dentro de un mismo cluster y al mismo tiempo minimizar la similaridad de los documentos de un cluster con respecto al resto. En este caso específico se ha utilizado una medida habitual en la literatura como es el índice de Silhouette [140], el cual calcula la distancia media de un elemento dado con el resto de elementos del cluster más cercano y le resta la distancia media del mismo elemento con respecto a los elementos de su mismo cluster y, finalmente se realiza un promedio para

todos los elementos del conjunto. Otro ejemplo de medida de calidad de un cluster es el índice de Davies-Bouldin [38], el cual representa el cociente entre las distancias dentro de un mismo cluster y las distancias entre el resto de clusters y, de la misma forma que en el caso anterior se hace un promedio para cada uno de los elementos del conjunto. Estas medidas cuyo principal objetivo es el de identificar como de compacto es un cluster específico y como de dispersos son los clusters entre sí se conocen como medidas de validación interna puesto que, se calculan solo y exclusivamente con la información que aporta el conjunto de datos y el clustering resultante tras aplicar el algoritmo de detección de grupos. Otra alternativa a esta es la de realizar una evaluación con medidas de validación externas las cuales dependen en gran medida del dominio de aplicación, es decir, en aquellos casos en los que el proceso de clustering es solo una parte del sistema que se construye y, en efecto, es necesario evaluar de alguna forma como el análisis de grupos afecta al comportamiento del sistema de forma general [36]. Siendo este último tipo de sistemas el que se trata en este caso de estudio, la validez de la calidad de los clusters se va a llevar a cabo de forma indirecta teniendo en cuenta también la calidad de las recomendaciones obtenidas usando medidas estandarizadas en el contexto de la Recuperación de Información.

5.3 Construcción de perfiles multifacéticos agrupando documentos

Tal y como se ha mencionado anteriormente en la sección de introducción a este capítulo, dado que un usuario del sistema que se pretende modelar puede estar interesado en varias temáticas distintas al mismo tiempo y, por lo tanto su perfil está compuesto por un conjunto de conceptos o temáticas que están representadas con términos con un peso asociado, se puede concluir que un perfil multifacético es aquel en el que se trata de capturar los distintos aspectos que contienen el conjunto de los documentos asociados a un usuario. En este caso de estudio, cada aspecto, concepto o faceta que se pueda extraer mediante la aplicación de técnicas de detección de grupos se va a considerar un subperfil del propio usuario y va a ser denominado como tal. De esta forma, los perfiles multifacéticos son, de forma alternativa, una representación intermedia a las utilizadas anteriormente donde se unía toda la información textual de un usuario en un mismo documento en el caso de la aproximación monolítica y en contrapartida, en la aproximación de perfiles por intervenciones, toda la información del usuario quedaba completamente separada en documentos independientes.

En la mayoría de las situaciones, las temáticas se encuentran de manera oculta puesto que pueden estar reflejadas de forma implícita en una colección de documentos y esto significa que es necesario diseñar un proceso que extraiga de forma automática y que sea capaz de aprender a partir de esta. En este caso de estudio, este proceso se va a llevar a cabo aplicando técnicas de análisis de clusters donde la principal idea es la de agrupar una colección de documentos independientes para obtener un total de k clusters de documentos. Para este propósito, los

distintos algoritmos de clustering que se van a utilizar tienen como entrada una matriz donde las filas se corresponden con cada uno de los documentos de la colección y las columnas representan todos y cada uno de los términos del vocabulario que existen en la colección de documentos, de esta forma los valores de cada celda representan la importancia de un término específico en un documento concreto tomando el valor 0 en el caso de no encontrarse ninguna ocurrencia de un término en el documento. Además, el número de clusters que se necesita extraer se introduce como un parámetro más en el algoritmo de clustering y serán estimados tal y como se explica más adelante en la sección de evaluación. Como salida, los algoritmos devuelven un conjunto de clusters donde, en el interior de cada uno de ellos, existe una gran similaridad entre todos los documentos que lo componen, es decir, todos los documentos que pertenecen a un mismo cluster tratan sobre una misma temática o concepto específico y, al mismo tiempo se presenta una baja similaridad entre los documentos de clusters distintos. Se puede asumir de esta forma que cada grupo representa un concepto distinto y que, reuniendo todos los términos de los documentos que conforman un mismo cluster se puede obtener una representación de la temática como una lista de términos con un peso o valor de importancia asociado.

Con el propósito de crear perfiles de usuario basados en la información contenida en los documentos que tienen asociados, se han considerado dos aproximaciones distintas para agrupar sus respectivos documentos. La primera de ellas es una aproximación local la cual consiste en encontrar grupos de documentos para cada usuario de forma independiente solo tomando en consideración sus propios documentos. Por otro lado, la otra aproximación que se plantea tiene un enfoque global puesto que lleva a cabo el proceso de clustering con todos los documentos de todo el conjunto de usuarios y posteriormente extrayendo los documentos de cada cluster para cada MP con el objetivo de construir los correspondientes subperfiles. En la primera aproximación local se capturan las temáticas de un usuario de forma específica mientras que, en la segunda aproximación de carácter global, se pretenden encontrar las temáticas comunes que generalmente comparten todos los usuarios. Esto significa que, en el clustering local, los grupos que se extraen tras el proceso son exclusivos de cada usuario y por lo tanto, solo y exclusivamente van a contener información del propio usuario. Por otro lado, en el clustering global, los clusters van a contener información de los documentos de distintos usuarios.

En la Figura 5.2 se muestran a modo de ejemplo como funcionan las aproximaciones que aquí se plantean. Comenzando por la izquierda, el primer esquema representa la distribución de los documentos de un usuario X y cómo se agrupan en tres cluster distintos con documentos similares de forma local y, como resultado de esa agregación, se obtienen tres subperfiles independientes que conformarán el perfil final del usuario. En el esquema central de la Figura 5.2 se muestra el enfoque de clustering global para todos los usuarios, en este caso X, Y y Z, y los hipotéticos grupos que se han detectado para estos usuarios. Se puede observar que los clusters c_2, c_3, c_5 y c_6 son heterogéneos en el sentido en el que integran documentos de usuarios distintos. Si se observa de nuevo al usuario X, el número de subperfiles que se construyen a partir de esta aproximación

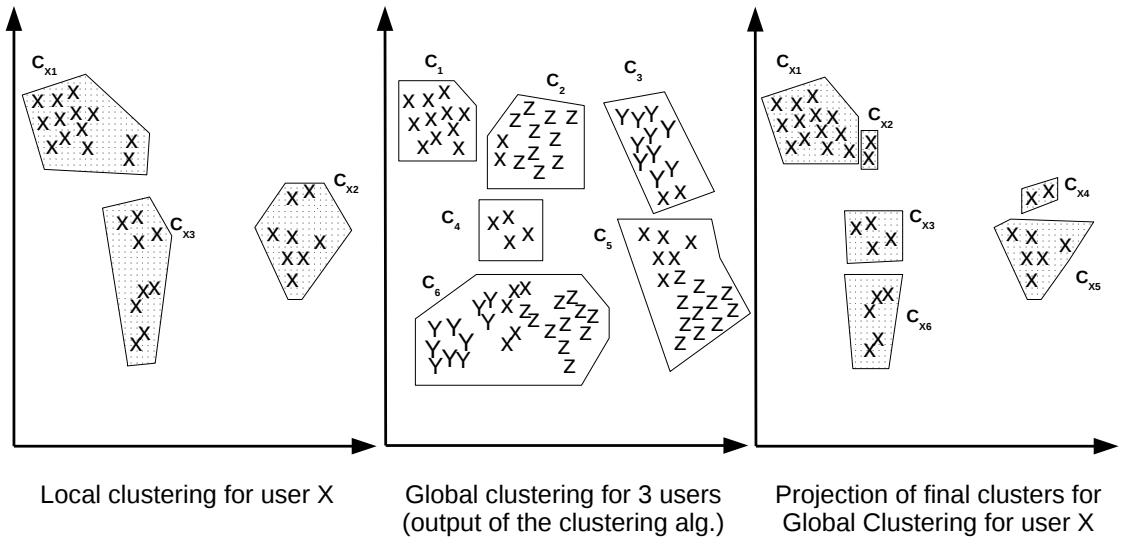

Figura 5.2: Esquema de clustering local y global

global va a depender del número de cluster al que pertenece en este enfoque, es decir, en el esquema de la derecha se puede observar que el número total de clusters considerados para el usuario X asciende a seis, y por lo tanto seis serán los diferentes subperfiles asociados que se construirán para dicho usuario X.

En ambos casos, la salida del proceso no supervisado de clustering va a ser una asociación de cada documento de un usuario específico a un cluster concreto. De esta forma, para cada usuario y para un cluster dado, se crea un "macro documento" donde se reúnen todos los documentos pertenecientes a un mismo cluster y, cada uno de los documentos generados por este proceso, va a ser considerado como un subperfil. Dicho esto, a partir de este proceso se obtiene una nueva colección de documentos que va a contener todos los documentos de subperfiles de todo el conjunto de usuarios, la cual pasa a ser indexada de forma normal por el correspondiente Sistema de Recuperación de Información. Así, cuando una consulta se lanza contra el sistema, esta devuelve una lista en forma de ranking con los diferentes subperfiles (uno o más de uno) de los usuarios que sean más afines a la consulta. Como este problema se trata desde el enfoque de la recomendación de expertos, el ranking final debe estar compuesto por usuarios únicos, por lo tanto es necesario establecer una estrategia de fusión de los distintos subperfiles que pueden aparecer en el ranking original y obtener un valor de score final para cada uno de los usuarios.

5.4 Evaluación experimental

Aunque en este capítulo se aborda el problema de la búsqueda de expertos y recomendación de documentos de forma general, la evaluación se va a realizar de forma concreta, tal y como se

viene realizando en los capítulos anteriores, en un contexto parlamentario. El objetivo básico de este caso de estudio es el de encontrar Miembros del Parlamento que sean relevantes para una consulta que puede ser formulada por un ciudadano (recomendación) o bien, el objetivo de determinar qué Miembros del Parlamento pueden estar interesados en recibir un documento nuevo que llega al sistema (filtrado). Con el objetivo de realizar esta labor, se ha optado por una representación de los intereses de los Miembros del Parlamento en la forma de perfiles que se construyen en base a sus intervenciones en los debates públicos de las sesiones parlamentarias. Más específicamente, se tiene en consideración que un Miembro del Parlamento puede ser integrante al mismo tiempo en varias comisiones parlamentarias distintas y que estas son más reducidas en cuestión del número de MPs que las componen y que además cubren temáticas más específicas. De este modo, teniendo en cuenta que un Miembro del Parlamento puede estar interesado en múltiples aspectos políticos (por ejemplo agricultura, economía y medio ambiente) que pueden no estar relacionados entre sí de forma explicita, el objetivo es crear subperfiles para cada Miembro del Parlamento para representar sus respectivos intereses en las diferentes temáticas.

De forma general, el objetivo de este proceso de evaluación es el determinar si el clustering de textos es una buena herramienta para identificar de forma automática las diferentes temáticas que pueden ser de interés para un usuario y además, discernir la utilidad de la aplicación de estas técnicas para la recomendación de expertos y el filtrado de información. Para conseguir esto se plantean una serie de cuestiones con el objetivo de que se puedan verificar tras el proceso de evaluación, las cuales se volverán a referenciar con el mismo enunciado en el apartado de resultados como recordatorio previo antes de ser resueltas:

- La primera cuestión que se plantea es la de si el clustering de documentos es una técnica apropiada para la extracción de forma automática de las temáticas en las que una persona pueda estar interesada.

- La segunda cuestión propone si la utilización de subperfiles basados en técnicas de clustering beneficia de algún modo las tareas de filtrado y recomendación.

- En la tercera cuestión se presenta la disyuntiva entre si existe alguna diferencia significativa entre las aproximaciones local y global que se proponen.

- En la cuarta cuestión surge el interrogante de si el número de grupos extraídos de la colección influye de forma relevante en el rendimiento de Sistema de Recomendación.

- Y finalmente, se plantea qué algoritmo o algoritmos de clustering se ajustan mejor para la tarea que se aborda en este caso de estudio.

En esta sección se describe, por tanto, el diseño de los experimentos que se han llevado a cabo y los resultados correspondientes obtenidos por cada uno de ellos.

5.4.1 Revisión del Sistema de Filtrado y Recomendación

En primer lugar, el conjunto de datos que se ha utilizado para llevar a cabo el proceso de experimentación han sido los mismos con los que se han llevado a cabo los anteriores experimentos. De la misma forma, el procesamiento de la colección sigue siendo el mismo que se explicó en el Capítulo 3.

Con el objetivo de recomendar un conjunto de MPs a partir de una consulta o un documento que se pretende filtrar, se ha utilizado la herramienta Apache Lucene Library[1], la cual consta con el modelo de recuperación de información conocido en el estado del arte BM25 [67]. Por cada MP_i, el modelo de indexación recibe como entrada un conjunto de los subperfiles del propio MP, por ejemplo, para el MP MP_5 lo documentos que se indexan son los subperfiles generados por tres clusters c_1, c_2 y c_3 (en definitiva hay 3 subperfiles en total, denominados a efectos prácticos como $MP_5\_c_1$, $MP_5\_c_2$ y $MP_5\_c_3$, respectivamente). Los términos contenidos en esos subperfiles son procesados eliminando stop words y reduciéndolos a su raíz semántica mediante la técnica de stemming implementada en Lucene Spanish Analyzer. Además, Cualquier término que aparezca en menos del 1% del total de intervenciones se elimina por ser considerado como irrelevante. Entonces, dada una consulta, el modelo devuelve una lista ordenada de forma decreciente a modo de ranking con los subperfiles de los MPs más afines a la consulta y el respectivo grado de afinidad. No obstante, como el objetivo final de sistema es el de obtener un ranking de MPs de acuerdo a su relevancia con respecto a la consulta, el ranking de subperfiles original se filtra considerando para el ello el método $CombLgDCS$, el cual viene definido en [40]. Este método basa su funcionamiento en calcular un valor de score único para cada MP_i a partir de agregar todos los valores de score de los subperfiles del MP en cuestión pero con una devaluación logarítmica en función de su posición en el ranking. La formula del método $CombLgDCS$ se muestra como:

$$score(MP_i, q) = \sum_{MP_{i\_c_j}} \frac{s(MP_i\_c_j)}{\log_2(rank(MP_i\_c_j)+1)}, \tag{5.1}$$

donde MP_i representa a un MP, $MP_i\_c_j$ es un subperfil del MP en cuestión dentro del ranking, $s(MP_i\_c_j)$ define el respectivo valor de score (similitud entre el perfil del MP y la consulta) y, finalmente $rank(MP_i\_c_j)$ es la posición del subperfil $MP_i\_c_j$ en el ranking. Y una vez los scores de los subperfiles se han agregado y recalculado para cada MP, los MPs se ordenan de nuevo de acuerdo al nuevo valor de score obtenido.

5.4.2 Algoritmos de clustering

Para llevar a cabo el proceso de experimentación se han utilizado las implementaciones de R de los siguientes algoritmos de clustering: AGNES y DIANA como métodos de clustering jerárquicos

[1] https://lucene.apache.org/

aglomerativo y divisivo respectivamente, K-MEANS y PAM como métodos de clustering basado en centroides y, finalmente, los algoritmos Latent Dirichlet Allocation (LDA) como modelo estadístico generativo y Self Organizing Maps (SOM) como modelo basado en redes neuronales. Los algoritmos que se han utilizado para llevar a cabo el proceso de agrupamiento se han seleccionado en base a ser los más utilizados en la literatura del estado del arte.

Estos algoritmos reciben como entrada una matriz donde los filas representan cada uno de los documentos y las columnas definen cada uno de los términos en dichos documentos. Tras el procesamiento, donde se eliminan stop words, caracteres de puntuación y numéricos y se realiza un proceso de stemming para obtener las raíces semánticas, a cada uno de los términos del conjunto resultante se le otorga un peso basado en el esquema TfIdf. De este modo, los documentos quedan representados en la matriz como un vector de términos presentes en la colección de forma que si un término específico aparece en el documento, este será representado por el valor TfIdf y si no aparece en el documento será representado con el valor 0.0.

Con respecto a la aproximación local, dado un MP cualquiera MP_i, el número de instancias a ser agrupadas se corresponde con el número de documentos asociados únicamente al MP MP_i. De esta forma, se repite el proceso de agrupamiento para todos y cada uno de los MPs con el objetivo de obtener un agrupamiento propio e independiente para cada uno de ellos. Por el contrario, en la aproximación global, el número de instancias a agrupar se corresponde con el número de todos los documentos de todos los MPs y por lo tanto, en esta aproximación el proceso de agrupamiento solo se ejecuta una única vez, obteniendo así un conjunto de clusters para cada MP.

Tanto en los métodos basados en centroides como en los métodos basados en agrupamiento jerárquico se utiliza la medida de similitud coseno para calcular la distancia entre los diferentes individuos. Con respecto al algoritmo LDA [14], adaptando su uso para hacer las veces de algoritmo de agrupamiento, una vez que se ha detectado la distribución de las temáticas para todos los documentos, cada uno de ellos se agrupa en torno al temática más probable, formando así un cluster. En referencia al algoritmo basado en redes neuronales SOM, se puede adaptar su uso de forma en la que se puedan agrupar instancias similares juntas. Así, una vez ejecutado el algoritmo SOM y obtenida como salida una estructura donde cada documento se corresponde con una neurona, un conjunto de vectores de pesos definen la posición de la neurona dentro del espacio de datos discretizado y, finalmente, son esos vectores los que se agrupan en función de su similaridad usando un algoritmo de clustering, creando así clusters de instancias similares contenidas en cada una de las neuronas que se han agrupado juntas. En este caso de estudio se ha usado SOM en combinación con el algoritmo K-MEANS para agrupar las neuronas resultantes (SOM-KM) tal y cómo se especifica en la literatura para tareas generales de agrupamiento [68, 106, 107].

5.4.3 Selección del número de clusters

Tal y como se ha mencionado anteriormente, el número de clusters, k, en el que se requiere agrupar un conjunto de datos es un problema generalizado en este tipo de problemas, más aún cuando para cada uno de los MPs no se puede establecer un único valor de k genérico, ni se puede analizar de forma independiente cada uno de ellos para intentar hallar el valor óptimo. Por lo tanto, la situación ideal es la de llevar a cabo una selección automática, basándose en aspectos objetivos y tangibles del propio MP, para obtener el mejor valor posible de grupos para cada MP, lo cual no es una tarea trivial.

En el proceso de experimentación, se han tenido en cuenta diversas aproximaciones, donde k se ha fijado o se ha calculado automáticamente teniendo en cuenta algunos datos propios relacionados con la colección. De forma más específica, se han llevado a cabo los experimentos planteando las siguiente alternativas:

- $k = \#Com \Rightarrow$ Para la aproximación global, este valor representa el número total de comisiones constituidas en la octava legislatura del Parlamento de Andalucía, en este caso $k = 26$. En cambio, para la aproximación local, el valor representa específicamente el número de comisiones en las que ha participado un MP en cuestión donde, de media, un MP participa en 6.02 con una desviación estándar de 4.52. El objetivo de usar este criterio para establecer el valor de k es el de determinar el poder que tiene el algoritmo de clustering de reproducir los conjuntos de iniciativas parlamentarias dentro de las diferentes comisiones parlamentarias en clusters independientes.

- $k = m * n/t \Rightarrow$ Donde m = número total de términos en la colección de documentos; n = número de intervenciones totales de todos los MPs en la colección; y t = número de entradas en la matriz documentos/términos distintas de cero. Este criterio para establecer el valor de k es aplicable de la misma forma para ambas aproximaciones de clustering, global y local, no obstante, los valores de m, n y t dependerán del tipo de clustering. En el caso del clustering global, $m = 4208$; el número total de intervenciones es de $n = 10025$ y por último $t = 1702296$. En el otro caso, el clustering local, los valores de m, n y t varían en función de las intervenciones y los términos que utilice cada MP, pero de media, $m = 3427.45 \pm 2056.15$, $n = 58.11 \pm 58.55$ y $t = 12106.66 \pm 12064.64$. El valor final para k para la aproximación global es de $k = 24$ y, para la aproximación local, el valor de k medio es 15.85 ± 9.67.

- $k = \sqrt{n/2} \Rightarrow$ Teniendo en cuenta que el valor de n se corresponde, tal como en el criterio anterior, con el número de intervenciones, para la aproximación de clustering global el valor de $k = 70$, mientras que para la aproximación local, el valor de k, el cual varía en función de las características específicas de cada MP, es de media 4.25 ± 2.60.

5.4.4 Contexto experimental

Tal y como se especifica en los procesos de experimentación anteriores, el conjunto de iniciativas se particiona de forma aleatoria en dos subconjuntos; un conjunto de entrenamiento (80%) y un conjunto de test (20%). Por lo tanto, el conjunto de entrenamiento se utiliza para construir los subperfiles asociados a cada MP a partir de los clusters obtenidos y el propósito del conjunto de test es meramente el de evaluar la calidad de las diferentes propuestas. Este proceso, se repite cinco veces, generando así cinco particiones, y los resultados generales que se muestran se corresponden con la media de los resultados obtenidos en cada una de las particiones.

En el proceso de filtrado de documentos, las consultas al sistema se realizan a partir del contenido de la iniciativa, es decir el texto completo, en este caso, el objetivo es el de distribuir una iniciativa a cualquier MP que pueda tener interés en recibirla. Por otra parte con respecto al proceso de recomendación de MPs, se utilizan como consultas solo los extractos de las iniciativas con el objetivo de encontrar un MP específico el cual se debe corresponder con el MP más alto en el ranking. En ambos casos y desde el punto de vista de los juicios de relevancia, puesto que el objetivo principal es el de encontrar a los MPs que puedan estar en cierto grado relacionados o interesados en una temática, se asume que por cada consulta al sistema solo se van a considerar como válidos aquellos MPs que hayan participado de forma activa en la iniciativa en cuestión que se haya utilizado como consulta. Aunque cobra sentido el suponer que una iniciativa va a ser relevante o de interés para ciertos MPs aún no habiendo participado de forma activa en ella, se ha tomado esta decisión para realizar los experimentos de la forma más conservativa posible incluso cuando puede penalizar los resultados, y de forma particular en la aproximación de filtrado.

Tal y como se plantea a grandes rasgos, llegada una consulta nueva al sistema, el motor de búsqueda de este devuelve una lista ordenada de MPs en orden decreciente, la cual representa el grado de similitud entre los MPs y el texto de la consulta. Por tanto, con el objetivo de medir la calidad de los rankings devueltos por el sistema, se van a utilizar como medidas las comúnmente conocidas en el ámbito de la Recuperación de Información *precision* y *recall*, solo tomando los 10 primeros elementos del ranking de MPs (p@10 y r@10, respectivamente). Además se va a considerar una vez más la métrica Normalized Discounted Cumulative Gain [63] (ndcg@10), con el objetivo de terner en consideración la posición en el ranking de los MPs relevantes.

Una vez establecidas las bases de la experimentación, para establecer si la aplicación de técnicas de aprendizaje automático funcionan correctamente para aprender subperfiles y construir un perfil que defina a un MP, se ha tomado la decisión de comparar los resultados con casos basales distintos entre los que se citan:

- Un perfil único e independiente para cada MP (perfil monolítico). Es decir, se construye un único documento donde se comprenden todas y cada una de las intervenciones que el MP haya podido realizar en las distintas iniciativas. En consecuencia, este perfil va a contener todas las temáticas en las que un MP pueda estar interesado. Se puede considerar que

este sería el caso donde el valor de k sea igual a 1, es decir, un único cluster con todas las instancias.

- Se construyen un número de subperfiles para cada MP de acuerdo con las comisiones en las que haya participado (subperfiles basado en comisiones). De esta forma, cada MP va a disponer de varios subperfiles compuestos cada uno de ellos de las intervenciones que haya realizado dentro de una comisión específica. Desde un punto de vista práctico, si un MP ha participado en k comisiones distintas, este debería tener k subperfiles distintos.

- Como propuesta extrema contraria al perfil monolítico, otro caso base sería el considerar cada intervención como un subperfil distinto (subperfiles basado en intervenciones). En este caso el número de subperfiles independientes asociados a un MP va a ser igual al número de veces que haya intervenido en alguna iniciativa parlamentaria. Por lo tanto, el valor de k será igual al número de documentos n de un MP.

El objetivo tras el planteamiento de estos tres casos base es el de comparar las aproximaciones que en esta sección se plantean con los casos extremos donde $k = 1$ y $k = n$ para intentar discernir si la aplicación de técnicas de agrupamiento obtienen mejores resultados. Además, se plantea un caso intermedio con la pretensión de ser utilizado como guía ya que se usa el conocimiento privilegiado de conocer en cuántas comisiones ha participado un MP. En definitiva, la situación que se espera alcanzar es que las aproximaciones de recomendación y filtrado de MPs tengan un mayor rendimiento con la construcción de subperfiles basado en clustering que el obtenido con las aproximaciones basales.

5.4.5 Resultados

Tras el proceso de experimentación de las diversas aproximaciones propuestas, se pueden dar respuestas a las diferentes hipótesis enumeradas en la Sección 5.4, donde la primera de ellas enunciaba si la aplicación de técnicas de clustering a colecciones de documentos es una técnica apropiada para la extracción de temáticas en las que puede estar interesado un individuo. Para dar respuesta a este planteamiento, en primer lugar se debe mostrar cómo los clusters se ajustan a los temas tratados en las comisiones parlamentarias realizando un análisis, desde el punto de vista cualitativo, para un MP específico y otro análisis cuantitativo desde un punto de vista más general, es decir, centrado en las comisiones.

- **La primera cuestión que se plantea es la de si el clustering de documentos es una técnica apropiada para la extracción de forma automática de las temáticas en las que una persona pueda estar interesada.**

En el análisis cuantitativo individual se ha realizado sobre un MP específico del grupo parlamentario de `Izquierda Unida`. Se ha seleccionado este MP de entre todo el conjunto puesto

que es uno de los que registra un mayor número de intervenciones parlamentarias (en la octava legislatura intervino hasta en 172 iniciativas parlamentarias distintas) y a su vez, tratando un amplio abanico de temáticas distintas (además de constar con 97 intervenciones correspondientes a sesiones plenarias participó en 14 comisiones especializadas y grupos de trabajo con un total de 75 intervenciones). Por tanto, ¿en qué temáticas está este MP verdaderamente interesado?. Se puede establecer que las temáticas en las que este MP está interesado están relacionadas con las comisiones en las que participa, pero a su vez, es lógico pensar que algunas temáticas tienen un peso mayor que otras desde el punto de vista del interés del MP. Para representar este caso, la Tabla 5.1 en su segunda columna muestra el tamaño, en las diferentes comisiones, en base al porcentaje de términos que tienen sus intervenciones (la mitad del porcentaje está dedicado a las sesiones plenarias, donde se pueden debatir temáticas muy diversas). A partir de estos datos se puede observar que la principal área que este MP trata en sus intervenciones es la relacionada con la comisión de Igualdad y Bienestar Social, la comisión de Cultura y la comisión de Sanidad (representando el 70% de las intervenciones que el MP realiza en comisiones especializas sin considerar las sesiones plenarias).

A partir de esta información, se consideran en primer lugar dos de las aproximaciones basales donde no se utilizan técnicas de clustering, en concreto, el perfil monolítico y los subperfiles basados en comisiones. Con respecto al perfil monolítico se puede observar que los términos relativos a los procedimientos parlamentarios distorsionan las temáticas en las que el MP pueda estar interesado haciendo más difícil la labor de detectarlas, tal y como se muestra en la nube de palabras a la izquierda de la Figura 5.3. Por el contrario, considerando la aproximación de subperfiles basados en comisiones, se puede observar en la nube de palabras de la Figura 5.3 a la derecha, la cual se ha obtenido de la comisión de Igualdad y Bienestar Social, que los términos relacionados con la comisión predominan en el cluster, aún así los términos más comunes a la jerga parlamentaria siguen presentes pero en una menor medida. Centrando la vista ahora en el gran número de intervenciones del MP en las sesiones plenarias, en principio no se dispone de ninguna asociación de documentos a una temática, así que todos los documentos pasan a formar parte de un único subperfil, comportándose de esta forma de manera muy similar a las características del perfil monolítico.

A continuación, observando los resultados obtenidos tras la aplicación de las distintas técnicas de clustering, en particular el algoritmo K-MEANS en la aproximación global, siendo el valor de $k = 26$, se aprecia que todas las intervenciones del MP (incluidas las sesiones plenarias) se distribuyen en 14 de entre los 26 clusters candidatos. El tamaño de cada cluster (en base al porcentaje de términos) se muestra esta vez en la tercera columna de la Tabla 5.1. Con el objetivo de identificar la temática predominante en cada cluster, un aproximación lógica es la de observar los términos más comunes en el cluster, es decir aquellos que aportan la mayor contribución, y asignarlos al cluster con la temática que los términos sugieren, pudiéndose producir diversas situaciones:

Figura 5.3: Representación de la nube de palabras para distintos perfiles: el gráfico de la izquierda muestra un perfil monolítico y el de la derecha muestra un perfil basado en comisiones obtenido a partir de la comisión de Igualdad y Bienestar Social.

- Es posible encontrar una relación de 1-a-1 entre los documentos que conforman un cluster y una comisión específica, tal y como se representa en la nube de palabras de la Figura 5.4, a la izquierda, donde las palabras representadas en rojo sugieren que el cluster pertenece a la Comisión de Cultura.

- También, una comisión puede encontrarse dividida en temáticas distintas, 1-a-n. Por ejemplo, el proceso de clustering es capaz de descubrir la temática "violencia de género" como una nueva temática dentro de la comisión de Igualdad y Bienestar Social, tal y como se muestra en la nube de palabras de la derecha de la Figura 5.4. Las intervenciones pertenecientes a esta comisión están relacionadas con la igualdad de género y bienestar social pero el clustering ha sido capaz de determinar dos temáticas claramente diferenciadas "violencia de género" y "bienestar social".

- Se puede producir que dos comisiones diferentes se unan en un único cluster, 2-a-1. Es decir, como en el caso de dos comisiones altamente relacionadas entre sí como pueden ser "Tecnología, Ciencia y Empresa" y "Comercio, Tecnología y Ciencia" donde todas sus intervenciones se recogen en el mismo cluster.

- Descubrimiento de temáticas transversales a todas las comisiones, n-a-1: se ha detectado la existencia de clusters que recogen intervenciones de diversas comisiones distintas, como es el caso de la temática relacionada con asuntos económicos, el cual representa un interés general para todos los MPs y en todas las comisiones (por ejemplo, en la comisión de Sanidad se tratan los presupuestos de los hospitales). Estas temáticas incluyen un amplio

Distribución Real		Clustering
Plenos	0,500	
Comisiones		
Igualdad de Género y Bienestar Social	0,128	0,286
Cultura	0,121	0,151
Sanidad	0,103	0,144
Presidencia	0,046	
Turismo y Empresa	0,018	0,021
Asuntos Europeos	0,015	0,052
Obras Públicas y Vivienda	0,011	0,013
Obras Públicas y Transportes	0,010	0,030
Tecnología, Ciencia y Empresa	0,009	0,063
Comercio, Tecnología y Ciencia	0,009	
Gobernación	0,008	
Justicia	0,008	
Radio y Televisión	0,007	0,016
Medio Ambiente	0,005	
Temática de Economía		0,139
Temática de Violencia de Género		0,066
Temática de Moviento Laboral		0,007
Temática de Educación		0,007
Temática de Juventud		0,006

Tabla 5.1: Distribución (en términos de tamaño de perfil) de las intervenciones de los MPs en la legislatura. La segunda columna muestra la distribución 'verdadera' considerando las sesiones reales del parlamento. La tercera columna muestra la distribución considerando los grupos aprendidos.

número de intervenciones de las sesiones plenarias y de las comisiones especializadas. Esto refleja que la economía es una temática multidisciplinar cuyo uso se comparte en toda la actividad parlamentaria a pesar de no estar definido de forma explícita en una comisión concreta.

- En otros casos, un cluster, desde el punto de vista de la aproximación global, contiene solo un documento del MP en cuestión, lo cual puede ser considerado como una representación de una temática marginal que el MP ha tratado en un momento específico (ultimas tres filas de la Tabla 5.1).

Dicho esto, se puede afirmar que la aplicación de técnicas de clustering es capaz de identificar las temáticas de interés para un MP en concreto, al margen de las comisiones en las que este participe. Además, puede servir para distribuir las intervenciones de las sesiones plenarias entre sus respectivas temáticas.

En el análisis cuantitativo general el objetivo que se propone es el de determinar de forma

Figura 5.4: Representación de la nube de palabras para dos perfiles distintos aprendidos a partir de la aproximación global de K-MEANS.

gráfica el grado en que la salida obtenida por los algoritmos de clustering está cercana a la realidad, la cual en este caso es cómo los clusters se ajustan a las divisiones políticas de las temáticas en comisiones. Al contrario que en el caso anterior donde se observaba a un MP en concreto, en este caso, el análisis se realiza sobre todos los MPs del sistema y todas las comisiones, y más específicamente el análisis se realizará tomando $k = \#Com$, en este caso 26. En las matrices de color que se muestran en las Figuras 5.5 y 5.6 las columnas representan los 26 clusters obtenidos tras el proceso de clustering y cada una de las filas definen una comisión del parlamento. Así, la intensidad de color de cada celda de la matriz representa el porcentaje de documentos de una comisión que se han alojado dentro de un cluster (más intenso, mayor número de documentos). De este modo, como muestra de los resultados, la Figura 5.5 representa la distribución de los documentos de las comisiones en los distintos clusters en el algoritmo K-MEANS y la Figura 5.6 muestra la misma información pero con el algoritmo DIANA (clustering jerárquico divisivo).

Aún cuando la Figura 5.5 muestra el resultado del clustering en la aproximación global de K-MEANS, la aproximación homóloga usando el algoritmo LDA tiene un comportamiento muy similar pudiéndose observar en ambos casos que existe un grado muy alto de emparejamiento entre las comisiones y los clusters resultantes, es decir, la gran mayoría de los documentos de intervenciones de una comisión han ido a parar solo y exclusivamente a un mismo cluster. Tal y como se ha mencionado con anterioridad, en la mayoria de los casos, los clusters son capaces de capturar la esencia de las temáticas que conforman una comisión. En este caso particular, por ejemplo, el $Cluster12$ (prácticamente contiene todos los documentos de la comisión de Sanidad), $Cluster2$ (claramente asociado a la comisión de Cultura) o el $Cluster4$ (correspondiente a la comisión de Educación). Además, la aplicación de técnicas de clustering es capaz de unificar

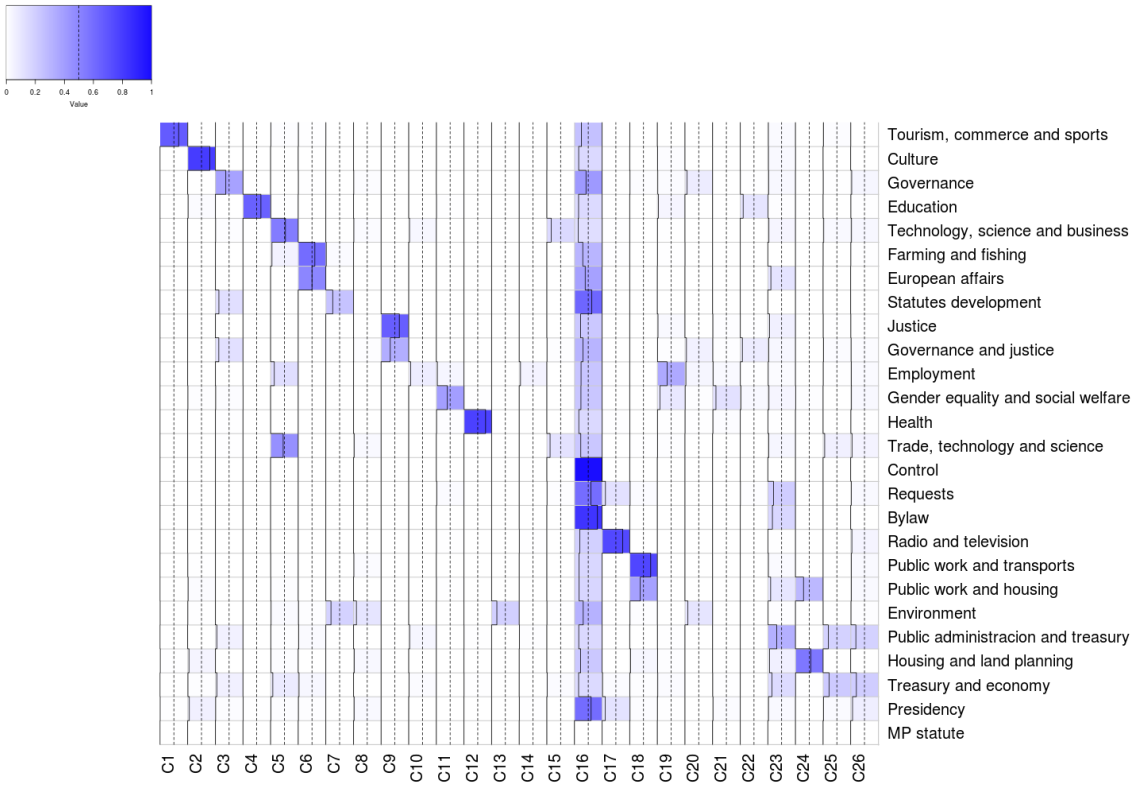

Figura 5.5: Distribución de las comisiones en clusters para la aproximación global de K-MEANS.

documentos de comisiones distintas pero que tratan esencialmente temáticas muy similares. Esta apreciación se puede observar, por ejemplo, entre la comisión de Obras Públicas y Transportes y la comisión de Obras Públicas y Vivienda, la cuales están contenidas juntas en su mayoría dentro del $Cluster18$. En otros casos se produce que el algoritmo de clustering detecta dos temáticas bien diferenciadas dentro de una misma comisión y las divide en varios clusters distintos como, por ejemplo, el $Cluster23$, $Cluster25$ y $Cluster26$ relacionados con asuntos económicos y de la administración pública. Por lo tanto, se puede afirmar que tanto Global K-MEANS y Global LDA capturan las temáticas de las comisiones con una precisión relativamente alta.

Otro patrón que se ha podido identificar en lo que respecta a cómo los clusters se ajustan a las comisiones es el que se presenta en la aproximación global del algoritmo DIANA. En la Figura 5.6 se muestra claramente un comportamiento distinto al de los algoritmos presentados en el párrafo anterior. Si se considera el mismo orden en el que se presentaba el eje de las comisiones para el clustering con el algoritmo K-MEANS y se ajustan los clusters con el objetivo de encontrar una diagonal como en la figura anterior, ocurre que esa diagonal no se presenta, o al menos de una forma tan clara, es decir, con este tipo de clustering los documentos de intervenciones de las comisiones se dividen en múltiples clusters de forma más homogénea. Por ejemplo, en este caso, los documentos de la comisión de Sanidad se distribuyen de forma equitativa entre un total de 10

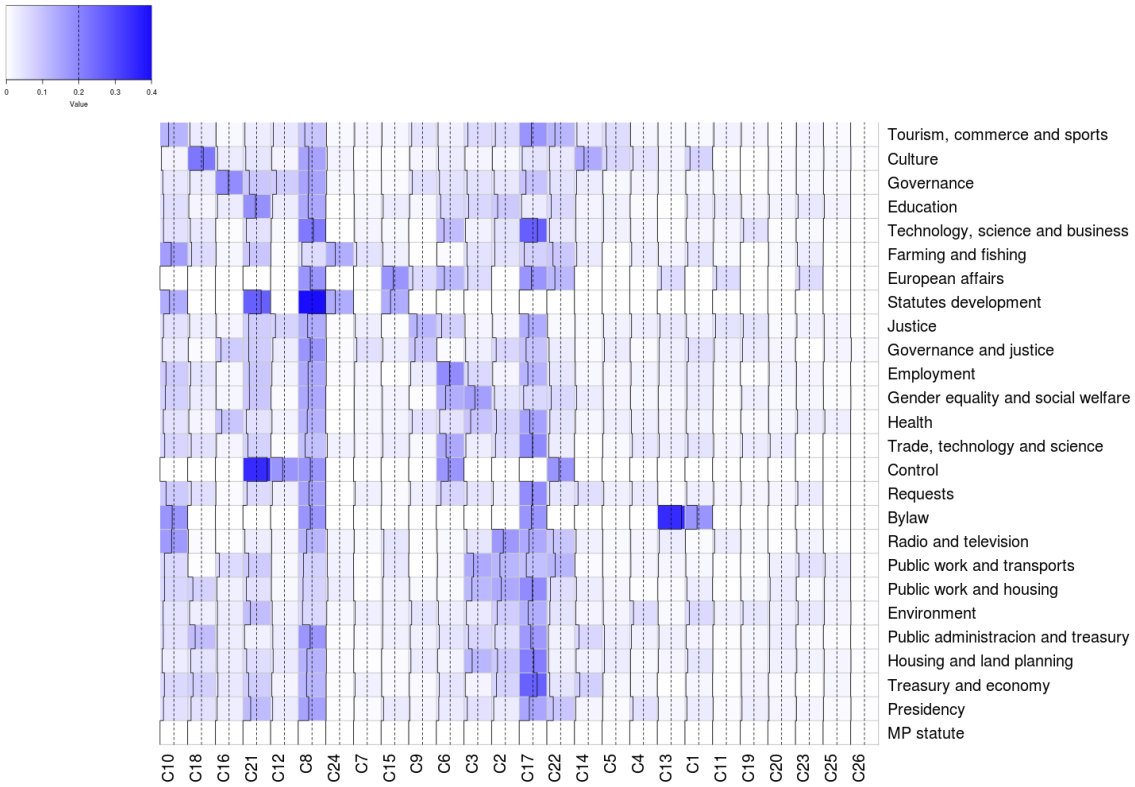

Figura 5.6: Distribución de las comisiones en clusters para la aproximación global de Diana.

clusters distintos al igual que otras muchas más comisiones que en el caso anterior estaban casi perfectamente definidas en un único cluster.

Tras este análisis, se puede concluir que aplicar técnicas de clustering sobre las intervenciones de un parlamento funcionan mejor para detectar distintas temáticas dentro de una misma intervención que las divisiones políticas artificiales en comisiones. Esto, probablemente ocurre puesto que los perfiles resultantes tras el proceso de clustering se comportan mejor que los perfiles generados a partir de las comisiones tal y como se puede ver en las Tablas 5.2 y 5.3.

- **La segunda cuestión propone si la utilización de subperfiles basados en técnicas de clustering beneficia de algún modo las tareas de filtrado y recomendación.**

Otra de las cuestiones que se lanzaban era la de si la utilización de subperfiles basados en técnicas de clustering beneficia de algún modo las tareas de filtrado y recomendación. Para responder a esta cuestión una vez llevado a cabo el proceso de experimentación y evaluación, los resultados se presentan en las Tablas 5.2 y 5.3. La primera de las tablas contiene los resultados para las tareas de filtrado y la segunda a su vez contiene los resultados para las tareas de recomendación. En ambas tablas, la primera columna representa el tipo de clustering ((T) Local o Global), la segunda columna (Alg.) indica el nombre del algoritmo de clustering que se ha

Tabla 5.2: Valores de las medidas de evaluación para los perfiles basados en clusters y los casos base para el filtrado (Etiquetas de las columnas: T = Tipo de clustering; (L)ocal o (G)lobal; k = método para estimar el número de clusters; $\#Clusters$ = número clusters; r@10 = recall 10 primeros; P-r = Posición en el ranking de recall; p@10 = precisión 10 primeros; P-p = Posición en el ranking de precisión; ndcg@10 = NDCG 10 primeros; P-ndcg = Posición en el ranking de NDCG; RRF = valor de Reciprocal Rank Fusion).

T	Alg.	k	#Clusters	r@10	P-r	p@10	P-p	ndcg@10	P-ndcg	RRF
G	AGNES	$\sqrt{n/2}$	70	0.7724	1	0.1779	1	0.6549	2	0.0489
G	AGNES	$\#Com$	26	0.7660	4	0.1754	7	0.6547	3	0.0464
G	PAM	$\sqrt{n/2}$	70	0.7698	3	0.1767	3	0.6391	10	0.0460
G	AGNES	$m*n/t$	24	0.7652	5	0.1752	8	0.6543	4	0.0457
L	DIANA	$\sqrt{n/2}$	4.25 ± 2.60	0.7630	6	0.1766	4	0.6379	12	0.0447
L	AGNES	$\sqrt{n/2}$	4.25 ± 2.60	0.7710	2	0.1775	2	0.6308	22	0.0445
G	DIANA	$\#Com$	26	0.7567	12	0.1744	13	0.6509	5	0.0430
L	KMEANS	$\sqrt{n/2}$	4.25 ± 2.60	0.7567	13	0.1749	9	0.6408	8	0.0429
L	SOM-KM	$\sqrt{n/2}$	4.25 ± 2.60	0.7559	14	0.1754	6	0.6377	14	0.0422
G	PAM	$m*n/t$	24	0.7613	7	0.1748	11	0.6366	16	0.0422
L	LDA	$\sqrt{n/2}$	4.25 ± 2.60	0.7595	10	0.1761	5	0.6303	23	0.0417
G	DIANA	$m*n/t$	24	0.7545	16	0.1739	16	0.6470	6	0.0415
G	PAM	$\#Com$	26	0.7610	8	0.1747	12	0.6353	19	0.0413
G	LDA	$\sqrt{n/2}$	70	0.7551	15	0.1740	15	0.6400	9	0.0412
G	LDA	$\#Com$	26	0.7570	11	0.1742	14	0.6354	18	0.0404
L	AGNES	$\#Com$	6.02 ± 4.52	0.7602	9	0.1748	10	0.6164	33	0.0395
G	DIANA	$\sqrt{n/2}$	70	0.7480	21	0.1723	24	0.6452	7	0.0392
M-Prof				0.7195	35	0.1724	23	0.6577	1	0.0390
G	KMEANS	$\sqrt{n/2}$	70	0.7484	20	0.1729	20	0.6377	13	0.0387
L	AGNES	$m*n/t$	15.85 ± 9.67	0.7476	22	0.1725	22	0.6380	11	0.0385
G	LDA	$m*n/t$	24	0.7507	17	0.1727	21	0.6269	28	0.0367
G	SOM-KM	$\#Com$	26	0.7497	19	0.1730	18	0.6246	31	0.0365
L	PAM	$\sqrt{n/2}$	4.25 ± 2.60	0.7466	24	0.1730	17	0.6277	27	0.0364
L	KMEANS	$m*n/t$	15.85 ± 9.67	0.7421	28	0.1722	26	0.6377	15	0.0363
G	SOM-KM	$\sqrt{n/2}$	70	0.7441	26	0.1720	27	0.6365	17	0.0361
G	SOM-KM	$m*n/t$	24	0.7475	23	0.1730	19	0.6254	30	0.0358
I-SubP				0.7505	18	0.1687	33	0.6283	25	0.0353
L	LDA	$m*n/t$	15.85 ± 9.67	0.7465	25	0.1722	25	0.6280	26	0.0352
L	SOM-KM	$m*n/t$	15.85 ± 9.67	0.7367	32	0.1711	29	0.6339	21	0.0345
L	DIANA	$m*n/t$	15.85 ± 9.67	0.7370	31	0.1702	32	0.6345	20	0.0344
G	KMEANS	$\#Com$	26	0.7429	27	0.1712	28	0.6265	29	0.0341
L	PAM	$m*n/t$	15.85 ± 9.67	0.7389	30	0.1707	31	0.6290	24	0.0340
G	KMEANS	$m*n/t$	24	0.7415	29	0.1710	30	0.6218	32	0.0332
C-SubP				0.7352	33	0.1655	34	0.6108	35	0.0319
L	DIANA	$\#Com$	6.02 ± 4.52	0.7214	34	0.1652	35	0.6131	34	0.0318
L	KMEANS	$\#Com$	6.02 ± 4.52	0.7079	36	0.1622	37	0.5996	36	0.0311
L	PAM	$\#Com$	6.02 ± 4.52	0.7062	37	0.1628	36	0.5861	37	0.0310
L	SOM-KM	$\#Com$	6.02 ± 4.52	0.6913	38	0.1583	38	0.5826	39	0.0305
L	LDA	$\#Com$	6.02 ± 4.52	0.6900	39	0.1576	39	0.5842	38	0.0304

Tabla 5.3: Valores de las medidas de evaluación para los perfiles basados en clusters y los casos base para el recomendación (Etiquetas de las columnas: T = Tipo de clustering; (L)ocal o (G)lobal; k = método para estimar el número de clusters; #$Clusters$ = número clusters; r@10 = recall 10 primeros; P-r = Posición en el ranking de recall; p@10 = precisión 10 primeros; P-p = Posición en el ranking de precisión; ndcg@10 = NDCG 10 primeros; P-ndcg = Posición en el ranking de NDCG; RRF = valor de Reciprocal Rank Fusion).

T	Alg.	k	#Clusters	r@10	P-r	p@10	P-p	ndcg@10	P-ndcg	RRF
L	LDA	$m*n/t$	15.85 ± 9.67	0.6529	1	0.1486	1	0.5195	2	0.0489
L	KMEANS	$m*n/t$	15.85 ± 9.67	0.6502	2	0.1482	3	0.5214	1	0.0484
L	SOM-KM	$m*n/t$	15.85 ± 9.67	0.6498	3	0.1482	2	0.5178	4	0.0476
L	PAM	$m*n/t$	15.85 ± 9.67	0.6481	4	0.1475	4	0.5183	3	0.0471
G	AGNES	$\sqrt{n/2}$	70	0.6438	5	0.1465	6	0.5065	9	0.0450
G	DIANA	$\sqrt{n/2}$	70	0.6408	7	0.1459	8	0.5163	5	0.0450
G	SOM-KM	$\sqrt{n/2}$	70	0.6437	6	0.1470	5	0.5023	10	0.0448
G	LDA	$\sqrt{n/2}$	70	0.6398	8	0.1459	7	0.5022	11	0.0437
L	DIANA	$m*n/t$	15.85 ± 9.67	0.6385	9	0.1453	11	0.5080	8	0.0433
G	DIANA	#Com	26	0.6357	13	0.1445	17	0.5113	6	0.0418
G	DIANA	$m*n/t$	24	0.6364	11	0.1445	18	0.5107	7	0.0418
G	KMEANS	$\sqrt{n/2}$	70	0.6380	10	0.1452	12	0.4961	16	0.0413
L	KMEANS	$\sqrt{n/2}$	4.25 ± 2.60	0.6350	15	0.1450	13	0.4976	12	0.0409
G	LDA	$m*n/t$	24	0.6342	16	0.1446	16	0.4962	13	0.0400
L	LDA	#Com	6.02 ± 4.52	0.6352	14	0.1457	9	0.4717	29	0.0392
L	AGNES	$m*n/t$	15.85 ± 9.67	0.6322	17	0.1443	19	0.4962	14	0.0392
L	LDA	$\sqrt{n/2}$	4.25 ± 2.60	0.6320	18	0.1442	20	0.4961	15	0.0387
G	LDA	#Com	26	0.6316	19	0.1440	22	0.4951	17	0.0378
L	SOM-KM	$\sqrt{n/2}$	4.25 ± 2.60	0.6313	20	0.1441	21	0.4950	18	0.0377
L	SOM-KM	#Com	6.02 ± 4.52	0.6295	21	0.1449	14	0.4665	32	0.0367
L	KMEANS	#Com	6.02 ± 4.52	0.6284	22	0.1447	15	0.4685	31	0.0365
C-SubP				0.6048	38	0.1457	10	0.4795	25	0.0363
G	SOM-KM	$m*n/t$	24	0.6281	23	0.1435	25	0.4807	22	0.0360
L	DIANA	$\sqrt{n/2}$	4.25 ± 2.60	0.6262	26	0.1430	27	0.4892	19	0.0358
G	KMEANS	$m*n/t$	24	0.6278	24	0.1437	24	0.4796	24	0.0357
G	SOM-KM	#Com	26	0.6254	27	0.1432	26	0.4804	23	0.0352
L	PAM	$\sqrt{n/2}$	4.25 ± 2.60	0.6247	29	0.1426	29	0.4880	20	0.0350
L	DIANA	#Com	6.02 ± 4.52	0.6262	25	0.1440	23	0.4687	30	0.0349
M-Prof				0.6358	12	0.1357	38	0.4546	35	0.0346
G	KMEANS	#Com	26	0.6248	28	0.1429	28	0.4791	26	0.0344
L	AGNES	$\sqrt{n/2}$	4.25 ± 2.60	0.6215	30	0.1422	32	0.4747	27	0.0335
G	PAM	$\sqrt{n/2}$	70	0.6151	33	0.1398	33	0.4721	28	0.0329
I-SubP				0.5959	39	0.1355	39	0.4868	21	0.0325
L	PAM	#Com	6.02 ± 4.52	0.6173	31	0.1423	31	0.4537	36	0.0324
L	AGNES	#Com	6.02 ± 4.52	0.6172	32	0.1423	30	0.4440	39	0.0321
G	AGNES	$m*n/t$	24	0.6123	34	0.1392	34	0.4548	34	0.0319
G	AGNES	#Com	26	0.6119	35	0.1389	35	0.4548	33	0.0318
G	PAM	$m*n/t$	24	0.6097	36	0.1387	36	0.4520	37	0.0311
G	PAM	#Com	26	0.6089	37	0.1384	37	0.4502	38	0.0308

empleado, la tercera columna contiene el criterio que se ha establecido para hallar el valor de k y finalmente, el valor de k (#Clusters[2]). Dentro de esas cuatro columnas se han incluido también los casos bases: los perfiles monolíticos (M-Prof), los subperfiles basados en comisiones (C-SubP) y los subperfiles basado en intervenciones (I-SubP). Las columnas etiquetadas con r@10, p@10 y $n$dcg@10 contienen los valores de las métricas de recall, precisión y NDCG respectivamente para los 10 primeros documentos del ranking. Puesto que el uso de diferentes medidas puede dar lugar sin duda a diferentes rankings para establecer que aproximación es mejor sobre el resto, se ha propuesto sintetizar las tres medidas en una única medida de carácter global para que se pueda observar de forma más clara a la hora de comparar los distintos métodos. Por lo tanto, se ha utilizado la medida Reciprocal Rank Fusion tal y como se define en [33], la cual es un método que se utiliza de forma original para combinar rankings de diferentes Sistemas de Recuperación de Información para ofrecer así un único ranking. En definitiva, se ha calculado la posición de cada uno de los métodos utilizados y los casos base en el ranking para cada una de las medidas (en las tablas, las columnas P-r, P-p y P-ndcg) y el valor final para la métrica Reciprocal Rank Fusion (RRF en las tablas). Finalmente, se han ordenado los métodos de manera decreciente de acuerdo al valor de RRF siendo esta forma la más justa de presentar los resultados con el objetivo de facilitar el análisis y extraer conclusiones de los mismos.

En primer lugar, centrando la atención en el rendimiento de los distintos métodos basales tanto en las tareas de recomendación como en las tareas de filtrado y teniendo en cuenta solo el ranking agregado de las distintas medidas de evaluación tenemos que, como primera conclusión los casos base se sitúan en la parte más baja de ambas tablas, quedando todas ellas por debajo de la mitad de cada tabla. Esto significa que existe un gran número de técnicas de clustering que superan en rendimiento a los casos base. En el contexto de las tareas de filtrado y en términos de rendimiento es destacable que los perfiles monolíticos y los subperfiles basados en intervenciones son mejores que la aproximación de los subperfiles basado en comisiones. Por otro lado, con respecto al problema de recomendación, el mejor caso base en este caso es la aproximación de los subperfiles basados en comisiones siendo el peor de los casos base la aproximación de los subperfiles basados en intervenciones.

En el caso del problema de filtrado, se ha considerado que la medida de recall es más importante que es resto, en el sentido en que el sistema debe ser capaz de encontrar el máximo número posible de MPs relevantes, es decir, identificar el mayor número de MPs que puedan estar interesados en un documento que quiere ser filtrado. Por otro lado, en el problema de búsqueda de expertos, la métrica mas importante en este caso podría ser la NDCG en lugar del recall, puesto que no existe tanto interés en encontrar un gran número de MPs interesados en la consulta sino que los que se encuentren sean correctos de forma unívoca estando situados en las partes más altas del ranking. En ambos casos, con sus respectivas métricas más importantes, se puede establecer que los perfiles monolíticos tiene un peor rendimiento y por el contrario los subperfiles

[2]Para el clustering local, como el #Clusters depende de cada MP, se muestra la media y la desviación estándar de todos los MP

basados en intervenciones actúan de mejor manera.

Observando los resultados que se presentan en las tablas se puede asegurar que hay un número considerable de algoritmos de clustering que tienen un rendimiento mejor que los casos base tanto en las tareas de recomendación como en las tareas de filtrado. En ambos casos, más de la mitad de las aproximaciones planteadas superan en el ranking agregado RRF al mejor caso base en sus respectivas tablas. Este número incrementa a dos tercios en el caso del filtrado con la métrica de recall y NDCG para la recomendación.

La Tabla 5.4 muestra el porcentaje de mejora de los algoritmos de clustering que mejor se comportan en las tareas de recomendación y filtrado, considerando la medida NDCG y recall respectivamente, con respecto a los casos base. Los porcentajes son moderados pero sirven para respaldar el hecho de que la aplicación de técnicas de clustering es una buena alternativa para la detección de temáticas implícitas y por lo tanto para la construcción de subperfiles. Cabe destacar, además, que las mayores diferencias porcentuales existen entre la utilización de técnicas de clustering y la aproximación base de perfiles monolíticos (M-Prof), lo cual es positivo, puesto que da soporte al hecho de que se obtiene un mejor rendimiento usando subperfiles basados en clustering que utilizando un único perfil con toda la información del MP. Estos porcentajes son menores cuando se comparan los algoritmos de clustering con la aproximación basada en comisiones (C-SubP) pero aún así son relevantes y, además esto respalda la hipótesis de que los perfiles basados en intervenciones (C-SubI) en algunos casos recogen de forma más eficiente las temáticas que las divisiones políticas en comisiones. También cabe mencionar que las diferencias entre los mejores métodos de clustering y las aproximaciones basales son, en todo caso, estadísticamente significativas (usando un t-test) tal y como ocurre como la mayoría de los algoritmos situados por encima de los casos base.

Tabla 5.4: Porcentajes de mejora de los métodos de clustering con respecto a los baselines. El símbolo $*$ representa que existe una diferencia estadísticamente significativa.

	Filtrado - Recall	**Recomendación - NDCG**
	Global AGNES $\sqrt{n/2}$	**Local LDA m*n/t**
M-Prof	7.35 % *	14.27 % *
C-SubP	5.06 % *	8.33 % *
I-SubP	2.91 % *	6.72 % *

Como conclusión general a este análisis y validación a la segunda cuestión que se planteaba en esta sección, las aproximaciones en las que se construyen subperfiles basados en la aplicación de técnicas de clustering son una buena opción a la hora de abordar los problemas de recomendación y filtrado dado que su rendimiento es apreciablemente mejor que el de los casos base propuestos. Incluso pueden llegar a ser mejores que la realización de una división artificial en distintas comisiones, las cuales han sido construidas por razones de contexto político y no de forma natural. De hecho, los subperfiles generados a partir de la aplicación de métodos de agrupamiento son capaces de representar las temáticas en las que un usuario puede estar interesado de

una forma más precisa. La aplicación de técnicas de clustering, además, permite la creación de diferentes grupos para temáticas pertenecientes a la misma comisión o, del mismo modo, permite la combinación de distintas facetas que habían sido separadas de forma artificial en dos o más comisiones diferentes y, de este modo, los subperfiles basados en clustering son una mejor aproximación comparada con el uso de un único perfil por MP donde todas las temáticas formaban parte de una amalgama heterogénea donde eran casi indetectables.

- **En la tercera cuestión se presenta la disyuntiva entre si existe alguna diferencia significativa entre las aproximaciones local y global que se proponen.**

Otra de las cuestiones que se lanzaba planteaba si existía alguna diferencia entre aplicar métodos de agrupamiento a MPs de forma independiente, es decir, la aproximación local, o bien aplicar dichos métodos sobre todo el conjunto de MPs a la vez, tal y como se define en la aproximación global. En el caso del filtrado de nuevos documentos y tomando como métricas el recall y la medida de rankings agregada RRF, es evidente de forma general que en la aproximación global se obtienen mejores resultados que en la aproximación local: la mayoría de los algoritmos de clustering en su versión global quedan en la parte superior del ranking mientras que la versión local se comporta de peor forma, de hecho, en algunos casos quedan incluso por debajo de los casos base. Esta distinción en términos de rendimiento no es tan evidente en el caso de los problemas de recomendación, donde en este caso se toma como referencia la métrica NDCG y la medida de rankings agregados RRF. En este caso, las apariciones en el ranking de las aproximaciones locales y globales de distintos algoritmos se encuentran distribuidas de forma más homogénea, aunque si es cierto que los algoritmos que encabezan la lista son los que construyen los subperfiles de forma local. Para argumentar esta conclusión se han calculado la media de las posiciones en el ranking de las aproximaciones locales y globales por separado tanto para las tareas de recomendación como para las tareas de filtrado. En el caso del problema de filtrado, el valor medio de las posiciones del ranking para las aproximaciones globales es de 15.28 y para las aproximaciones locales 23.67 (menor valor de media significa mejor resultado). Por otro lado, para el problema de recomendación, el valor de las medias para las aproximaciones globales es de 20.94 y para el caso de las aproximaciones locales es de 17.22, así que en este caso la diferencia no se puede considerar tan evidente como en el caso de las tareas de filtrado.

Una posible explicación a este comportamiento podría ser que en las aproximaciones locales se obliga de alguna forma a las intervenciones de un MP a distribuirse entre exactamente k clusters y esto provoca entre otras cosas que los documentos se dividan en más temáticas de las que realmente existen, creando así subperfiles muy difusos en lo referente a la información que contienen. Sin embargo, en la aproximaciones globales esto no se produce ya que los documentos de un MP no tienen que distribuirse necesariamente entre todos los k clusters sino solo a un subconjunto de ellos, por lo tanto se obtienen subperfiles más cohesionados y menos artificiales. Esto se traduce en que el tamaño de los subperfiles en las aproximaciones locales en general

sea menor que el de los subperfiles generados en las aproximaciones globales. A su vez, en las tareas de filtrado de documentos, dado que las consultas se corresponden con el texto completo de una iniciativa, estas contienen muchos más términos con respecto a las consultas en las tareas de recomendación que son básicamente un conjunto más pequeño de términos. Dicho esto, la conclusión que se obtiene al respecto es que las consultas con muchos términos, como en el caso del problema de filtrado, tienen un mejor rendimiento con los subperfiles obtenidos de manera global, los cuales tienen un mayor número de términos. En el caso de que la consulta sea más corta, como ocurre en las tareas de recomendación, el tamaño de los subperfiles no tiene tanta importancia y por lo tanto las aproximaciones locales globales se comportan de forma similar. En definitiva, para dar una respuesta a qué aproximación tiene un mayor rendimiento se puede afirmar que las aproximaciones de clustering globales son mejores para tratar con problemas de filtrado y, por otro lado, aunque no sea tan evidente se puede afirmar que en el caso de los problemas de recomendación, las aproximaciones de clustering local aportan, de manera general, mejores resultados.

- **En la cuarta cuestión surge el interrogante de si el número de grupos extraídos de la colección influye de forma relevante en el rendimiento de Sistema de Recomendación.**

Otra de las cuestiones planteaba si la elección del criterio para establecer el número de grupos y el valor de k en sí influía de forma relevante en el rendimiento del Sistema de Recomendación y Filtrado. Tal y como se ha mencionado con anterioridad, uno de los problemas menos triviales con los que se ha tratado en esta parte de la investigación ha sido el de seleccionar el número correcto de grupos en los que dividir las intervenciones de un MP, o de todo el conjunto de ellos, para poder extraer las correspondientes temáticas. Este problema se produce siempre que sea necesario aplicar técnicas de clustering independientemente del problema con el agravante de que en este caso de estudio, el análisis individual de cada uno de los MPs para intentar encontrar su valor de k específico no era viable. Por lo tanto, en este caso se han definido tres criterios distintos para calcular de forma objetiva un valor para el número de grupos sin que ello signifique encontrar un valor de forma exhaustiva probando con valores de forma consecutiva hasta llegar a una solución aceptable. Además de los métodos más utilizados en el estado del arte, $\sqrt{n/2}$ y más específicamente para clustering de colecciones documentales $m * n/t$, se ha usado también el número de comisiones $\#Com$ como caso base para el valor de k. Cabe destacar que, en el caso del clustering global, los valores de $\#Com$ y $m * n/t$ están muy cercanos (26 y 24 respectivamente) y por lo tanto, los resultados obtenidos con ambos criterios son muy parecidos con independencia de qué método de clustering o el tipo de problema en el que se apliquen.

En el caso de las aproximaciones globales en las tareas de filtrado de documentos, todos los criterios para establecer el número de grupos obtienen un mejor resultado que la generación de subperfiles basado en comisiones C-SubP tanto para la métrica de recall como para el ranking

agregado RRF y la mayoría son a su vez mejores que las aproximaciones basales de construcción de subperfiles basados en intervenciones I-SubP y los perfiles monolíticos M-Prof. Los mejores resultados para estos casos se obtienen estableciendo el valor de k con el criterio de $\sqrt{n/2}$ donde, al ser el valor más alto $k = 70$ cabe más espacio para encontrar temáticas más específicas que quedan diluidas en los casos donde el número de grupos es menor. Para las aproximaciones locales, por otro lado, es destacable el hecho de que el rendimiento de la mayoría de algoritmos de clustering es malo cuando se toma como criterio para establecer el número de grupos $k = \#Com$ el cual es incluso peor en todos los casos que la generación de subperfiles basados en comisiones C-SubP. En contrapartida, $\sqrt{n/2}$ es el criterio que se comporta mejor también en este caso y la razón de esto puede ser que en las aproximaciones locales, este criterio genera el menor número de grupos en media (4.25) lo cual conlleva a su vez que los subperfiles tengan un mayor tamaño, lo cual beneficia a las tareas de filtrado de documentos.

Para el caso de las tareas de recomendación y para las aproximaciones globales, una vez más $\sqrt{n/2}$ obtienes los mejores y más robustos resultados al margen del algoritmo de clustering que se utilice. Por otro lado, los resultados que se obtienen con los criterios $\#Com$ y $n * m/t$ dependen claramente del algoritmo de clustering donde se apliquen aunque de forma general son peores (para NDCG y el ranking agregado RRF). Centrándose ahora en el clustering de tipo local en las tareas de recomendación, parece ser que los criterios $\#Com$ y $\sqrt{n/2}$ no consiguen detectar el número suficiente de temáticas por generar un valor demasiado bajo para k siendo por lo tanto el criterio $n * m/t$ la mejor opción para estos casos (los clusters que se generan son mejores y los resultados obtenidos con este criterio presentan valores más altos). De hecho, en estos casos, cualquier algoritmo combinado con $n * m/t$ mejora el rendimiento tanto a la construcción de subperfiles basados en comisiones C-SubP como al resto de aproximaciones basales.

Con respecto a los algoritmos que se utilizan, AGNES (clustering jerárquico aglomerativo) muestra un comportamiento menos volátil con independencia del criterio para establecer el valor de grupos que se aplique para las aproximaciones tanto locales como globales en el problema de filtrado de documentos, y por el contrario, en el problema de recomendación, el rendimiento para este método varía claramente en función del criterio. Para el resto de algoritmos, no es posible establecer una conclusión tan obvia puesto que el rendimiento de los todos ellos cambia notablemente con respecto al número de clusters que se establezca, tanto si la aproximación es local o global o si el problema es de recomendación o filtrado.

Como conclusión para responder al interrogante que se plantea en esta cuestión sobre la importancia del valor de k se puede decir que para realizar un cluster de calidad es importante establecer un número de grupos adecuado tanto si el problema es de recomendación como de filtrado de documentos. Finalmente, cabe destacar que se han encontrado valores de k que mejoran los resultados obtenidos por el caso base donde se generan subperfiles basados en comisiones. Esto arroja la conclusión de que es bueno no restringir los subperfiles de los MPs solo a las comisiones donde participan, ya que pueden existir temáticas ocultas dentro de una misma

comisión o incluso presentarse varias temáticas de forma unificada por tratar los mismos temas en comisiones distintas.

Por último, se han llevado a cabo dos test de análisis de la varianza ANOVA, con un $\alpha = 0.05$, con el objetivo de concluir qué algoritmos se ajustan mejor con respecto a las tareas que se abordan. Uno de los test ANOVA se va a llevar a cabo los resultados de recall@10 para la tarea de filtrado, y el otro test ANOVA se va a ejecutar sobre la medida NDCG@10 para el problema de recomendación. La conclusión que se extrae del test ANOVA es que existen diferencias estadísticamente significativas entre las dos medidas (p-values de $4,9309E-37$ y $1.1842E-24$, respectivamente). Por lo cual, tomar una buena decisión con respecto a qué algoritmo de clustering se va a seleccionar en combinación de si se va a hacer con una aproximación global o local y el correcto número de clusters, se convierte en la tarea previa más importante para el rendimiento del problema de este caso de estudio.

- **Y finalmente, se plantea qué algoritmo o algoritmos de clustering se ajustan mejor para la tarea que se aborda en este caso de estudio.**

Las técnicas de clustering que funcionan mejor dependen del problema que se tenga que tratar. Si se toma el problema desde el enfoque de filtrado, las diferentes técnicas de clustering jerárquico funcionan mejor que el resto, específicamente, el clustering aglomerativo (AGNES) es el algoritmo que mejores resultados devuelve para este caso en la aproximación de clustering global con respecto a todas las medidas de evaluación que se han tenido en cuenta. Para el problema de recomendación de expertos, obteniéndose también buenos resultados con AGNES, en definitiva la mejor aproximación para este tipo de problema es la versión local del algoritmo LDA para la mayoría de las métricas que se han utilizado, seguido de cerca por el algoritmo basado en centroides K-MEANS , la versión SOM-KM Y PAM, todos ellos también en su versión local. En esta caso, se puede ver claramente cómo el problema de recomendación de expertos puede ser abordado satisfactoriamente por múltiples técnicas de clustering, lo cual lleva a intuir que para este tipo de problema la correcta elección de un algoritmo no es tan relevante siempre y cuando estos sean en su versión local. Aplicando un test ANOVA para las cinco primeras combinaciones de aproximaciones de las Tablas 5.2 y 5.3, una vez más sobre valores de las medidas recall@10 y ndcg@10 respectivamente, el resultado es que no existen diferencias estadísticamente significativas entre ellos (p-values de 0.8154 y 0.9691, respectivamente). Esto significa que cualquiera de estas combinaciones puede utilizarse para sus respectivas tareas con las garantía de que se van a obtener buenos resultados. De todos modos, es importante tal y como se ha referido con anterioridad, que la obtención de buenos resultados con los algoritmos de clustering está estrechamente relacionada con el valor del parámetro k que define el número de grupos.

Finalmente, se han representado gráficamente los valores de las medidas de recall@10 (para filtrado) y ndcg@10 (para recomendación) de todas las variaciones de aproximaciones de clustering

que se han llevado a cabo en la Figura 5.7. El objetivo de observar esta representación es el de descubrir fácilmente qué aproximaciones, si existen, tienen un mejor rendimiento, tanto para las tareas de recomendación como las de filtrado, al mismo tiempo. En la figura se han utilizado distintos símbolos para la representación del clustering global (círculo) y para el clustering local (aspa). Además, se han utilizado distintos colores para representar el criterio para establecer el valor de k, asignando el color verde para $\#Com$, el color rojo para el criterio $\sqrt{n/2}$ y por último $m * n/t$ con el color azul. Por último, el algoritmo de clustering en sí, se representa con la primera letra de su nombre junto al símbolo, es decir, Kmeans, Lda, Diana, Agnes, Som, Pam. Además se han añadido al gráfico los tres casos base (MONolítico, COMisiones e INTervenciones), representadas con un triángulo.

En la Figura 5.7 se puede observar que el uso de $\#Com$ como criterio para establecer el número de grupos no es, en definitiva, una buena forma de asignarle un valor al parámetro k para las tareas de filtrado, donde se han obtenido los peores resultados, pero tampoco es buena idea usarlo en las tareas recomendación. Este hecho se puede considerar como un evidencia de que el número de comisiones que se establecen en el parlamento no se ajusta debidamente a las distintas temáticas que en realidad se tratan en el parlamento, es más, para obtener unos mejores resultados en el sistema no es tan siquiera necesario conocer el número de comisiones parlamentarias. Esto puede resultar interesante en el sentido en que estas aproximaciones pueden ser aplicadas de la misma forma en escenarios donde, o bien no se disponga de esa información, o bien no exista.

Además, con la pretensión de encontrar una estrategia que funcione de manera aceptable tanto frente a los problemas de filtrado como a los de recomendación, las aproximaciones que tienen mejor rendimiento que la mejor aproximación basal (subperfiles basados en intervenciones) serían las más recomendables de utilizar, es decir, aquellas que quedan situadas en el sector superior derecha del gráfico (delimitados con líneas negras). En este caso, se puede observar claramente que unas alternativas de razonable calidad podrían ser aplicando tanto métodos jerárquicos como el algoritmo LDA en su versión global y tomando $\sqrt{n/2}$ como criterio para establecer el número de grupos, obteniendo los mejores resultados con el algoritmo de clustering AGNES en su versión global y tomando $\sqrt{n/2}$ como valor de k. Sin embargo, si solo se tiene en cuenta el problema de recomendación, esta última aproximación referida tiene un comportamiento muy dependiente del valor del resto de parámetros, siendo de vital importancia una estimación correcta de los parámetros previamente. Dicho esto, si se busca una estrategia de clustering robusta, que se ajuste de igual forma a los problemas de recomendación y filtrado y que lo haga de la forma más estable posible, tanto LDA como DIANA en su versión de clustering global podrían ser las mejores alternativas.

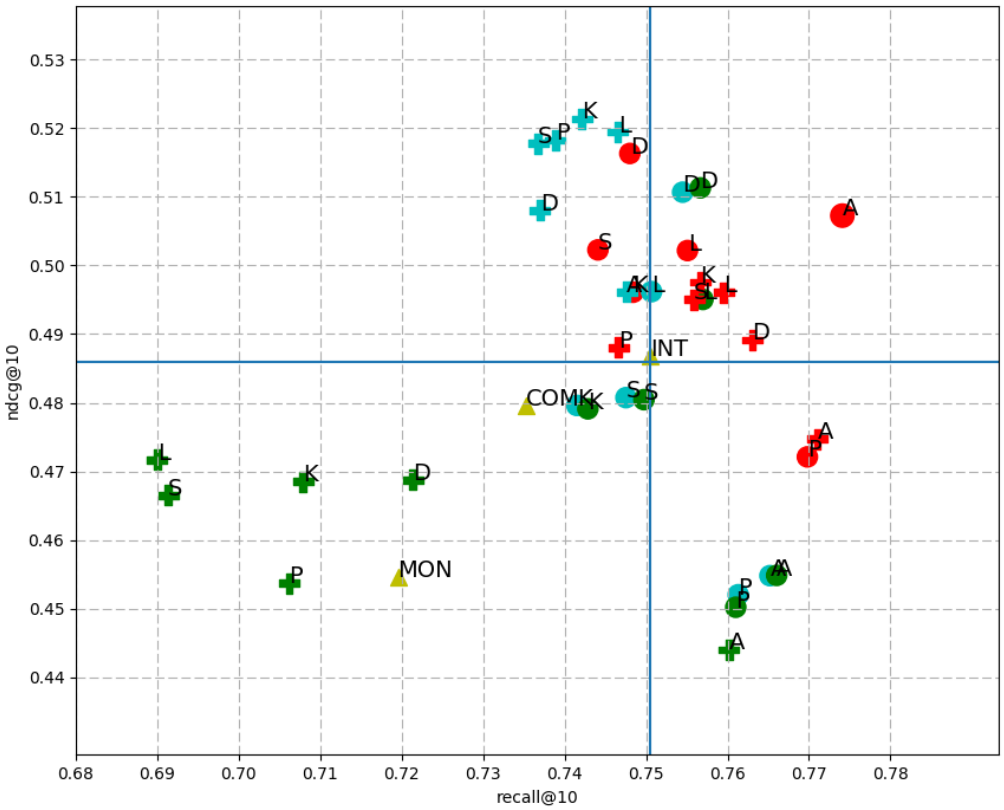

Figura 5.7: recall@10 vs ndcg@10 para todas las combinaciones de clustering y las aproximaciones basales.

5.5 Conclusiones

En este capítulo se ha presentado una propuesta basada en clustering de documentos textuales para construir de forma automática perfiles compuestos de usuarios expertos con el objetivo de reflejar de forma apropiada las temáticas en las que el usuario puede estar interesado. Para ello se han considerado dos dominios de aplicación ampliamente referidos en la literatura como pueden ser los campos de filtrado y recomendación de expertos. En el primer caso, la tarea que se aborda consiste en decidir qué expertos pueden estar interesados en recibir un nuevo documento que llega al sistema de acuerdo a sus intereses o campo profesional. En el segundo caso se lidia con el problema de encontrar un conjunto expertos que sean más afines a las necesidades de información expresadas en forma de consulta por un usuario del sistema. El ámbito específico sobre el que se ha llevado a cabo la evaluación experimental ha sido desde el punto de vista político, donde los Miembros del Parlamento jugaban el papel de expertos y la fuente de información utilizada para

construir los perfiles de usuario del sistema se corresponde con las transcripciones literales de las intervenciones de los MPs en los debates de las iniciativas parlamentarias.

A pesar de que ambos problemas, filtrado y recomendación, se pueden formular de forma unificada; dada una consulta, ya sea un documento que necesita ser filtrado o una necesidad de información que debe ser cumplida, se obtiene una lista expertos en forma de ranking, los cuales pueden estar interesados tanto en recibir información de un nuevo documento como bien de satisfacer una demanda de información. Y aunque ambos problemas se pueden abordar utilizando aproximaciones similares, los resultados obtenidos en el proceso de experimentación sugieren que existen diferencias significativas entre ambos. Estas diferencias sirven para determinar que existen diferentes herramientas para resolver los problemas que aquí se plantean de forma apropiada, ya sea el tipo de clustering que se realiza, abordando el problema desde un enfoque local o global o bien el número de grupos en los que se divide el conjunto de documentos.

Tal y como se ha referido, se han propuesto dos alternativas distintas para la extracción de grupos: un enfoque local y alternativamente un enfoque global. En el primer enfoque los cluster de documentos se realizaban de forma independiente para cada uno de los Miembros del Parlamento, mientras que en el enfoque global se realizaba un único proceso de clustering sobre todos los documentos de todo el conjunto de Miembros del Parlamento. En ambos casos se han aplicado algoritmos de clustering de muy diversa naturaleza: jerárquicos aglomerativos y divisivos, basados en centroides y medoides, basados en modelos estadísticos generativos y basados en redes neuronales, además de diferentes métodos para estimar un valor adecuado para el número de clusters. Se han considerado también tres casos base distintos: dos casos extremos, un perfil único (monolítico) para cada Miembro del Parlamento y un subperfil independiente por cada intervención de un Miembro del Parlamento determinado, además una aproximación alternativa donde los subperfiles de un MP no se construían a partir de un proceso de clustering sino que el subperfil se componía a partir de la extracción de las intervenciones de un Miembro del Parlamento en cada una de la comisiones en las que participaba puestas de forma conjunta.

Dicho esto, las principales conclusiones que se pueden extraer del proceso de experimentación que se ha llevado a cabo se pueden enumerar como sigue. En primer lugar se puede afirmar que de forma general, aplicar técnicas de clustering en este tipo de problemas, es una buena opción para la detección y descubrimiento de grupos de documentos en los que se relatan distintas temáticas de interés, mejorando incluso situaciones donde los grupos está previamente impuestos de forma explícita y externa, tal y como puede ser una comisión parlamentaria. Por otro lado, otra conclusión a destacar podría ser que la mayoría de las alternativas aplicadas basadas en clustering superan, en cuestión de rendimiento, a los tres casos basales propuestos tanto en el enfoque de filtrado como en el de recomendación, con diferencias positivas estadísticamente significativas. Otra conclusión que se ha extraído desde el punto de vista de la elección del tipo de clustering, para el problema de filtrado de documentos, la elección preferible es la aproximación de clustering global de forma clara, sin embargo, para tratar el problema desde el enfoque de la recomendación

la situación no es tan evidente, a pesar de que las primeras cuatro alternativas, desde el punto de vista del rendimiento, se han llevado a cabo con carácter local. Este comportamiento está relacionado de forma directa con el tamaño de los clusters que cada aproximación genera y el hecho de que en un problema de filtrado el tamaño de la consulta, en este caso el documento completo, es generalmente mayor que en el mismo problema con un enfoque de recomendación, donde la consulta es la información que requiere el usuario que la formula. Desde el punto de vista de la selección del número de clusters, la conclusión que se extrae es la de que de nuevo existen diferencias entre tratar el problema con un planteamiento de filtrado o con uno de recomendación, es decir, para el filtrado de documentos el mejor método para establecer el número de clusters es $\sqrt{n/2}$ y, para recomendación sin embargo, mn/t funciona mejor. Finalmente, observando de forma concreta los algoritmos de clustering que se han utilizado se puede concluir que los métodos jerárquicos, en particular el aglomerativo, funcionan bastante mejor para problemas de filtrado, mientras que LDA, SOM y los métodos basados en centroides obtienen un mejor rendimiento para el problema de recomendación. No obstante, los resultados indican que la decisión sobre escoger un algoritmo específico de clustering u otro no es crítica, puesto que no se han encontrado diferencias estadísticamente significativas entre las cinco mejores aproximaciones llevadas a cabo para cada tipo de problema.

A la vista de los resultados obtenidos y las conclusiones extraídas a partir de ellos, en el siguiente capítulo de esta Tesis Doctoral se plantea como nueva línea de investigación explorar las capacidades potenciales del algoritmo LDA para construir subperfiles más robustos desde la perspectiva del contenido semántico de la información, en la cual este algoritmo está concretamente especializado.

5.6 Trabajos relacionados

En esta sección se van a relatar algunos de los trabajos de investigación más relevantes en materia de construcción de perfiles de usuarios. A pesar de que algunas de estas publicaciones no aplican técnicas de clustering, podría ser interesante que figurasen aquí a modo de revisión puesto que, en este capítulo se construyen perfiles multifacéticos y la mayoría de los autores coinciden en que este enfoque afecta de forma positiva a la representación de los perfiles de usuarios. Con el objetivo de sintetizar los trabajos que se han publicado al respecto, en la Tabla 5.6 que se incluye en esta sección se representan las siguientes principales características: propósito específico de los perfiles (Purpose); fuente de la información que se ha utilizado para construir los perfiles (Inf. Source); si se ha aplicado, qué algoritmo de clustering se ha utilizado (Clust. Alg.); entidades que se han tenido en consideración (Entity cons.) y sus correspondientes características (Features); y finalmente, el tipo de perfil compuesto que se ha obtenido (Type of Profile). Además, para poder realizar la comparativa, se ha incluido la aproximación propuesta en este capítulo en la última línea de la tabla.

En un primer grupo de publicaciones relacionadas estrechamente entre sí, se aborda el problema con una forma estructurada de representar los perfiles basándose en diferentes fuentes de información. En el contexto de la búsqueda de expertos, [108] concibe un perfil estructurado donde se reúne información personal de los expertos, intereses y ámbito profesional sobre el que operan, y los representa en forma de árbol con tres ramas independientes desde las que cuelgan los términos contenidos en cada una de fuentes de donde se extrae la información. Un caso alternativo a este tipo de representación se muestra en [58] donde, basándose en la lista de temáticas de Twitter, se generan dos subperfiles para los usuarios: el primero engloba una serie de etiquetas que definen al usuario en sí y el segundo subperfil contiene etiquetas que vienen definidas por los contactos del usuario. Un tercer ejemplo se plantea en [157] donde los autores recogen la información del usuario de distintas fuentes (páginas Web, usuarios, elementos, consultas, etc), y estas se agrupan en un grafo multinivel, creando así conexiones entre los diferentes cluster aplicando refuerzo mutuo. Finalmente, en [102], con el objetivo de recomendar elementos de carácter social se crean tres subperfiles diferentes usando como fuente de información las palabras claves de las publicaciones sociales de los propios usuarios, una serie de etiquetas asociadas a dichas palabras claves y nuevos términos que conectan conceptos de forma subyacente.

Otras publicaciones relacionadas con el caso de estudio que aquí se plantea presentan aproximaciones donde la información tiene como origen una única fuente, generalmente documentos de texto, y se organiza en diferentes subperfiles. Un primer caso de diseño para la recomendación de noticias [55] propone un método para la representación de un perfil en dos facetas: un subperfil a largo plazo que comprende los términos y categorías de un historial de documentos relevantes; y un subperfil a corto plazo con la misma información pero creado después de que el primer subperfil se haya construido. Un segundo ejemplo a relatar es [16], donde el perfil, al igual que en el caso anterior, está definido por dos subperfiles: una lista de términos extraída a partir de los documentos que se han considerado verdaderamente relevantes, enriquecida a su vez por términos pertenecientes al cluster al cual pertenece el usuario en cuestión tras aplicar K-means a cada usuario y la lista de hiperónimos de Wordnet, y por otro lado el segundo subperfil contiene los términos de los documentos que han sido juzgados como no relevantes por el usuario.

Otro tipo de perfiles estructurados son aquellos que se basan en jerarquías. En la publicación [128] se construyen perfiles de expertos que contienen una jerarquía temporal, donde en cada nodo se representan temáticas con un peso asociado. En [28] se presenta un sistema de personalización basado en mantener una jerarquía sobre los intereses de los usuarios a partir de las páginas web que estos visitan. En ambos casos, las jerarquías no están aprendidas como tal sino que son impuestas de forma explícita.

Considerando una estructura donde los perfiles contengan una lista de categorías, temáticas o conceptos que no estén relacionado entre sí, y cada uno de ellos está representado por un conjunto de palabras claves, se puede citar el Sistema de Recomendación de *Syskill & Webert* [109], el

Sistema de Personalización *Alipes* [147] y el Sistema de Recomendación Webmate [29]. En los dos primeros casos, la lista de categorías viene dada por el sistema de una forma o otra, es decir, no se aprende de forma automática como en el tercer sistema que sí lo hace aplicando algoritmos de clustering. Otro ejemplo en esta categoría se presenta en [75] donde se generan unos perfiles basados en categorías no a partir de documentos completos sino de los términos de las consultas que se formulan al sistema.

Las siguientes publicaciones ofrecen una estructura de los perfiles similar pero aprendida de forma automática mediante la aplicación de técnicas de detección de grupos: [139] aplica un clustering incremental con un enfoque local para generar clusters de temáticas aplicado al problema de búsqueda personalizada; *Web Personae* [96] utiliza técnicas de clustering jerárquico sobre el conjunto de páginas web visitadas por el usuario, también con un enfoque local para el mismo propósito de búsqueda personalizada.

En [4], una vez que se han extraído los términos de las distintas fuentes de información, se aplica el algoritmo *Induced Bisecting K-means* para agrupar dichos términos en conceptos relacionados semánticamente, representando a cada usuario en este caso como un conjunto de áreas de investigación las cuales están caracterizadas por un conjunto de términos relacionados. Una aproximación similar a esta se plantea en [6], donde se utiliza una técnica de clustering de documentos basadas en los métodos de detección de comunidades, los autores crean grupos de etiquetas para representar a los usuarios. Mientras que este enfoque es de carácter local para el clustering basado en etiquetas, también se ha considerado el enfoque global pero utilizando términos como características. [113] plantea en esencia la misma idea que en [6], agrupar documentos similares, pero con la diferencia significativa de que en el primer caso no se representa de forma explícita los intereses del usuario sino que usa la estructura de clusters para recomendar directamente artículos científicos.

Reference	Purpose	Inf. Source	Clust. Alg.	Entity Cons.	Features	Type of profile
[108]	Expert finding	Heterogeneous documents	–	Documents	Keywords	Tree with nodes containing keywords
[58]	User modeling	Lists in Twitter	–	–	Tags assigned to lists	Intentional (weighted tags in lists the user follows) and Extensional profiles (weighted tags in lists the friends follows)
[157]	User modeling	Web heterogeneous objects	Probabilistic clustering	Web Objects	Keywords	Multi-layered graph with nodes in different layers representing different type of objects
[102]	Content-based recommendation	Social items from Facebook & Instagram	–	–	Keywords & Concepts	Wikipedia Concepts & Extended keywords
[16]	Content-based recommendation	News	Variation of K-Means	Users	Keywords	Terms in positive documents (and in clusters they belong to), in negative documents, plus WordNet hypernyms
[55]	Content-based recommendation	News	–	News	Keywords	Keywords in observed News, keywords in concepts
[128]	IR Personalization	Web pages	–	User's interest	Keywords	Hierarchy of web pages
[28]	IR Personalization	Web pages	–	Web pages	Keywords	Hierarchy of the user's interests
[109]	Content-based recommendation	Web pages & their categories	–	–	Keywords & Categories	List of categories and the associated terms
[147]	News personalization	News	–	–	Keywords and categories	List of categories comprising three lists of keywords, respectively
[29]	Content-based recommendation	Web pages	–	Keywords	–	List of categories comprising keywords
[75]	Collaborative tagging	Documents	Community Discovery-based	Documents	Keywords	Subprofiles comprising keywords
[139]	IR Personalization	Web pages	Incremental clustering	Web pages	Keywords	List of topics
[96]	IR Personalization	Web pages	Hierarchical clustering	Web pages	Keywords	List of cluster centroids
[113]	Content-based recommendation	Scientific articles	–	Scientific articles	Keywords	Clusters of articles
[6]	Collaborative tagging	Documents	Community discovery technique	Documents	Keywords	Subprofiles comprising tags (extracted from clustered documents)
[4]	Expert finding	Heterogeneous sources	Bisecting K-Means	Different sources	Keywords	Subprofiles of research areas
de Campos et al. (2018)	Content-based recommendation & filtering	Parliament initiatives	AGNES, DIANA, LDA, K-Means, PAM, SOM	Initiatives	Keywords	List of subprofiles, containing weighted keywords

Tabla 5.5: Resumen de los trabajos relacionados sobre perfiles compuestos.

Las cuatro primeras aproximaciones que se han presentado en esta revisión del estado del arte ([58, 102, 108, 157]) en el contexto de representación de perfiles de una forma estructurada basándose en la obtención de información de diversas fuentes, difieren con respecto a la aproximación que se propone en este capítulo en que mientras estos autores consideran múltiples fuentes de información para construir los perfiles de usuarios, en esta aproximación solo se considera una única fuente, además la estructura de los perfiles es relativamente compleja como para soportar la diversidad de fuentes de información. Otra diferencia a destacar es que a excepción de [102], ninguna de las publicaciones muestra mayor interés en detectar y representar temáticas subyacentes tal y como se lleva a cabo en este estudio. Por otro lado, los trabajos que se han referenciado usan conceptos que han sido extraídos de una fuente de información externa y no aprendidos de forma automática tal y como se hace en la propuesta de este capítulo de forma implícita. Finalmente, ninguno de los trabajos revisados utiliza técnicas de clustering para cumplir su propósito con excepción de [157].

En el caso donde la información proviene de una única fuente, generalmente documentos de texto, y se organiza posteriormente en diferentes perfiles [55], en la aproximación de este capítulo no se consideran documentos positivos ni documentos negativos y tampoco ningún tipo de retroalimentación por parte del usuario. La principal diferencia con las aproximaciones de las propuestas [28, 128], además de que no aplican ninguna técnica de clustering, es que en el planteamiento de este estudio no se usan jerarquías de conceptos para representar los perfiles tal y como hacen los autores citados. En este caso, los conceptos no están relacionados entre sí.

Moviendo la vista a las aproximaciones que construyen los perfiles a partir de una lista de categorías, temáticas o conceptos [29, 75, 109, 147], a pesar de que la estructura del perfil es muy similar a la que se presenta en [75], la principal diferencia es que, en la propuesta de este capítulo, se aplican técnicas de clustering de forma automática para crear las categorías de forma implícita. Mientras que estas aproximaciones solo abordan el problema desde un punto de vista local, en este caso de estudio también se considera un enfoque global utilizando toda la información de todos los usuarios de forma conjunta. Además, en la aproximación de este capítulo se usan los términos propios de los documentos como atributos mientras que en [75] se usan los términos de las consultas.

Con respecto a las publicaciones [139] y [96], las cuales construyen perfiles estructurados de forma similar por medio de la aplicación de técnicas de detección de grupos, una ve más, la principal diferencia con las aproximaciones de este capítulo es que solo se considera el enfoque local para la construcción de perfiles. La aplicación de los perfiles es otra de las diferencias más destacables con respecto a esta aproximación y las publicaciones referenciadas, mientras que en el estado del arte se considera a los perfiles como herramientas de personalización, en esta Tesis Doctoral por el contrario se consideran para construir un Sistema de Recomendación Basado en Contenido. Finalmente, una última diferencia se basa en la selección de perfiles, mientras que en los trabajos referenciados se usa solo el perfil más relevante de forma única, en esta propuesta se

combinan todos los perfiles con el objetivo de determinar que usuario debe ser recomendado.

Considerando las publicaciones [4, 6, 113], la principal diferencia es que, en este caso de estudio, los clusters que se han extraído de forma no supervisada contienen documentos de texto y además se considera un enfoque clustering global utilizando los términos de los documentos como atributos.

Además de las diferencias descritas en esta sección, cabe mencionar que en la experimentación de este caso de estudio se han probado la adecuación de los distintos algoritmos de clustering que se ha utilizado y los diferentes métodos para establecer el número de clusters. Es muy difícil encontrar alguna publicación específica que aborde el problema de cómo se puede determinar de forma unívoca el número de grupos a extraer de un conjunto de datos.

CAPÍTULO 6

CAPÍTULO

PERFILES DE TÉRMINOS BASADOS EN LDA PARA BÚSQUEDA DE EXPERTOS EN EL ÁMBITO PARLAMENTARIO

Una actividad bastante usual en algunas instituciones políticas, como por ejemplo en un Parlamento, es la de encontrar políticos expertos en un campo específico. Con el objetivo de abordar este problema, el primer paso es el de definir los perfiles de los políticos de manera que queden registrados sus intereses, los cuales se pueden aprender de forma automática a partir de sus discursos en el parlamento. Como los políticos pueden ser expertos en varios campos, una alternativa es construir varios subperfiles donde se recojan cada una de las distintas temáticas. En este capítulo, se propone una nueva aproximación para este propósito basada en Latent Dirichlet Allocation (LDA), para aprender las temáticas de cada intervención y distribuir los términos específicos de cada una de ellas en distintos subperfiles. Con este objetivo, se propone el uso de múltiples medidas de distancias y similitud para determinar de forma automática el número óptimo de temáticas que contiene un documento y demostrar así que todas las medidas utilizadas convergen a las estrategias Euclídea, Dice, Sorensen, Coseno y Overlap. Los resultados de los experimentos muestran que los valores de rendimiento obtenidos tienden a ser más altos que los casos base para las tareas de recomendación de expertos, por lo tanto elegir correctamente el número de temáticas de forma apropiada resulta ser relevante para esta tarea.

6.1 Introducción

Al igual que en los capítulos anteriores, se plantea el problema de construir un Sistema de Recomendación Basado en Contenido desde el punto de vista de las aproximaciones de recomendación y filtrado. En la aproximación de recomendación, el objetivo es el de, a partir de la información contenida en los documentos, recomendar al usuario una serie de individuos de acuerdo a la consulta que se lanza contra el sistema. A este tipo de problema se le denomina de forma más específica como búsqueda de expertos [9], puesto que los elementos que el sistema devuelve como respuesta a la consulta son personas físicas. Por otro lado, y también a partir de la información que se puede extraer de un texto se puede abordar el problema de filtrado de documentos, es decir, hacer llegar a los individuos únicamente aquella información que les puede resultar relevante. En este caso de estudio particular, tal como se viene realizando en esta tesis doctoral, el ámbito donde se va construir el sistema está relacionado con la esfera política, más concretamente, en el ámbito parlamentario donde los individuados que maneja el sistema son políticos y diputados. De esta forma, los elementos de donde se extrae la información para construir el sistema son las intervenciones de los diputados en los discursos parlamentarios, de donde se pueden observar las temáticas e intereses en los que los diputados pueden estar interesados de acuerdo a su actividad parlamentaria. Por ejemplo, en el caso concreto de un parlamento regional donde existe un Miembro del Parlamento que pertenece a la Comisión de Agricultura, se espera de él que tenga conocimientos relacionados con dicha comisión pero a su vez con comisiones relacionadas, es decir, legislación Europea y nacional, iniciativas donde se interpele la comisión, subsidios agrarios, presupuestos, exportaciones y comercio, etc.

Por un lado, cuando un usuario tiene una cuestión específica para plantear a los políticos correspondientes o requiere algún tipo de información sobre un tema concreto que se ha tratado en un pleno del parlamento, la principal tarea es la de encontrar de la forma más unívoca posible, qué Miembro del Parlamento desempeña su labor política en torno a ese tema. Por otro lado, los Miembros del Parlamento necesitan están informados sobre los asuntos del territorio donde desarrollan su labor política sin que la cantidad de información recibida sea abrumadora. A menudo, y mayormente en el ámbito político, la información relacionada con los diputados está distribuida en distintas fuentes o de difícil acceso, convirtiendo las tareas de búsqueda de expertos o filtrado de documentos en una labor tediosa y desalentadora. Para lidiar con estos problemas se plantea la construcción de un sistema especializado donde se recoja la información textual de los diputados y se almacene de forma que los usuarios puedan acceder a ella con mayor facilidad. De esta forma, el sistema devuelve sea cual sea el propósito, un ranking con aquellos Miembros del Parlamento a los que la información del usuario o el documento a filtrar les puede resultar más relevante.

La información de cada Miembro del Parlamento junto con las áreas de interés donde es experto se puede recoger de una amplia variedad de fuentes, por ejemplo comunicados o documentos que el diputado haya escrito, las transcripciones literales de sus intervenciones en el parlamento,

información periodística, etc. Pero en este caso en particular, tal y como se ha referido en los capítulos anteriores, la fuente de información que se va a utilizar para construir los perfiles de los diputados del sistema van a ser, las transcripciones de sus intervenciones de las sesiones plenarias y de comisiones del parlamento.

Dicho esto, puesto que los elementos que el sistema debe devolver son Miembros del Parlamento, los intereses de los posibles candidatos en los problemas de recomendación y filtrado se deben representar como perfiles. La forma de perfil más común en estos casos es la de reunir de alguna forma el conjunto de términos que determinan los intereses del individuo, pero en el caso de que el individuo tenga interés en temáticas diversas o heterogéneas, lo ideal no sería reunir todos los términos juntos en un único perfil, como era el caso del perfil monolítico (Sección 3.2.2). El motivo de esto es que algunas temáticas puede quedar ocultas en la medida en la que el MP se refiera a una temática en mayor medida que el resto, lo cual conllevaría desde el punto de vista del sistema, a que las temáticas a las que el diputado se refiere en menor medida se diluyesen y perdieran dentro de su perfil. Por tanto, una buena alternativa a este tipo de perfil monolítico sería la de distribuir la información del diputado en varios subperfiles, más claros y más homogéneos.

Por esta razón, el objetivo de este capítulo es el de encontrar una forma de dividir un perfil heterogéneo monolítico pero de una forma alternativa a la planteada en el Capítulo 5, siendo esta más específica al construir los subperfiles a nivel de términos. Así se puede lanzar la hipótesis, basándose en las conclusiones del capítulo anterior, de que la construcción de subperfiles que recojan de forma más concisa los intereses de un Miembro de Parlamento, conlleva a obtener un mejor rendimiento en el sistema. En este capítulo se va a utilizar la técnica Latent Dirichlet Allocation (LDA) [14] para intentar extraer las temáticas de la colección de documentos, pero a diferencia del Capítulo 5 donde se utilizaba LDA para agrupar los documentos en torno a la temática más probable, en este caso la aplicación de LDA sobre los documentos es algo distinta.

Tal y como se ha mencionado, la forma más extendida de construir los documentos en los que se recoge perfil de un individuo es mediante la utilización de vectores de términos ($bag-of-words$). Otra opción podría ser la de usar técnicas de modelado de temáticas, por ejemplo LDA, para que estos documentos pasen a estar construidos con vectores de los términos que están relacionados con cada una de las temáticas. Algunas aproximaciones de las que se relatan en la sección de trabajos relacionados 6.5, construyen los perfiles a partir de extracción de términos, otras aproximaciones lo hacen a partir de la extracción de temáticas, o incluso combinando ambas formas. En la aproximación que aquí se propone, se utiliza LDA para distribuir los términos correspondientes a cada uno de las temáticas extraídas de un documento en diversos subdocumentos. Por tanto, los subdocumentos que se hayan generado a partir de la misma temáticas se aglomeran en un único subperfil. Este procedimiento, sin embargo, puede generar un conjunto de perfiles mayor de lo esperado para algunos individuos (Miembros del Parlamento en este caso), lo cual tiene como consecuencia que algunos de los subperfiles obtenidos estén

compuestos por un conjunto muy reducido de términos, los cuales son prácticamente inservibles. Para lidiar con esta problemática se ha desarrollado un método general para reducir el número de subperfiles seleccionando solo y exclusivamente aquellos que se han generado a partir de las temáticas más relevantes de cada documento y redistribuyendo los términos entre ellos.

En definitiva, la aproximación que se plantea en este capítulo pretende vislumbrar si el uso de LDA, de una forma alternativa a como se utiliza en el Capítulo 5, es un buen método para construir subperfiles de diputados para las tareas de recomendación y filtrado. Para ello, se ha realizado un estudio sobre cómo utilizar LDA para construir subperfiles basados en términos asociados a una misma temática; se ha propuesto un método para distribuir las apariciones de los términos de un documento en distintos subdocumentos asociados cada uno de ellos a una temática específica, usando para ello las matrices generadas como salida por LDA; se ha planteado también un método general basado en medidas de distancia y similitud para asignar a cada documento un subconjunto óptimo de los temáticas asociadas originalmente al documento por LDA; por último, se ha llevado a cabo un proceso exhaustivo de experimentación usando, al igual que en capítulos anteriores, la colección de intervenciones parlamentarias y comparando los resultados con algunos de los obtenidos con otros modelos.

6.2 Aplicación de LDA para obtener subperfiles homogéneos

Se tiene un conjunto de Miembros del Parlamento (MP) para el problema de Búsqueda de Expertos y una colección $\mathscr{D}$ de documentos, donde cada uno de estos documentos está asociado a un MP. Dicho esto, cada uno de los documentos que componen esta colección se corresponde con cada una de las intervenciones que el MP ha llevado a cabo en los debates parlamentarios y de donde se pueden extraer, por tanto, los intereses propios del MP en cuestión. Cómo se ha mencionado en la Sección 6.1, el objetivo principal de esta aproximación es el de distribuir todos los términos que se aglomeran en un perfil monolítico heterogéneo, en un conjunto de subperfiles más homogéneos basados en temáticas. Para este propósito, se aplica el algoritmo LDA sobre toda la colección de documentos para la aproximación global, la cual será la única considerada en este estudio. De esta forma, como salida, LDA para un número de temáticas k devuelve dos matrices para una colección de documentos compuesta de n documentos y m términos distintos. Esta matrices tienen unas dimensiones de $m \times k$ y $k \times n$, donde cada una de las entradas de la matriz representa $p(t|x)$ (probabilidad de pertenencia de un término t a una temática x) y $p(x|d)$ (probabilidad de que en un documento d se trate la temática x), respectivamente.

6.2.1 Distribución de documentos en subdocumentos homogéneos

Una vez aplicado LDA a una colección de documentos y obtenidas las correspondientes matrices de probabilidad, el segundo paso es distribuir cada documento d de la colección en varios (inicialmente k) subdocumentos d_i de forma que, cada subdocumento d_i comprenda la parte

de d asociada a la parte específica correspondiente a la temática x_i. Para llevar a cabo esto, por cada término $t \in d$, se necesita un método para determinar a qué subdocumento debería estar destinado t. Por ejemplo, en un documento que trate sobre la escolarización de niños hospitalizados, LDA probablemente detecte de forma automática 2 temáticas diferentes: Sanidad y Educación. Por tanto, los términos asociados a la temática de Sanidad (hospital, paciente, etc) deberían ir destinados a un subdocumento y, por otro lado, los términos asociados a la temática de Educación (escuela, alumno, etc) se deberían reflejar en otro subdocumento distinto. Esta forma de dividir un documento no es exclusiva, en el sentido en que ciertos términos pueden pertenecer a varias temáticas en mayor o menor medida (el término $niño$, si nos ceñimos al ejemplo anterior). Por tanto, para cada término t que aparece en d, se debe discernir el número de instancias (apariciones del término en d) que se deben asignar a cada uno de los subdocumentos d_i. En cierto modo, esto se correspondería con el cálculo de la importancia de un término t en cada una de las temáticas de las que se compone el documento d. Cuando en la colección se dispone de información externa o privilegiada este problema se puede resolver de forma supervisada, pero en este caso se debe realizar de forma no supervisada.

En esta aproximación se plantea usar el algoritmo LDA de forma que se puedan distribuir cada una de las apariciones de un término t en un documento d, con notación $freq(t,d)$, entre una serie de subdocumentos asociados a d de forma proporcional a las probabilidades $p(x_i|t,d)$, $i = 1,\ldots,k$. Por tanto, es esencial poder estimar de una forma eficiente estas probabilidades a partir de las matrices obtenidas como salida del algoritmo LDA. Para este objetivo, se va a asumir la relación de independencia donde los términos y los documentos son condicionalmente independientes dada una temática, es decir, $p(t|x,d) = p(t|x)$.

Se usa esta relación de independencia para dado

$$p(t,x|d) = p(t|x,d)p(x|d) = p(t|x)p(x|d)$$

y, además dado

$$p(t|d) = \sum_{j=1}^{k} p(t,x_j|d) = \sum_{j=1}^{k} p(t|x_j)p(x_j|d)$$

para obtener

$$p(x|t,d) = \frac{p(t,x|d)}{p(t|d)} = \frac{p(t|x)p(x|d)}{\sum_{j=1}^{k} p(t|x_j)p(x_j|d)} \tag{6.1}$$

Y de esta forma, se obtiene la manera de calcular $p(x|t,d)$ a partir de las matrices obtenidas como resultado del algoritmo LDA.

Una vez que se han obtenido las probabilidades de una temática $p(x|t,d)$ para los términos t en cada uno de los documentos d, se distribuyen las instancias de t entre todos los subdocumentos en los que se divide d a partir del cálculo del producto $freq(t,d) * p(x|t,d)$ y redondeando este valor al entero más cercano. Puede ocurrir que al redondear el número de apariciones de un término al entero más cercano, en el cómputo global de todos los subperfiles se pierdan o se

encuentren más apariciones del término de las que existen en el documento original. Para lidiar con este problema, se ha diseñado un método por el cual se eliminan las apariciones adicionales de un término de los subdocumentos asociados a las temáticas menos probables y, por otro lado se añaden apariciones a los subdocumentos asociados a las temáticas más probables si el caso es que se han perdido instancias. Por ejemplo, si las probabilidades $p(x|t,d)$ para $(k=6)$ son $(0.390, 0.225, 0.157, 0.077, 0.076, 0.075)$ y $freq(t,d)=7$, tras realizar el producto $freq(t,d)*p(x|t,d)$ y redondear al entero más cercano, se tiene $(3,2,1,1,1,1)$, es decir, dos instancias del término de más (9 en lugar de 7). Por lo tanto, se puede eliminar una instancia de cada uno de los dos documentos asociados a las temáticas menos probables, quedando así $(3,2,1,1,0,0)$. Por el contrario, si usando las mismas probabilidades, se tiene que $freq(t,d)=3$, se obtiene $(1,1,0,0,0,0)$ perdiéndose en este caso una instancia del término. En este caso, se añade una instancia al documento asociado a la temática más probable, quedando así, $(2,1,0,0,0,0)$. El algoritmo que genera los subdocumentos asociados a un documento d se muestran en la Figura 6.1.

```
for each document d {
  for each term t in d {
    sr=0
    for each topic x {
      s[x]=freq(t,d)*p(x|t,d)
      r[x]=round(s[x])
      sr+=r[x]
    }
    if (sr<freq(t,d))
      add 1 to r[x] for the freq(t,d)-sr topics x having
      greater values of s[x]
    else if (sr>freq(t,d))
           substract 1 from r[x] for the sr-freq(t,d) topics
           x having smaller values of s[x] but having r[x]>0
    for each topic x
      add r[x] instances of term t to the subdocument of d
      associated to topic x
   }
}
```

Figura 6.1: Algoritmo para generar subdocumentos a partir de las probabilidades $p(x|t,d)$.

6.2.2 Unión de subdocumentos para obtener subperfiles homogéneos

Tras la aplicación del algoritmo de la Figura 6.1 a la colección de documentos, cada documento d queda dividido en un conjunto de como máximo k subdocumentos (puede ocurrir que la probabilidad $p(x|d)$ de una temática dado un documento d sea cero, por lo tanto el subdocumento

correspondiente estará vacío), cada uno de ellos asociado a una temática específica y que contiene los términos más relevantes de d para dicha temática. Dicho esto, el tercer paso es el de usar esos subdocumentos para construir los subperfiles asociados a cada diputado.

Sea $\mathscr{D}_i$ el conjunto de documentos de la colección $\mathscr{D}$ asociados al candidato i, $\mathscr{D}_i = \{d_{i1}, \ldots, d_{in_i}\}$. A su vez, cada documento $d_{ij} \in \mathscr{D}_i$ se separa en subdocumentos $d_{ij1}, \ldots, d_{ijk}$ ($d_{ij} = \cup_{l=1}^{k} d_{ijl}$), donde cada d_{ijl} está asociado a una temática x_l. El subperfil, por tanto, del MP i asociado a la temática x_l, $\mathscr{S}(i,l)$, se construye como

$$\mathscr{S}(i,l) = \cup_{j=1}^{n_i} d_{ijl} \tag{6.2}$$

Es decir, se van concatenando todos los términos de los subdocumentos correspondientes a la temática x_l asociadas a todos los documentos del MP i.

Cuando este método se lleva a la práctica, se observa que tiende a generar un número relativamente alto de subperfiles para cada uno de los diputados, a pesar de que este número debe ser menor que k. Cabe destacar que, en el momento en que se genera un subdocumento que no esté vacío para una temática x_l a partir de un documento asociado al diputado i, este diputado va a tener un subperfil para la temática x_l. Esto puede provocar en consecuencia, que algunos de estos subperfiles carezcan de relevancia en el sentido en que van a estar compuestos de un número muy reducido de términos, y esto puede resultar problemático cuando esos subperfiles se indexen para construir un Sistema de Recuperación de Información, que es en definitiva el objetivo primordial. Por tanto, en la siguiente sección, se muestra el estudio de diversos métodos basados en el cálculo de distancias y similitudes entre subperfiles, para reducir el número de temáticas asociadas a cada documento.

6.2.3 Selección del número óptimo de subdocumentos

Los distintos valores para establecer el número de temáticas, k, que se han considerado, $k = 24, 70, 300$, son demasiado altos en algunos casos, puesto que pretender dividir un único documento en ese número de partes no es muy realista a efectos prácticos, ya que se pueden dar problemas como la generación de subperfiles con un único termino en todos los casos, o con un número muy pequeño, los cuales no serían nada informativos. En este sentido, si la probabilidad $p(x|d) > 0$ entonces, por baja que sea, el documento d se va a distribuir proporcionalmente sobre el subdocumento que genera la temática x. Por este motivo, con el objetivo de mejorar esta aproximación, se ha propuesto un método capaz de seleccionar un subconjunto con las temáticas más probables para obtener una mejor distribución de los términos de d en los distintos subdocumentos.

De esto modo, el problema se puede enunciar de la siguiente forma: dada una distribución de probabilidad sobre k temáticas, $p = (p_1, p_2, \ldots, p_k)$, con $p_1 \geq p_2 \geq \ldots \geq p_k$, se pretende encontrar el mejor índice, i_p, por el cual, las temáticas seleccionadas en el conjunto $x_1, x_2, \ldots, x_{i_p}$, sean las temáticas más probables. Dicho esto, cabe mencionar que solo existen k soluciones posibles, $i_p =$

$1,2,\ldots,k$, las cuales están asociadas al vector $I_1 = (1,0,0,\ldots,0)$ (seleccionando solo la temática más probable), $I_2 = (1,1,0,\ldots,0)$ (seleccionando las dos temáticas más probables), $I_3 = (1,1,1,0,\ldots,0)$, hasta alcanzar $I_k = (1,1,\ldots,1)$ (seleccionando todos las temáticas).

De esta manera, se puede formular el problema como el de encontrar el vector

$$I_{i_p} = \arg\min_{I_j} Dist(p, I_j)$$

donde $Dist$ es una medida de distancia entre los vectores p y I_j. O, de forma alternativa,

$$I_{i_p} = \arg\max_{I_j} Sim(p, I_j)$$

donde Sim es una medida de similitud entre los vectores p y I_j.

Por tanto, se van a considerar algunas propuestas basadas en diferentes funciones de distancia y similitud para resolver el problema en cuestión. Suponiendo que se tiene dos vectores k-dimesionales $w_1 = (w_{11}, w_{21}, \ldots, w_{k1})$ y $w_2 = (w_{12}, w_{22}, \ldots, w_{k2})$, en [25], se definen diferentes formas de calcular la distancia o similitud entre esos dos vectores.

Se puede utilizar la medida de similitud Coseno.

$$Cos(w_1, w_2) = \frac{\sum_{i=1}^{k} w_{i1} w_{i2}}{\sqrt{\sum_{i=1}^{k} w_{i1}^2} \sqrt{\sum_{i=1}^{k} w_{i2}^2}}$$

la cual, adaptada para este caso, sería

$$Cos(p, I_j) = \frac{\sum_{i=1}^{j} p_i}{\sqrt{j} \sqrt{\sum_{i=1}^{k} p_i^2}}$$

Y puesto qué, el valor de la expresión $\sqrt{\sum_{i=1}^{k} p_i^2}$ es siempre constante para todos los I_j, la solución en este caso, sería

$$I_{i_p} = \arg\max_{I_j} \frac{\sum_{i=1}^{j} p_i}{\sqrt{j}} \tag{6.3}$$

Otra medida de similitud que se ha tenido en cuenta ha sido el coeficiente de Dice, el cual se define como

$$Dic(w_1, w_2) = \frac{2 \sum_{i=1}^{k} w_{i1} w_{i2}}{\sum_{i=1}^{k} w_{i1}^2 + \sum_{i=1}^{k} w_{i2}^2}$$

la cual, adaptada para este caso, sería

$$Dic(p, I_j) = \frac{2 \sum_{i=1}^{j} p_i}{j + \sum_{i=1}^{k} p_i^2}$$

Por tanto, la solución al problema aplicando este método, se definiría como

$$I_{i_p} = \arg\max_{I_j} \frac{\sum_{i=1}^{j} p_i}{j + \sum_{i=1}^{k} p_i^2} \tag{6.4}$$

También se puede usar el índice de similitud de Jaccard, el cual se define como

$$Jac(w_1,w_2)=\frac{\sum_{i=1}^{k} w_{i1}w_{i2}}{\sum_{i=1}^{k} w_{i1}^2+\sum_{i=1}^{k} w_{i2}^2-\sum_{i=1}^{k} w_{i1}w_{i2}}$$

En este caso, esta medida de similitud se transforma en

$$Jac(p,I_j)=\frac{\sum_{i=1}^{j} p_i}{j+\sum_{i=1}^{k} p_i^2-\sum_{i=1}^{j} p_i}$$

Se puede comprobar que, aunque $Dic(p,I_j)$ y $Jac(p,I_j)$ devuelven diferentes valores, los rankings que generan estas medidas sobre los vectores I_j son siempre idénticos y por lo tanto se genera la misma solución con ambos métodos.

Por otro lado, la medida de similitud de Czekanowski que se define como

$$Cze(w_1,w_2)=\frac{2\sum_{i=1}^{k}\min(w_{i1},w_{i2})}{\sum_{i=1}^{k}(w_{i1}+w_{i2})}$$

En este caso, esta medida de similitud se transforma en

$$Cze(p,I_j)=\frac{2\sum_{i=1}^{j} p_i}{j+1}$$

Y la solución en este caso, sería

$$I_{i_p}=\arg\max_{I_j}\frac{\sum_{i=1}^{j} p_i}{j+1} \tag{6.5}$$

La similaridad de Ruzicka definida como

$$Ruz(w_1,w_2)=\frac{\sum_{i=1}^{k}\min(w_{i1},w_{i2})}{\sum_{i=1}^{k}\max(w_{i1},w_{i2})}$$

Que se transforma para este problema en

$$Ruz(p,I_j)=\frac{\sum_{i=1}^{j} p_i}{j+1-\sum_{i=1}^{j} p_i}$$

Al igual que en el caso de las medidas de similitud de Dice y Jaccard, se ha comprobado que los rankings generados por $Cze(p,I_j)$ y $Ruz(p,I_j)$ sobre los vectores I_j son siempre los mismos, y por lo tanto, generan la misma solución.

Por último, se ha considerado como medida de similitud el coeficiente de Overlap, que se define como

$$Ove(w_1,w_2)=\frac{\sum_{i=1}^{k} w_{i1}w_{i2}}{\min(\sum_{i=1}^{k} w_{i1},\sum_{i=1}^{k} w_{i2})}$$

El cual, en este caso se reduce a

$$Ove(p,I_j)=\sum_{i=1}^{j} p_i$$

Coeficiente que siempre obtiene el valor máximo cuando $j = k$. En este caso, por tanto, la solución siempre va a ser $I_{i_p} = I_k$, es decir, usando todas las temáticas. Este método puede ser un buen caso base para observar las diferencias entre seleccionar un número óptimo de temáticas o seleccionarlas todas.

En lo referente a las medidas de distancia que se han utilizado, la primera de ellas es la clásica distancia Euclídea

$$Euc(w_1, w_2) = \sqrt{\sum_{i=1}^{k} (w_{i1} - w_{i2})^2}$$

En este caso, se obtiene

$$Euc(p, I_j) = \sqrt{j + \sum_{i=1}^{k} p_i^2 - 2\sum_{i=1}^{j} p_i}$$

Para esta medida de distancia es sencillo probar que $\forall j = 1, \ldots, k-1, Euc(p, I_j) \leq Euc(p, I_{j+1})$. Por tanto, la mejor solución será siempre $I_{i_p} = I_1$, es decir, solo usar la temática más probable. De la misma forma, este método también sería interesante considerarlo como un caso base para observar las diferencias entre seleccionar la temáticas más probable o seleccionar un conjunto óptimo. De esta forma, ambos casos extremos para la solución del problema quedarían cubiertos.

También se puede utilizar la medida de distancia de Hamming (distancia Manhattan)

$$Ham(w_1, w_2) = \sum_{i=1}^{k} |w_{i1} - w_{i2}|$$

En este caso, se obtendría

$$Ham(p, I_j) = \sum_{i=1}^{j} (1 - p_i) + \sum_{i=j+1}^{k} p_i = j + 1 - 2\sum_{i=1}^{j} p_i$$

Que como en el caso de la distancia Euclídea, se puede observar que $Ham(p, I_j) \leq Ham(p, I_{j+1})$ y, por tanto la solución obtenida por la distancia de Hamming es $I_{i_p} = I_1$, al igual que con la distancia Euclídea.

La distancia de Chebyshev definida como

$$Che(w_1, w_2) = \max_{i=1}^{k} |w_{i1} - w_{i2}|$$

Adaptada para este caso de estudio como

$$Che(p, I_j) = \max(\max_{i=1}^{j}(1 - p_i), \max_{i=j+1}^{k}(p_i)) = 1 - p_j$$

Que una vez más, la distancia mínima se alcanza cuando $j = 1$, es decir, la solución sería $I_{i_p} = I_1$.

También se ha utilizado la medida de distancia de Sorensen, que se define como

$$Sor(w_1, w_2) = \frac{\sum_{i=1}^{k} |w_{i1} - w_{i2}|}{\sum_{i=1}^{k} (w_{i1} + w_{i2})}$$

Que en este caso se reduce a

$$Sor(p,I_j) = \frac{j+1-2\sum_{i=1}^{j} p_i}{j+1} = 1 - \frac{2\sum_{i=1}^{j} p_i}{j+1}$$

Y esta medida es igual a $1 - Cze(p,I_j)$, donde Cze es la medida de similitud Czekanowski considerada anteriormente. Por tanto, la distancia de Sorensen, la cual será utilizada como representante, genera los mismos resultados que $Cze(p,I_j)$.

La distancia de Soergel

$$Soe(w_1,w_2) = \frac{\sum_{i=1}^{k} |w_{i1} - w_{i2}|}{\sum_{i=1}^{k} \max(w_{i1},w_{i2})}$$

En este caso se obtiene

$$Soe(p,I_j) = \frac{j+1-2\sum_{i=1}^{j} p_i}{j+1-\sum_{i=1}^{j} p_i} = 1 - \frac{\sum_{i=1}^{j} p_i}{j+1-\sum_{i=1}^{j} p_i}$$

Y esta medida es igual que $1 - Ruz(p,I_j)$, donde Ruz es la medida de similitud Ruzicka que se ha considerado previamente. Por tanto, la medida de distancia Soergel genera las mismas soluciones que Ruzicka, que Czekanowski y que Sorensen.

La medida de distancia de Kulczynski definida como

$$Kul(w_1,w_2) = \frac{\sum_{i=1}^{k} |w_{i1} - w_{i2}|}{\sum_{i=1}^{k} \min(w_{i1},w_{i2})}$$

En este caso

$$Kul(p,I_j) = \frac{j+1-2\sum_{i=1}^{j} p_i}{\sum_{i=1}^{j} p_i} = \frac{j+1}{\sum_{i=1}^{j} p_i} - 2$$

Cabe destacar que, $Kul(p,I_j) = 2(1/Cze(p,I_j) - 1)$. Y como esta expresión es una función monótona decreciente de la similitud de Czekanowski, la medida de distancia de Kulczynski genera las misma solución que Czekanowski.

La medida de distancia de Camberra

$$Cam(w_1,w_2) = \sum_{i=1}^{k} \frac{|w_{i1} - w_{i2}|}{w_{i1} + w_{i2}}$$

Adapatada como

$$Cam(p,I_j) = \sum_{i=1}^{j} \frac{1-p_i}{1+p_i} + \sum_{i=j+1}^{k} \frac{p_i}{p_i} = \sum_{i=1}^{j} \frac{1-p_i}{1+p_i} + k - j$$

De la cual se puede comprobar fácilmente que $\forall j = 1,\ldots,k-1$, $Cam(p,I_j) \geq Cam(p,I_{j+1})$, y por tanto, la medida de distancia de Camberra obtiene el valor mínimo en $j = k$ y la solución siempre sería $I_{i_p} = I_k$.

La distancia de divergencia

$$Div(w_1,w_2) = 2\sum_{i=1}^{k} \frac{(w_{i1} - w_{i2})^2}{(w_{i1} + w_{i2})^2}$$

Y para este caso concreto

$$Div(p,Ij) = 2\sum_{i=1}^{j}\frac{(1-p_i)^2}{(1+p_i)^2} + 2(k-j)$$

Que como en el caso anterior, se puede observar que $\forall j = 1,\ldots,k-1, Div(p,I_j) \geq Div(p,I_{j+1})$ y por lo tanto la solución es otra vez $I_{i_p} = I_k$.

Finalmente, la última medida distancia que se ha utilizado es la distancia de Neyman, la cual queda definida como

$$Ney(w_1,w_2) = \sum_{i=1}^{k}\frac{(w_{i1}-w_{i2})^2}{w_{i1}}$$

Y en este caso quedaría como

$$Ney(p,I_j) = \sum_{i=1}^{j}\frac{(1-p_i)^2}{p_i} + \sum_{i=j+1}^{k}\frac{p_i^2}{p_i} = \sum_{i=1}^{j}\frac{1}{p_i} + 1 - 2j$$

De la misma forma se puede demostrar que $\forall j = 1,\ldots,k-1$, $Ney(p,I_j) \leq Ney(p,I_{j+1})$, y una vez más, la solución en este caso sería siempre $I_{i_p} = I_1$.

Aunque se ha referido el uso de múltiples medidas de distancia y similitud (15 en total), solo se han encontrado cinco soluciones distintas de entre todas ellas. Una de ellas representa el caso en el que se considera como solución utilizar todos las temáticas y se va a denominar Overlap ($I_{i_p} = I_k$). El otro caso extremo, tal y como se ha dicho anteriormente, es la distancia Euclídea, la cual siempre sugiere el uso únicamente de la temática más probable ($I_{i_p} = I_1$). Los otros tres casos se corresponden con la medida de similitud Coseno ($I_{i_p} = \arg\max_{I_j}\frac{\sum_{i=1}^{j}p_i}{\sqrt{j}}$), la medida de similitud Dice (($I_{i_p} = \arg\max_{I_j}\frac{\sum_{i=1}^{j}p_i}{j+\sum_{i=1}^{k}p_i^2}$) y la medida de distancia Sorensen ($I_{i_p} = \arg\max_{I_j}\frac{\sum_{i=1}^{j}p_i}{j+1}$). De forma experimental se ha comprobado que Dice es un método más selectivo a la hora de determinar el número óptimo de temáticas que Sorensen (menos temáticas), siendo el método basado en la medida de similitud Coseno el más permisivo (más temáticas). Obviamente, la distancia Euclídea es el método más restrictivo. Por ilustrar la situación con un ejemplo, para un valor de $k = 5$ y considerando la distribución de probabilidad (0.50,0.29,0.19,0.01,0.01). Usando la similitud Coseno, se seleccionarían las tres primeras temáticas ($i_p = 3$), mientras que con Sorensen se seleccionarían solo las dos primeras ($i_p = 2$), y Dice, al igual que la distancia Euclídea, seleccionarían únicamente la primera temática más probable ($i_p = 1$). Por supuesto, en este ejemplo, el método Overlap seleccionaría todas las temáticas ($i_p = 5$).

Dicho esto, se van a utilizar únicamente estas cinco alternativas como estrategias para distribuir los términos en el proceso de experimentación. Con el objetivo de poder observar de forma más precisa cómo los métodos que se han considerado desempeñan su labor en el contexto político del estudio, en la Figura 6.2 se muestra el número medio de subdocumentos generados por cada método para cada documento asociado a un Miembro del Parlamento. Del mismo modo, en la Tabla 6.1 la media de todos los subdocumentos, el máximo y el mínimo número de subdocumentos generados para cada documento de cada Miembro del Parlamento. Estos datos se han obtenido

a partir de aplicar el algoritmo LDA con $k = 70$ temáticas. En el caso del método Overlap, aún cuando se deberían seleccionar todas las temáticas, es decir 70, los valores cambian puesto que no se han considerado las temáticas cuya probabilidad $p(x|d)$ sea igual a cero.

Finalmente, se presentan las demostraciones que evidencian con que estrategias, de las que se han planteado en esta sección, se alcanzan las mismas soluciones.

En primer lugar, la demostración de que las medidas de similitud **Dice** y **Jaccard** devuelven la misma solución.

$$Dic(p,I_j) \leq Dic(p,I_{j+r}) \Leftrightarrow Jac(p,I_j) \leq Jac(p,I_{j+r})$$

De este modo se tiene, por un lado:

$Dic(p,I_j) \leq Dic(p,I_{j+r}) \Leftrightarrow \frac{2\sum_{i=1}^{j} p_i}{j+\sum_{i=1}^{k} p_i^2} \leq \frac{2\sum_{i=1}^{j} p_i + 2\sum_{i=j+1}^{r} p_i}{j+r+\sum_{i=1}^{k} p_i^2} \Leftrightarrow$
$2(j+\sum_{i=1}^{k} p_i^2)\sum_{i=1}^{j} p_i + 2r\sum_{i=1}^{j} p_i \leq 2(j+\sum_{i=1}^{k} p_i^2)\sum_{i=1}^{j} p_i +$
$2(j+\sum_{i=1}^{k} p_i^2)\sum_{i=j+1}^{r} p_i \Leftrightarrow r\sum_{i=1}^{j} p_i \leq (j+\sum_{i=1}^{k} p_i^2)\sum_{i=j+1}^{r} p_i$

Y por otro lado:

$Jac(p,I_j) \leq Jac(p,I_{j+r}) \Leftrightarrow \frac{\sum_{i=1}^{j} p_i}{j+\sum_{i=1}^{k} p_i^2-\sum_{i=1}^{j} p_i} \leq \frac{\sum_{i=1}^{j} p_i+\sum_{i=j+1}^{r} p_i}{j+r+\sum_{i=1}^{k} p_i^2-\sum_{i=1}^{j} p_i-\sum_{i=j+1}^{r} p_i}$
$\Leftrightarrow \sum_{i=1}^{j} p_i(j+\sum_{i=1}^{k} p_i^2-\sum_{i=1}^{j} p_i) + r\sum_{i=1}^{j} p_i - \sum_{i=1}^{j} p_i \sum_{i=j+1}^{r} p_i \leq$
$\sum_{i=1}^{j} p_i(j+\sum_{i=1}^{k} p_i^2-\sum_{i=1}^{j} p_i)+(j+\sum_{i=1}^{k} p_i^2)\sum_{i=j+1}^{r} p_i - \sum_{i=1}^{j} p_i \sum_{i=j+1}^{r} p_i \Leftrightarrow r\sum_{i=1}^{j} p_i \leq (j+\sum_{i=1}^{k} p_i^2)\sum_{i=j+1}^{r} p_i$

Y de esta forma se puede observar que las condiciones son las mismas en ambos casos.

Demostración de que las medidas de similitud **Czekanowski** y **Ruzicka** devuelven la misma solución.

$$Cze(p,I_j) \leq Cze(p,I_{j+r}) \Leftrightarrow Ruz(p,I_j) \leq Ruz(p,I_{j+r})$$

De este modo se tiene, por un lado:

$Cze(p,I_j) \leq Cze(p,I_{j+r}) \Leftrightarrow \frac{2\sum_{i=1}^{j} p_i}{j+1} \leq \frac{2\sum_{i=1}^{j} p_i + 2\sum_{i=j+1}^{r} p_i}{j+1+r} \Leftrightarrow$
$2(j+1)\sum_{i=1}^{j} p_i + 2r\sum_{i=1}^{j} p_i \leq 2(j+1)\sum_{i=1}^{j} p_i + 2(j+1)\sum_{i=j+1}^{r} p_i \Leftrightarrow$
$r\sum_{i=1}^{j} p_i \leq (j+1)\sum_{i=j+1}^{r} p_i$

Y por otro lado:

$Ruz(p,I_j) \leq Ruz(p,I_{j+r}) \Leftrightarrow \frac{\sum_{i=1}^{j} p_i}{j+1-\sum_{i=1}^{j} p_i} \leq \frac{\sum_{i=1}^{j} p_i+\sum_{i=j+1}^{r} p_i}{j+1+r-\sum_{i=1}^{j} p_i-\sum_{i=j+1}^{r} p_i} \Leftrightarrow$
$\sum_{i=1}^{j} p_i(j+1-\sum_{i=1}^{j} p_i) + r\sum_{i=1}^{j} p_i - \sum_{i=1}^{j} p_i \sum_{i=j+1}^{r} p_i \leq$
$\sum_{i=1}^{j} p_i(j+1-\sum_{i=1}^{j} p_i) + (j+1)\sum_{i=j+1}^{r} p_i - \sum_{i=1}^{j} p_i \sum_{i=j+1}^{r} p_i \Leftrightarrow$
$r\sum_{i=1}^{j} p_i \leq (j+1)\sum_{i=j+1}^{r} p_i$

Y así se puede ver que las condiciones son las mismas.

Demostración para la distancia **Euclídea**,

$$\forall j = 1,\ldots,k-1, Euc(p,I_j) \leq Euc(p,I_{j+1})$$

de cómo $\sum_{i=1}^{k} p_i^2$ es un factor constante en la expresión de,

$$Euc(p,I_j) = \sqrt{j + \sum_{i=1}^{k} p_i^2 - 2\sum_{i=1}^{j} p_i},$$

sólo se tiene que demostrar que $j - 2\sum_{i=1}^{j} p_i \leq j+1-2\sum_{i=1}^{j+1} p_i$. Y esto es cierto sí y sólo sí $2p_{j+1} \leq 1$, es decir $p_{j+1} \leq 1/2$. Y esto es siempre cierto porque $p_{j+1} \leq p_j$.

Demostración para la distancia de **Hamming**,

$$\forall j = 1,\ldots,k-1, Ham(p,I_j) \leq Ham(p,I_{j+1})$$

de cómo $Ham(p,I_j) = j + 1 - 2\sum_{i=1}^{j} p_i$, entonces $Ham(p,I_j) \leq Ham(p,I_{j+1}) \Leftrightarrow$
$j + 1 - 2\sum_{i=1}^{j} p_i \leq j + 1 + 1 - 2\sum_{i=1}^{j+1} p_i \Leftrightarrow 2p_{j+1} \leq 1.$

Esto es siempre cierto porque $p_{j+1} \leq p_j$.

Demostración para la distancia **Camberra**,

$$\forall j = 1,\ldots,k-1, Cam(p,I_j) \geq Cam(p,I_{j+1})$$

$Cam(p,I_j) \geq Cam(p,I_{j+1}) \Leftrightarrow \sum_{i=1}^{j} \frac{1-p_i}{1+p_i} + k - j \geq$
$\sum_{i=1}^{j} \frac{1-p_i}{1+p_i} + \frac{1-p_{j+1}}{1+p_{j+1}} + k - j - 1 \Leftrightarrow 1 \geq \frac{1-p_{j+1}}{1+p_{j+1}} \Leftrightarrow p_{j+1} \geq 0.$

Demostración para la distancia de **divergencia**,

$$\forall j = 1,\ldots,k-1, Div(p,I_j) \geq Div(p,I_{j+1})$$

$Div(p,I_j) \geq Div(p,I_{j+1}) \Leftrightarrow 2\sum_{i=1}^{j} \frac{(1-p_i)^2}{(1+p_i)^2} + 2(k-j) \geq$
$2\sum_{i=1}^{j} \frac{(1-p_i)^2}{(1+p_i)^2} + 2\frac{(1-p_{j+1})^2}{(1+p_{j+1})^2} + 2(k-j) - 2 \Leftrightarrow 1 \geq \frac{(1-p_{j+1})^2}{(1+p_{j+1})^2}.$

La últime desigualdad es obviamente cierta.

Demostración para la distancia de **Neyman**,

$$\forall j = 1,\ldots,k-1, Ney(p,I_j) \leq Ney(p,I_{j+1})$$

$Ney(p,I_j) \leq Ney(p,I_{j+1}) \Leftrightarrow \sum_{i=1}^{j} \frac{1}{p_i} + 1 - 2j \leq \sum_{i=1}^{j} \frac{1}{p_i} + \frac{1}{p_{j+1}} + 1 - 2j - 2 \Leftrightarrow 2 \leq \frac{1}{p_{j+1}} \Leftrightarrow p_{j+1} \leq 1/2$, la cual es siempre cierta porque $p_{j+1} \leq p_j$.

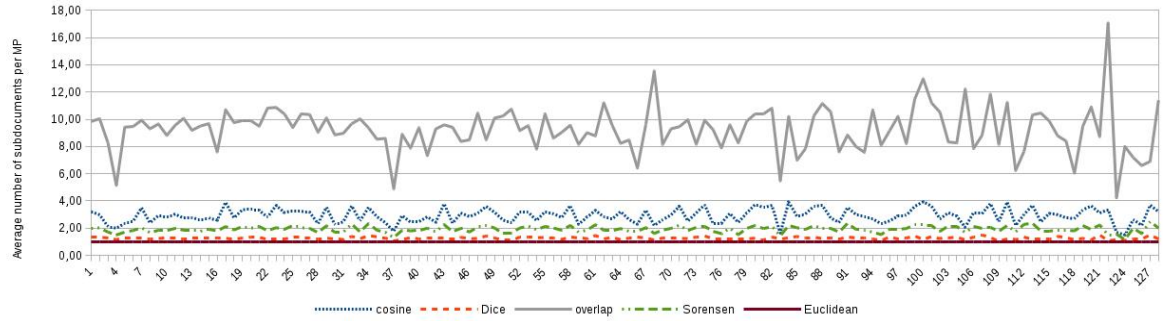

Figura 6.2: Número medio de subdocumentos generados a partir de cada uno de los documentos asociados a cada MP, usando los cinco métodos ($k = 70$).

Tabla 6.1: Promedios de la media, máximo y mínimo número de subdocumentos generados a partir de todos los documentos asociados a cada MP ($k = 70$).

	media	max	min
Overlap	9.25	18.66	2.39
Cosine	2.93	7.27	1.0
Sorensen	1.92	4.16	1.0
Dice	1.26	2.66	1.0
Euclidean	1.0	1.0	1.0

6.2.4 Construcción del número óptimo de subdocumentos

Puesto que el algoritmo de la Figura 6.1 no contempla los métodos para obtener el número óptimo de temáticas en las que distribuir los términos de un documento, este debe ser modificado en función del método que se vaya a utilizar. Es importante destacar que esta selección de temáticas está basada en la distribución de probabilidad de las temáticas en un documento. Pero, en este estudio existe la posibilidad de usar tanto la distribución de temáticas por documento, $p(x|d)$, generada por LDA, como la distribución de temáticas por documento y término $p(x|t,d)$, es decir, seleccionar las temáticas tanto a nivel de documento como a nivel de documento y término. En la experimentación se ha llevado a cabo el proceso de generación de subperfiles teniendo en cuenta las distribución $p(x|t,d)$, pero se han obtenido unos malos resultados. La razón de esto es que los términos más representativos de las temáticas menos probables, es decir aquellos con un alto $p(x|t,d)$ pero un bajo $p(x|d)$, generan subdocumentos con muy pocos términos y, en consecuencia, subperfiles muy pequeños y muy poco informativos.

Si se usa la distribución $p(x|d)$ para seleccionar el número óptimo de temáticas y sus correspondientes subdocumentos, solo es necesario reemplazar la probabilidad original $p(x|t,d)$ en la Ecuación (6.6) por $p_o(x|t,d)$, de forma que:

$$p_o(x_i|t,d) = \frac{p(t|x_i)p_n(x_i|d)}{\sum_{j=1}^{k} p(t|x_j)p_n(x_j|d)} \tag{6.6}$$

donde $p_n(x_i|d)$ está definido como

$$p_n(x_i|d) = \begin{cases} \frac{p(x_i|d)}{\sum_{j=1}^{i_d} p(x_j|d)} & i = 1,\ldots,i_d \\ 0 & i = i_d+1,\ldots,k \end{cases} \tag{6.7}$$

donde i_d representa el número de las temáticas más probables según el método de selección que se haya considerado, empezando por las probabilidades $p(x|d)$. En este proceso, simplemente se han vuelto a calcular las probabilidades en la Ecuación (6.1) pero usando una versión normalizada, $p_n(x|d)$, de $p(x|d)$ y considerando solo y exclusivamente las temáticas seleccionadas. En este caso, se va a asumir que un documento solo se va a distribuir entre un número pequeño de temáticas, y que las instancias de los términos que originalmente se asignarían a otras temáticas y que después se eliminarían, son reasignadas a las temáticas restantes. De este modo, cada documento d esta asociado únicamente a sus i_d temáticas más probables (de acuerdo a $p(x|d)$) y cada uno de los subdocumentos está asociado exactamente a una de esas temáticas.

Por tanto, si se combinan las Ecuaciones (6.6) and (6.7) se puede escribir $p_o(x_i|t,d)$ como

$$p_o(x_i|t,d) = \begin{cases} \frac{p(t|x_i)p(x_i|d)}{\sum_{j=1}^{i_d} p(t|x_j)p(x_j|d)} & i = 1,\ldots,i_d \\ 0 & i = i_d+1,\ldots,k \end{cases} \tag{6.8}$$

Procediendo de esta forma, los subdocumentos asociados a las temáticas que se han seleccionado no solo van a contener los términos propios de la temática, sino también los términos de las temáticas que se han descartado. Una opción podría ser la de descartar también los términos de las temáticas que se hayan descartado y no incluirlos en ningún subdocumento pero, en ese caso, la unión de todos los subdocumentos no daría como resultado el documento original lo que conllevaría a un pérdida de información que no se produciría en el caso base donde se seleccionan todas las temáticas. No obstante esta experimentación se ha llevado a cabo pero, a la vista de los resultados y por la razón previamente mencionada se ha considerado que esos experimentos no sean tenidos en cuenta.

Para ilustrar cómo se construyen los subperfiles se va a proponer el siguiente ejemplo: uno de los documentos contiene una intervención donde un diputado, miembro de la Comisión de Vivienda, habla sobre viviendas destinadas a ciertos grupos sociales, por ejemplo, protección oficial, personas en riesgo de exclusión social o víctimas de violencia de género. El discurso del parlamentario, por tanto, va a contener términos representativos como $mujer$, $abuso$, etc. y esos términos en concreto son característicos de la temática relacionada con *Igualdad de Género*. Esta intervención por tanto se podría particionar en al menos dos temáticas: *Vivienda* e *Igualdad de Género*. No obstante, es este caso, las referencias a la violencia de género son marginales en el sentido en que la temática principal de la intervención está relacionada con la vivienda. Tras usar este método para reducir el número de temáticas que se pueden asociar a un documento, solo se va a mantener como temática dominante *Vivienda* y por tanto, los términos $mujer$, $abuso$, etc. van a permanecer asociados a esa temática dominante. De esta

forma, como el diputado es miembro de la Comisión de Vivienda, cabe esperar que su subperfil esté compuesto principalmente por términos relacionado con vivienda y con relativamente pocos términos marginales. Y, en consecuencia, si un usuario lanza una consulta contra el sistema sobre violencia de género, cabe esperar que este diputado en cuestión no sea uno de los más relevantes a la consulta si lo comparamos con otros diputados que traten específicamente esta temática, mientras que podrían considerarse erróneamente más relevantes si estas palabras se asignaran a un subperfil específico y más pequeño que trate sobre *Igualdad de Género*. Esto ilustra la importancia de seleccionar únicamente aquellas temáticas que se presentan de forma predominante en cada documento. No obstante, en el caso de que la consulta fuese más específica, por ejemplo si la consulta esta relacionada con las viviendas que se facilitan a las víctimas de violencia de género, entonces tendría sentido que el diputado que se propone como ejemplo fuese uno de los más relevantes a dicha consulta, puesto que ambas temáticas están combinadas en su subperfil. En ese caso, cualquier término que no sea específico de la temática pero que esté incluido en el subperfil va a servir para agregar contexto al resto de términos para que estos encajen mejor en relación a la consulta. Si se descartan esos términos, el sistema va a perder la capacidad de encontrar la relevancia del diputado del ejemplo para consultas más precisas. En otras palabras, los términos añadidos a un subperfil, que no son propios de la temática a la que están asociados no afectan de forma negativa al rendimiento del sistema sino que, por el contrario, ayudan a mejorarlo.

6.3 Evaluación Experimental

El objetivo principal de este capítulo es el de determinar de forma empírica si la creación de subperfiles de términos basados en la extracción de temáticas de la colección a partir de la aplicación del algoritmo LDA, es una buena opción para mejorar el rendimiento de los Sistemas de Búsqueda de Expertos. Al igual que en procesos experimentales de esta Tesis Doctoral, se va a considerar la aplicación real sobre un contexto parlamentario. Del mismo modo, se va a utilizar la colección de intervenciones parlamentarias procesada tal y como se define en la Sección 3.3 y usando como medidas de evaluación las tres medidas clásicas en Recuperación de Información; *precision* y NDCG, explicadas también en la Sección 3.3, sobre los diez primeros elementos del ranking que se obtenga con la consulta. Y por otro lado, la medida de *recall* sobre el número total de MPs relevantes para cada consulta. A diferencia de en experimentos anteriores, en este caso no se ha utilizado el recall sobre los diez primeros elementos del ranking puesto que en muchos casos existen más de diez MPs relevantes para una consulta. También cabe destacar que considera la métrica con el número total de elementos relevantes, tanto el valor de *recall* como el de *precision* son el mismo.

6.3.1 Implementación de LDA

La implementación de LDA que se ha utilizado para el proceso de experimentación de este capítulo ha sido la que se recoge en el paquete $topicmodels$ en R. La función que llama a la ejecución del algoritmo necesita, entre otros parámetros, el número de temáticas que se quiere extraer de la colección, es decir k. Para ello se ha considerado tres formas distintas de establecer el valor de ese parámetro:

- $k = m * n/t$: método clásico [23] que se aplica en el estado del arte para clustering de colecciones de documentos y que ya se ha utilizado en anteriores experimentos de esta tesis. Para más detalle, m se corresponde con el número de términos únicos en la colección ($m = 4208$), n es el valor para el número de intervenciones de todos los diputados ($n = 10025$, 80% del total del número de intervenciones), y t que toma el valor del número de entradas en la matriz de documentos/términos distintas de cero ($t = 1,702,296$). Y por tanto, el valor de k es 24.

- $k = \sqrt{n/2}$: otra aproximación clásica para distribuir las instancias de un conjunto de datos en grupos de forma homogénea [72] y que también se ha utilizado en experimentos anteriores. Para más detalle, siendo n el número de instancias, documentos en este caso, el valor de $k = 70$.

- $k = 300$: de forma alternativa se utiliza este valor, el cual es más alto que los encontrados habitualmente en la literatura relacionada con LDA [30, 56, 85, 87, 101, 148, 151], para observar cómo se comporta el sistema. Aún así en [53] se puede observar como los valores altos de k obtienen mejores resultados en comparación con valores más pequeños.

Cabe destacar que en este capítulo no se ha utilizado el criterio $\#Com$, tal y como se hacía en el Capítulo 5, por dos motivos: el primero es que el valor de $\#Com$ es muy similar al de $k = m * n/t$ y por tanto no aporta información nueva y relevante y, en segundo lugar, con vistas a aplicar estos métodos en distintos ámbitos, se ha establecido que conocer a priori el número de comisiones es tener información privilegiada de la que no siempre se va disponer, por lo tanto sería injusto usarla aquí. Con respecto a otros parámetros, tal y como sugiere [53], el hiperparámetro α se va a ajustar a $50/k$, donde k es el número de temáticas a extraer de la colección, y el valor de β se va a fijar a 0.1 para todos los experimentos de este capítulo. Cabe destacar que cuanto mayor es el valor del hiperparámetro α, mayor es la posibilidad de que se extraigan más temáticas de un documento. Por otro lado, el valor de β determina la distribución de términos en las diferentes temáticas.

6.3.2 Aproximaciones base a este problema, juicios de relevancia y consultas

Antes de llevar a cabo la experimentación y con el objetivo de realizar una comparativa entre diferentes aproximaciones se van a considerar dos casos base desde los que partir en el proceso

experimental, los cuales son las aproximaciones clásicas en la literatura de Búsqueda de Expertos y que ya se han utilizado en experimentos anteriores.

- **Perfiles monolíticos** ($TermMon$): se construye un perfil único para cada MP el cual contiene el conjunto completo de términos de todas las intervenciones del MP en cuestión.
- **Perfiles de intervenciones** ($TermInt$): por cada MP se tiene que hay un subperfil por cada intervención y en consecuencia cada subperfil va a contener todos los términos de dicha intervención.

Con respecto a los juicios de relevancia que se van a considerar es este experimento, durante todo el desarrollo de esta Tesis se ha determinado que un MP es relevante cuando ha participado activamente en la iniciativa que se utiliza como consulta. Este juicio de relevancia, ciertamente, es muy restrictivo y se ha venido teniendo en cuenta para poner a prueba las virtudes de las diferentes aproximaciones que se han evaluado para observar su comportamiento en los ámbitos más estrictos. En este caso de estudio particular, en lugar de determinar que una consulta es relevante sólo a los individuos que han participado activamente en ella y con el objetivo de evaluar el sistema en un contexto más práctico, se ha tomado como juicio de relevancia que una consulta sea relevante a todos los miembros de la comisión a la que hace referencia la iniciativa que se usa cómo consulta. De este modo, si las diferentes aproximaciones que aquí se proponen funcionan correctamente, sería de esperar que las medidas de NDCG, la cual muestra que los expertos a los que realmente les es relevante la consulta estén a la cabeza del ranking, y la *precision*, que en este caso determina que el sistema ha encontrado a los expertos relevantes, sean más altas en detrimento del *recall*, que al ser menos restrictivos en los juicios de relevancia se verá perjudicado.

Por otro lado, en lo referente a las consultas que se van a utilizar en este experimento, puesto que es un sistema de búsqueda de expertos sólo se van a utilizar las consultas que se venían usando en las aproximaciones de recomendación, es decir, los extractos de las iniciativas. Además, con la intención de realizar consultas más fieles a la realidad, cada una de las iniciativas contiene una serie de materias que determinan en cierto modo las temáticas sobre lo que se va a debatir en ella. Puesto que en el contexto de esta experimentación se presta especial atención a la detección de temáticas con el algoritmo LDA, sería interesante proponer como consultas la unión del extracto de la iniciativa con las materias que se relatan en la iniciativa para poder descubrir así si las temáticas que se aprenden con LDA son similares a las impuestas en las iniciativas que se utilizan como consulta.

6.3.3 Conjunto de entrenamiento, generación de perfiles y Sistema de Recuperación de Información

Si centramos la atención en el conjunto de datos de entrenamiento, en la Figura 6.3 se muestra cómo cada documento de entrada contiene las intervenciones de los Miembros del Parlamento en

una iniciativa, las cuales se han procesado previamente como en casos anteriores, eliminando $stopwords$, realizando un proceso de $stemming$ y suprimiendo todos los términos que aparezca en menos del 1% del total de las intervenciones. Por tanto, se ha utilizado este conjunto de documentos (intervenciones) para entrenar un conjunto de modelos LDA con el objetivo de obtener las matrices de distribución de temáticas en documentos $p(x|d)$ y de términos en temáticas $p(t|x)$.

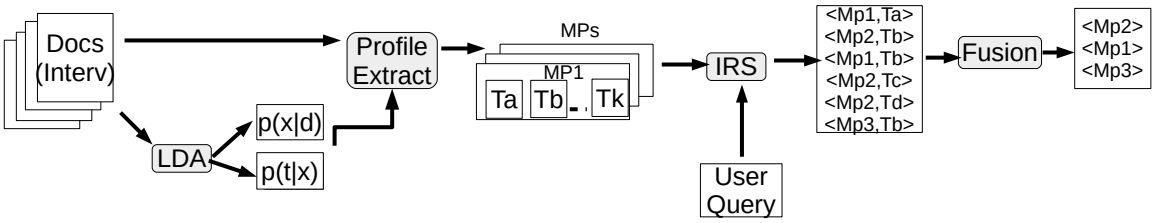

Figura 6.3: Proceso global.

Estas matrices de distribuciones de probabilidad se utilizan para extraer los diferentes subperfiles de los diputados, tal y como se ha explicado en al Sección 6.2 y usando los criterios de reducción de temáticas Euclídea, Dice, Sorensen, Coseno y Overlap. Donde cada uno de los subperfiles que se obtienen por este procedimiento puede ser considerado como un documento ($bag-of-words$) que representa un aspecto particular de los intereses de un diputado. Estos subperfiles se pueden construir de forma más heterogénea, como los generados a partir de usar la distancia Euclídea, o construidos de forma más pura y precisa, como los generados mediante el criterio Overlap.

A continuación, estos perfiles de términos basados en temáticas se utilizan como entrada en la construcción de un Sistema de Recuperación de Información. De forma más específica, dado un MP_i, entonces se indexa cada subperfil $< MP_i, T_j >, j = 1 \ldots k$ utilizando para ello la librería Apache Lucene (https://lucene.apache.org/). De este modo, el Sistema de Recuperación de Información obtiene la utilidad de determinar qué MPs son más relevantes dada una consulta q, en el caso de la Recomendación de Expertos. Con el objetivo de emparejar de algún modo las consultas con estos subperfiles, se ha utilizado el modelo *Language Model*(LM) implementado también en la librería Apache Lucene tanto con el método *Jelinek-Mercer Smoothing* como con el método *Dirichlet Prior Smoothing*. Ciertamente se han llevado a cabo experimentos siguiendo con el modelo BM25, que es que se ha usado hasta ahora, pero los resultados eran bastante similares, por tanto, en este capítulo solo se van a presentar los resultados de LM con el método *Jelinek-Mercer Smoothing*.

6.3.4 Análisis de los efectos de las estrategias de distribución de temáticas

Dependiendo de la forma en la que los términos de cada intervención estén distribuidos entre las diferentes temáticas tiene un impacto directo en los subperfiles que se aprenden, los correspon-

diente subperfiles que posteriormente van a ser indexados por el Sistema de Recuperación de Información y, en consecuencia, en el ranking de MPs que este sistema va a devolver en función de una consulta. Por tanto, en esta sección se va a llevar a cabo un análisis detallado de los resultados obtenidos de los experimentos para su mayor compresión.

- **Análisis del tamaño de los subperfiles**: Puesto que en todas las particiones de datos donde se ha llevado a cabo la experimentación tienen características muy similares y las tendencias se mantienen, a efectos prácticos, solo se va a analizar una única partición de datos, siendo la información extraída de esta válida para el resto de particiones. Dicho esto, para la partición que se ha tomado como muestra para el análisis se van a calcular ciertas estadísticas relacionadas con el tamaño de los subperfiles que esta genera. Con ese propósito, en la Tabla 6.2 se puede observar información con respecto al número de subperfiles totales obtenidos ($\#SP$) para cada una de las estrategias de distribución de temáticas (Euclídea, Dice, Sorensen, Coseno y Overlap) y para cada valor de k, además la tabla muestra el valor medio del número de subperfiles por MP (Avg. #SP) y el número medio de términos en cada subperfil (Avg.SP-size).

Tabla 6.2: Análisis del tamaño de los subperfiles por estrategia de distribución, donde $m * n/t = 24$ y $\sqrt{n/2} = 70$.

k		**Overlap**	**Coseno**	**Sorensen**	**Dice**	**Euclídea**
$m * n/t$	$\#SP$	2987	2147	1919	1664	1438
	Avg. #SP	9.73	6.99	6.25	5.42	4.68
	Avg. SP-size	1573.87	2190.0	2450.48	2822.10	3266.65
	#tinySP	559	172	130	98	57
	#tinySP/#SP	18.71	8.01	6.77	5.89	3.96
$\sqrt{n/2}$	$\#SP$	10772	5774	4340	3254	2762
	Avg. #SP	35.09	18.81	14.14	10.60	9.00
	Avg. SP-size	442.125	819.90	1087.50	1448.28	1704.62
	#tinySP	3696	784	375	196	118
	#tinySP/#SP	31.34	13.58	8.64	6.02	4.27
300	$\#SP$	56122	15644	7486	4182	4122
	Avg. #SP	182.81	50.96	24.38	13.62	13.43
	Avg. SP-size	164.8	381.75	709.52	1203.13	1219.08
	#tinySP	35986	5864	1209	223	225
	#tinySP/#SP	64.12	37.48	16.15	5.33	5.45

Si se toma en consideración el número de temáticas, k, se puede afirmar que, tal y como era de esperar, cada MP no trata todas y cada una de las temáticas, es decir, los términos de sus intervenciones no se distribuyen en tantos subdocumentos como temáticas haya, sino en un subconjunto de temáticas donde el MP esté especializado. Del mismo modo, sí es cierto que

el número de subperfiles incrementa a medida que lo hace el valor de k, pero eso no ocurre como consecuencia directa de que el valor de k sea más alto, sino de que el algoritmo LDA es capaz de detectar temáticas más específicas. Aunque en estos casos, se obtienen unos subperfiles con pocos términos, que aún siendo más precisos, en muchos casos carecen de contexto y dejan de ser representativos como subperfil del MP.

Si se observa el número de subperfiles que se obtienen en función de la estrategia de distribución de temáticas que se utilice, se puede ver que con la estrategia Overlap se obtiene el mayor número de subperfiles, siguiendo en orden decreciente las estrategias Coseno, Sorensen y Dice, siendo la Euclídea la estrategia que genera el menor número de subperfiles. Esto tiene sentido con respecto a la manera en la que se distribuyen los términos en la temáticas a partir de estas estrategias. Si el número de temáticas es menor, es obvio que los subperfiles que se generen van a ser más grandes en número de términos y mas heterogéneos. Esto se puede apreciar claramente en el caso extremo de la estrategia Euclídea, donde todos los términos de un documento se asignan sólo a un subdocumento, el cual está asociado a la temática más probable. Por el contrario, con la estrategia Overlap se obtienen subperfiles más puros puesto que cada una de las ocurrencias de los distintos términos tiene que asignarse a cada una de las temáticas en las que se distribuye un documento.

Como cabe esperar, las temáticas más probables son propensas a recibir un mayor número de términos, pero por otro lado, las temáticas menos probables también reciben un número, generalmente pequeño, de términos. Para ilustrar esta situación, se van a considerar como subperfiles pequeños aquellos que tienen 50 términos (incluyendo repeticiones) o menos. El problema de este tipo de subperfiles pequeños es que en la gran mayoría de los casos no son útiles puesto que carecen de representatividad y contexto y puede llevar a una interpretación errónea por parte del sistema, puesto que este puede considerar como relevantes este tipo de subperfiles cuando en realidad solo es una temática que el MP ha tratado de forma marginal. Para tener una clara referencia a esto, en la Tabla 6.2 se han añadido el número absoluto de subperfiles pequeños que se generan con cada estrategia de distribución y valor de k (#tinySP) y los valores relativos con respecto al número total de subperfiles que se han generado (#tinySP/#SP). En este caso, se puede apreciar claramente que el número de subperfiles pequeños aparecen de manera más común cuando el valor de k o el número de temáticas seleccionadas por la estrategia de distribución va incrementando.

- **Análisis de los efectos en el ranking de salida del Sistema de Recuperación de Información**: Una vez analizado el comportamiento de los subperfiles con respecto a su tamaño, en función del valor de k y la estrategia de distribución, se va a estudiar cómo las características de los subperfiles influyen sobre el comportamiento del Sistema de Recuperación de Información y los correspondientes rankings de MP que se obtienen a partir de las consultas que se lanzan contra él. Cabe destacar que en el análisis que se va a

llevar a cabo en esta sección, no se va a tener en cuenta ni el proceso de fusión de instancias de un mismo diputado en el ranking ni ningún criterio de relevancia de un diputado con respecto a la consulta.

En primer lugar, se va a considerar un ejemplo de forma ilustrativa: si se analiza, el ranking que se obtiene con la consulta *"Concesiones de transporte público para personas mayores"*, cabe esperar que, en el ranking, los subperfiles relacionados con las temáticas "Transporte Público" y "Bienestar Social" de los MPs se sitúen en la parte alta. Por tanto, sería interesante estudiar la distribución de temáticas en el ranking de salida y cómo influye en el rendimiento del sistema que las temáticas estén más concentradas en un número menor de subperfiles o por el contrario más dispersas en un mayor número de subperfiles.

Cabe aclarar, que en estos experimentos, cuando se habla tanto de MPs como de temáticas, se está hablando de lo mismo, es decir, se obtiene un ranking de subperfiles de MPs y cada subperfil está asociado a una temática. Y para la evaluación de este tipo de problema es necesario encontrar una medida que sea capaz de capturar la diversidad o variabilidad de los resultados obtenidos. En teoría probabilística, la medida que se utiliza para este propósito es el cálculo de la entropía de una distribución. Para ello, el primer paso es el de usar las frecuencias con las que aparece una temática y las frecuencias de los MPs y calcular su distribución para poder estudiar así cómo están relacionadas. La situación que se espera es que obtenga unos valores de entropía bajos para la distribución de temáticas (la consulta solo debería devolver un conjunto pequeño y preciso de temáticas) y, al mismo tiempo, unos valores altos de entropía en la distribución de MPs (el sistema debe devolver el mayor número posible de MPs).

Los resultados del cálculo de la entropía se muestra en la Figura 6.4 cuando se usa el valor de $k = \sqrt{n/2} = 70$ y tomando para el cálculo de la entropía solo los 20 primeros subperfiles $< MP_i, T_j >$, no obstante, para otros valores de k los resultados obtenidos son muy similares. De forma más concreta, con el objetivo de poder relacionar el cálculo de las entropías con las estrategias de distribución, se ha usado la formula de la entropía normalizada de una distribución sobre n posibles resultados, con probabilidades $p_1, p_2, ..., p_n$, la cual se define como:

$$H_n(p) = -\sum_i \frac{p_i \log_b p_i}{\log_b n}. \tag{6.9}$$

Esta fórmula devuelve valores comprendidos en el intervalo [0,1] y, de esta forma los resultados de la entropía están normalizados independientemente del número de temáticas o el número de MPs.

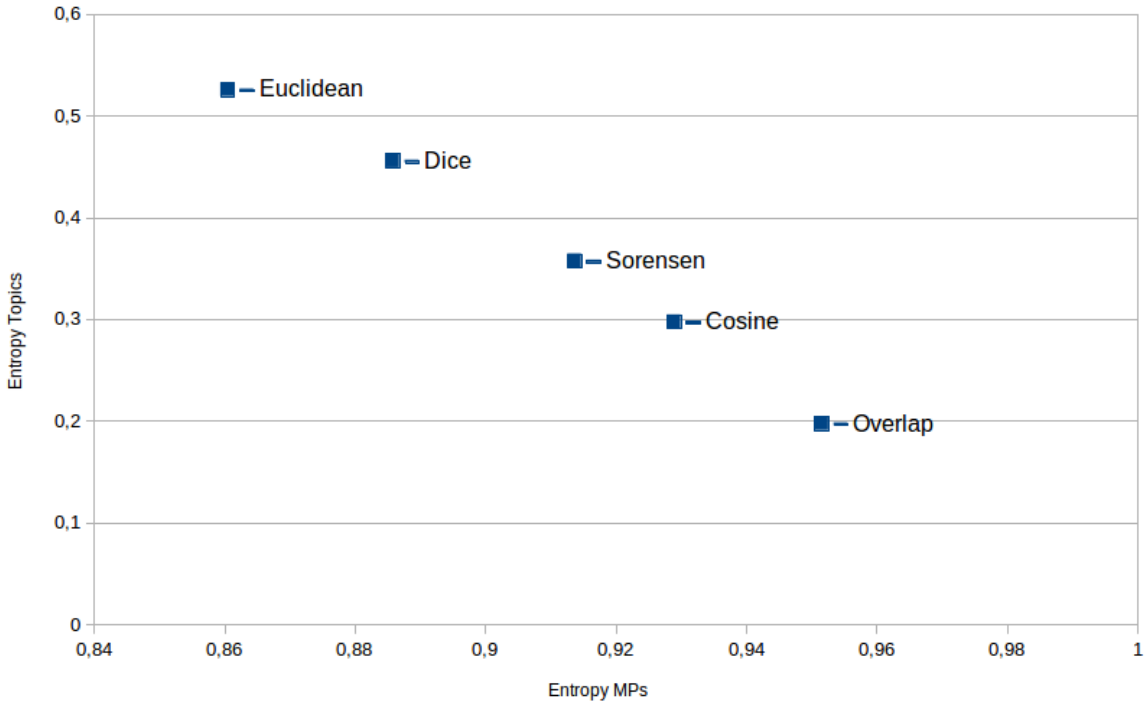

Figura 6.4: Entropía normalizada de la distribución de MPs y la distribución de temáticas considerando $k = \sqrt{n/2}$.

En la Figura 6.4, además se puede observar que los resultados obtenidos por la medida Overlap, con la cual se generan los subperfiles más puros, generalmente devuelve como salida un ranking con un número muy reducido de temáticas (probablemente las más representativos) y además obteniendo una alta diversidad en el conjunto de MPs. Aunque esta estrategia podría ser considerada como una buena opción si no la mejor, realmente no sería una buena elección puesto que, los subperfiles que quedan en la parte más alta del ranking no representan en absoluto los intereses reales de los MPs. Los subperfiles que genera esta estrategia tienen un número muy reducido de términos y muchos de ellos coinciden con los términos de la consulta, y ese es el motivo de que obtengan valores de score tan altos. En contrapartida, la estrategia Euclídea, con la que se generan subperfiles más heterogéneos y donde no es posible distinguir las temáticas dentro de un mismo documento, devuelve rankings con un número alto de temáticas. Esto ocurre porque los términos de una consulta pueden pertenecer a diferentes subperfiles al mismo tiempo independientemente de la temática que se trate en el subperfil. En consecuencia, el ranking de salida muestra una amplia variabilidad de temáticas y al mismo tiempo concentradas en un conjunto muy pequeño de MPs. El resto de estrategias, sin embargo, presentan un comportamiento más balanceado entre ambas situaciones, pero ocurre una tendencia de forma clara: cuando se elige una estrategia donde se aumentan el número de subperfiles, es decir que se obtienen

Tabla 6.3: Casos base: Aproximaciones basadas en términos, donde $m * n/t = 24$ y $\sqrt{n/2} = 70$ con extractos.

	NDCG@10	Precision@10	Recall@nr
TermInt	**0.6849**	**0.6478**	**0.4825**
TermMon	0.6317	0.6058	0.4391

Tabla 6.4: Resultados para las diferentes estrategias de distribuión (E)uclidea, (D)ice, (S)orensen, (C)oseno y (O)verlap, donde $m * n/t = 24$ y $\sqrt{n/2} = 70$ con extractos.

	NDCG@10			Precision@10			Recall@nr		
	$m * n/t$	$\sqrt{n/2}$	300	$m * n/t$	$\sqrt{n/2}$	300	$m * n/t$	$\sqrt{n/2}$	300
E	*0.6942*	0.7049	0.6935	*0.6688*	0.6764	0.6612	0.5058	0.5027	*0.4957*
D	0.6887	0.7088	0.6943	0.6656	***0.6808***	*0.6626*	***0.5063***	*0.5054*	0.4956
S	0.6817	***0.7096***	*0.6953*	0.6579	0.6792	0.6622	0.5038	0.5042	0.4949
C	0.6711	0.7022	0.6943	0.6472	0.6685	0.6543	0.4991	0.5000	0.4881
O	0.6264	0.6453	0.6100	0.6001	0.6072	0.5709	0.4682	0.4652	0.4477

perfiles más puros, el comportamiento del sistema tiende a reducir la variabilidad de las temáticas en los rankings que devuelve y aumenta la variabilidad de los MPs.

6.3.5 Resultados

El objetivo de esta sección es el de determinar si la aplicación del algoritmo LDA y las posteriores estrategias de distribución son una buena aproximación para construir subperfiles de MPs basado en términos y, en consecuencia, mejores recomendaciones de MPs.

En primer lugar se van a analizar los resultados obtenidos por los casos base. En la Tabla 6.3 y Tabla 6.5 se muestran los valores para las distintas métricas y los mejores resultados se muestran en negrita tanto usando los extractos como consultas como usando los extractos en conjunción con las materias. De esta tabla se puede concluir, que los perfiles basados en intervenciones, en ambos casos, obtienen de manera general mejores resultados que los perfiles monolíticos.

A continuación se van a discutir los resultados obtenidos considerando las diferentes estrategias de distribución (una vez que se ha aplicado el algoritmo LDA sobre la colección). La Tabla 6.4 y la Tabla 6.6 muestra los valores obtenidos para las diferentes métricas, con el mejor resultado global marcado en negrita y el mejor valor para cada k representado en cursiva tanto usando los extractos como consultas como usando los extractos en conjunción con las materias.

Con respecto a las tablas que determinan los resultados de los experimentos usando como consulta solo los extractos de las iniciativas, es decir, las Tablas 6.3 y 6.4 se puede apreciar que las distintas estrategias de distribución de los términos funcionan mejor que el mejor de los casos base tanto para las medidas de NDCG, *precision* y *recall* a excepción de la aproximación Overlap que funciona considerablemente peor incluso si es comparada con el peor de los casos base. Sin embargo, el resto de estrategias de distribución de términos entre sí devuelven resultados muy similares y, aunque se podría decir que para este caso las estrategias Dice y Sorensen son las

mejores, un test estadístico determina que no existen diferencias estadísticamente significativas entre las distintas aproximaciones independientemente de la medida que se elija o el valor de k. Si es cierto, por otro lado, que se observa una tendencia al alza cuando se toma como valor de $k = \sqrt{n/2}$, es decir, independientemente de la medida que se elija, este valor de k destaca por encima del resto aunque, en este caso, sea de manera muy poco apreciable. Dicho esto, tal y cómo se ha comentado previamente, resultaría interesante, dado que esta experimentación se realiza en el contexto de la extracción de temáticas, llevar a cabo la experimentación usando como consultas los extractos en conjunción con las respectivas materias de las iniciativas (Tablas 6.5 y 6.6).

Los resultados obtenidos usando como consultas los extractos y las materias siguen el mismo patrón que en el caso anterior pero en este caso las diferencias entre las distintas estrategias de distribución se aprecian de manera más clara. Una vez más, si comparamos los resultados con los casos base, donde no se aplica ninguna técnica de distribución de términos se tiene que, a excepción de la estrategia Overlap, el resto de estrategias supera tanto en la medida NDCG como en *precision* y *recall* a los casos base independientemente del valor de k. De hecho, en el mejor de los casos para la medida NDCG se obtiene una mejora de 5.31% y, en el caso de *precision* y *recall* se obtiene una mejora del 6.46% y 9.71% respectivamente, comparado con el mejor caso base y arrojando diferencias estadísticamente significativas (usando un t-test) para todos los casos.

Con respecto a cómo influye la elección del número de temáticas k que LDA debe extraer de la colección, se puede apreciar que con la elección de valores altos, como es el caso de $k = 300$, se obtienen peores resultados. Esto se produce puesto que con este valor los subperfiles que se obtienen para algunos MPs no son representativos de sus intereses reales ya que están formados sólo por un conjunto muy pequeño de términos‘sin contexto alguno, lo cual empeora la efectividad de la recuperación. Por otro lado, cuando se utiliza un valor pequeño para k, con es el caso de $m * n/t = 24$, muchas de las temáticas de la colección se mezclan entre sí quedando diluidas en un aglomerado de múltiples temáticas. Esto conlleva inevitablemente a una pérdida de la expresividad de los subperfiles y por tanto a peores valores. Dicho esto, se puede concluir que la selección de un valor de k adecuado es una tarea delicada y esencial para evitar o bien que haya demasiadas temáticas perdiéndose así el contexto de la información, o bien que haya muy pocas quedando estas aglomeradas. Por eso, a la vista de los resultados, una buena aproximación para determinar el número de temáticas por parte del algoritmo LDA es $k = \sqrt{n/2}$ donde se obtiene mejoras en las medidas de NDCG y *precision* con diferencias estadísticamente significativas mientras que en el *recall* gana $k = 300$ pero sin diferencias estadísticamente significativas.

Si se tuviera que determinar una estrategia que funcionase mejor que el resto, dado que a la vista de los resultados se puede apreciar claramente que con el uso de estrategias de distribución de términos se obtiene un mayor rendimiento y aunque las diferencias entre la distintas estrategias no son muy significativas al fin y al cabo, se podría escoger la estrategia Sorensen con $k = \sqrt{n/2}$. La elección de la estrategia de distribución de términos Sorensen viene

Tabla 6.5: Casos base: Aproximaciones basadas en términos, donde $m * n/t = 24$ y $\sqrt{n/2} = 70$ con extractos y materias.

	NDCG@10	Precision@10	Recall@nr
TermInt	**0.7098**	**0.6732**	**0.4973**
TermMon	0.6762	0.6514	0.4919

Tabla 6.6: Resultados para las diferentes estrategias de distribuión (E)uclidea, (D)ice, (S)orensen, (C)oseno y (O)verlap, donde $m * n/t = 24$ y $\sqrt{n/2} = 70$ con extractos y materias.

	NDCG@10			Precision@10			Recall@nr		
	$m * n/t$	$\sqrt{n/2}$	300	$m * n/t$	$\sqrt{n/2}$	300	$m * n/t$	$\sqrt{n/2}$	300
E	*0.7372*	0.7391	0.7269	*0.7131*	0.7135	0.6991	0.5349	0.5274	0.5156
D	0.7321	0.7418	0.7268	0.7095	0.7166	0.6998	***0.5363***	*0.5295*	*0.5162*
S	0.7273	***0.7482***	*0.7288*	0.7052	***0.7197***	*0.7008*	0.5321	0.5273	0.5152
C	0.7206	0.7423	0.7245	0.6976	0.7102	0.6943	0.5263	0.5210	0.5056
O	0.6753	0.6829	0.6512	0.6445	0.6448	0.6116	0.4895	0.4796	0.4613

determinada puesto que en ambos experimentos con distintos tipos de consultas, los resultados obtenidos han quedado a la cabeza en la mayoría de los casos, aunque también es cierto que la elección de estrategias es muy dependiente del número de temáticas que se establece a priori, ya que, cuando el número de temáticas es pequeño, la estrategia Euclídea parece funcionar mejor. Las diferencias entre las dos estrategias de distribución son estadísticamente significativas en ambos casos para las medidas de NDCG y *precision*: los p-values de un t-test son 0.003 y 0.004 para $k = m * n/t$ y $k = \sqrt{n/2}$ respectivamente. Sin embargo, desde el punto de vista de la medida de *recall* se puede comprobar con un test estadístico que no hay diferencias estadísticamente significativas entre las estrategias Euclídea, Dice y Sorensen. Dicho esto, también cabe destacar que con la estrategia Sorensen se obtienen perfiles más homogéneos permitiendo así una mejor representatividad de los intereses de los distintos MPs obteniendo además buenos resultados en la recuperación.

6.4 Conclusiones

En este capítulo se ha demostrado cómo la construcción de subperfiles basados en términos aplicando para ello el algoritmo LDA con el objetivo de descubrir y extraer temáticas subyacentes afecta, de manera positiva, al rendimiento de un sistema de recomendación de expertos en un ámbito parlamentario. De este modo, se puede aplicar este modelo estadístico generativo a un colección de documentos para obtener un conjunto de k temáticas. Las distribuciones de probabilidad que se generan en este proceso se utilizan para poder distribuir los diferentes términos que se encuentran en un documento original entre las distintas temáticas. Además, se han propuesto hasta cinco métodos distintos basados o bien en medidas de distancia, o bien en medidas de similitud para poder determinar el número correcto de temáticas en las que distribuir

los términos de un documento. Tras este proceso, finalmente, se construyen tantos subperfiles como temáticas se hayan seleccionado.

En los experimentos que se han llevado a cabo se puede observar de manera clara que todas las estrategias de redistribución de términos (a excepción de la estrategia Overlap) resultan ser útiles para ser utilizadas una vez que se ha aplicado LDA para construir los perfiles de los MP. De hecho, la aplicación de estas estrategias mejoran los resultados obtenidos con los casos base. Por otro lado, también se ha obtenido que la correcta selección previa del número de temáticas a extraer de la colección es relevante en el contexto de este problema, siendo $k = \sqrt{n/2}$ la mejor opción. Tal y como ya se ha mencionado, es importante la utilización de una estrategia de redistribución de los términos para poder generar unos subperfiles coherentes con respecto a la distribución de los términos entre las distintas temáticas, pero la selección de la mejor alternativa no es tan obvia, en el sentido en que depende del número de temáticas que se han considerado previamente para el algoritmo LDA (un múmero grande o pequeño). No obstante, si se tuviera que definir una de estas estrategias como la mejor, se seleccionaría la estrategia Sorensen puesto que con ella se obtienen perfiles más homogéneos.

6.5 Trabajos relacionados

La principal meta de la búsqueda de expertos (también referida cómo recomendación de expertos o identificación de expertos) es la de encontrar individuos o elementos expertos en ciertas áreas especificas, y además es un campo muy importante dentro de la Recuperación de Información. Una señal del creciente interés en este tipo de métodos es la existencia de diversas y recientes revisiones de la materia [2, 81, 104, 154].

Los métodos de búsqueda de expertos tienen numerosas aplicaciones: tal y cómo pueden ser la búsqueda de revisores apropiados para artículos enviados a una conferencia o revista [99], detectar personas que puedan responder a ciertas preguntas en el ámbito de los sistemas CQA (community question answering) [117], encontrar colaboradores para un proyecto [2], o de manera general, la identificación de expertos del mundo académico [80, 116], redes sociales [17, 146], compañías, instituciones y organizaciones [11, 71] o incluso en todo internet [54]. Otras aplicaciones potenciales para la búsqueda de expertos se relatan en [104]. La única aplicación de búsqueda de expertos enfocada en un ámbito político son las publicadas por los autores de los trabajos [40, 41].

Las dos aproximaciones básicas de búsqueda de expertos en términos de representación de los expertos candidatos son los métodos basados en perfiles y los métodos basados en documentos [10, 101]. El primer método también se denomina aproximación independiente de la consulta y el segundo también se define como aproximación dependiente de la consulta [111]. Los métodos basados en perfiles (o independientes de la consulta) basan su funcionamiento en la construcción de un perfil para cada candidato experto y para ello se concatenan todos los documentos relevantes

de este experto. Por tanto, dada una consulta, estos "macro documentos" (cada uno de ellos asociado a un experto) se pueden representar como un ranking usando técnicas clásicas de recuperación de documentos. Esta aproximación *monolítica* también se conoce cómo modelos de autor de documento único (*single-document author model*) [99]. Los métodos basados en documentos (o dependientes de la consulta) mantienen los documentos asociados a cada experto y se recuperan solo aquellos documentos que son relevantes a la consulta. Entonces, los candidatos expertos se representan cómo un ranking donde la posición de estos vienen definida mediante la combinación de la relevancia de los documentos asociados a cada candidato. Una instancia específica de este tipo de método usa el criterio del máximo para agregar los valores de relevancia de los documentos de cada candidato y se conoce como el modelo de autor de documento máximo (max-document author model) [99]. Otras formas diferentes de agregar los valores de relevancia de los distintos documentos asociados a un candidato se estudian en [93]. Dado que el hecho de que en esta investigación, la colección documentos que se ha utilizado son las intervenciones asociadas a cada MP, se ha utilizado el método basado en documentos pero renombrando la aproximación a *basado en intervenciones*. Mientras que los métodos basados en documentos se consideran de forma general como mejores que los métodos basados en perfiles [10], en ciertos estudios se llega a la conclusión opuesta [86]. En un contexto político, cómo el que acontece en esta investigación, la relaciones entre los candidatos expertos (MPs en este caso) no se muestran de forma tan natural como las que se presentan en los sistemas CQA donde hay un individuo que pregunta y otro que responde, o entre individuos que envían un email e individuos que lo reciben como en otro contextos. Por esta razón, en esta revisión no se han considerado métodos de Búsqueda de Expertos basados en análisis de enlaces [70, 88].

Existen varias aproximaciones basadas en el análisis probabilístico de semánticas latentes (pLSA) [115, 149, 150] en el contexto de los sistemas CQA. En [149], se aplica pLSA sobre una colección de preguntas, los usuarios se modelan de forma indirecta, y cada usuario se representa como el promedio de la distribución de las temáticas de las preguntas en las que se ha involucrado. Por el contrario, en [115], los usuarios se modelan de forma directa, es decir, todas las preguntas relativas a un usuario se agrupan en un único documento, y por tanto se aplica pLSA sobre esta colección de documentos. [150] distingue entre dos papeles que los usuarios pueden desempeñar, tanto como individuo que pregunta como individuo que responde y, estos papeles se modelan de manera conjunta utilizando una extensión de pLSA.

En los modelos basados en documentos, normalmente, las probabilidades $p(q|d)$ de una consulta dado un documento se estiman usando la máxima verosimilitud y un suavizado de Dirichlet. En [148], sin embargo, las probabilidades $p(q|d)$ se calculan de forma diferente usando las temáticas aprendidas por LDA a partir de la colección de documentos, de cierta manera como puentes, es decir $p(q|d) = \sum_x p(q|x)p(x|d)$, donde x representa las diferentes temáticas. El modelo basado en documentos se usa esencialmente en [84] en conjunción con una distribución *a priori* no uniforme para cada uno de los candidatos expertos la cual se construye a partir de una

modificación del algoritmo LDA. En [85] se combinan modelos de lenguaje y LDA para predecir a los mejores expertos para responder cuestiones en los sistemas CQA. En ambos casos las probabilidades de cada consulta dado un usuario se calculan a partir del perfil del usuario el cual engloba todas las respuestas (documentos) asociadas al usuario (perfil monolítico). En el caso de LDA, este cálculo se lleva a cabo de manera que $p(t|userprofile) = \sum_x p(t|x)p(x|userprofile)$, donde $p(t|x)$ representa la distribución de términos dado una temática, y $p(x|userprofile)$ es la distribución de las temáticas dado un documento (perfil) asociado a un usuario, y ambas distribuciones se obtienen con LDA. En [101], se aplica LDA sobre la colección de documentos, donde se incluyen, además, los nombres de los propios expertos. Por tanto el hecho de que las consultas y los nombres de los expertos aparezcan relacionados con un alto grado de probabilidad a las mismas temáticas indica qué el candidato es un experto para las temáticas de la consulta. LDA se usa para calcular la distribución de probabilidad de la consulta dadas unas temáticas y la distribución de probabilidad de los expertos dadas unas temáticas y, estas distribuciones se combinan en una fórmula que es capaz de calcular la probabilidad de una consulta dado un experto. En [163] se calcula el grado de relevancia entre categorías usando LDA sobre sistemas CQA (donde las categorías se representan como una distribución de temáticas y la relevancia se mide mediante la divergencia KL). Esto se combina con un análisis de enlaces para encontrar expertos en una categoría dada a partir de considerar la información de otras categorías relevantes. [162] utiliza LDA sobre una colección de documentos, y cada uno de estos documentos está asociado a cada usuario de un sistema CQA y en ellos se comprenden todas la preguntas del usuario, con el objetivo de detectar similitudes entre las temáticas de los usuarios que preguntan con los usuarios que contestan. Esta similitud entre temáticas se integra en un algoritmo de $page\ rank$ sensible a temáticas en una red que conecta usuarios que preguntan con usuarios que responden.

En [30], se consideran tanto el modelo de LDA tradicional (estático) como el modelo de LDA dinámico. En ambos casos, la distribución de temáticas de los documentos asociados a un mismo candidato experto se promedian. La distribución de temáticas de una consulta también se calcula con el objetivo de obtener una medida de similitud entre la distribución de temáticas asociada a un candidato y la distribución de la consulta. [37] considera además el modelado de temáticas a lo largo del tiempo para la búsqueda de expertos en literatura científica. En [99], se propone el modelo de temáticas APT (Author-Persona-Topic) para la recomendación de revisores para artículos científicos sometidos a revisión. En este modelo, cada autor puede adoptar el papel de diferentes "personas", y estas se representan como distribuciones de temáticas independientes. Entonces, cada "persona" representa combinaciones de temáticas distintas. Otro modelo de temáticas es el ATM (Author-Topic Model) [125] y este es una extensión del algoritmo LDA pero incluyendo información sobre la autoría considerando las relaciones entre los autores, documentos, temáticas y términos en sí. En [117], se usan tanto LDA como STM (Segmented Topic Model), se aplican sobre un sistema CQA. El STM permite a un perfil de usuario (el cual contiene el texto de las preguntas contestadas por el usuario) ser modelado como una estructura jerárquica, donde

cada pregunta en el perfil puede tener una distribución distinta de temáticas. En [56], se propone un nuevo modelo de temáticas: el modelo UQA (User-Question-Answer), el cual es una extensión de LDA para descubrir temáticas latentes contenidos en términos, categorías y usuarios en un sistema CQA. En este caso, los usuarios se modelan como un pseudo documento combinando para ello todas las preguntas y respuestas asociadas a los mismos. Además de las distribuciones comunes de temáticas por documento y términos por temáticas de LDA, en este caso también se considera una distribución de categorías por temática. En cuestiones de rendimiento, el mejor modelo considerado es la combinación lineal de los resultados obtenidos con la utilización del modelado de temáticas junto con modelos basado en términos (un modelo basado en documentos usando BM25).

[160] propone el modelo TEL (Topic-level Expert Learning), el cual combina el análisis de enlaces basado en grafos y el análisis de semántica basada en contenido para el modelado de expertos en un sistema CQA. También, en el contexto de CQA, [78] propone otra combinación de LDA (para medir la similitud entre candidatos expertos) y un análisis de enlaces. El perfil de cada candidato se obtiene así a partir de todas las respuestas del usuario en el sistema. [151] combina modelos de temáticas con los campos donde los usuarios son expertos para obtener, de esta forma, el modelo probabilístico TEM (Topic Expertise Model) en un sistema CQA. Dicho modelo también hace uso de información sobre las categorías de las preguntas y sobre la puntuación registrada por los usuarios, la cual que mide la relevancia tanto para las preguntas como para las respuestas. Entonces, se plantea el modelo CQArank, el cual combina los intereses del usuario en ciertas temáticas y el modelo TEM para aprender las áreas donde el usuario es experto con una estructura de enlaces. [78], por otro lado, propone un modelo híbrido para la búsqueda de expertos en sistemas CQA el cual aprovecha la mayoría de la información disponible en estos sistemas, como puede ser el etiquetado de preguntas, las votaciones de los usuarios y las mejores respuestas. Esta información se considera para una combinación entre un análisis de enlaces y un análisis de temáticas y, de este modo, se consideran tanto la autoría como la similitud entre los usuarios y las preguntas. En términos del análisis de temáticas, el contenido textual de las preguntas asociadas a cada usuario se suplementa con las diez palabras más similares (extraídas de Wikipedia usando word2vect) para cada etiqueta asociada a la pregunta en sí. Finalmente, una extensión de LDA, denominada tag-LDA, se usa para encontrar las temáticas asociadas a cada usuario. En el contexto de CQA, [132] extiende el funcionamiento del LDA supervisado considerando el paradigma $learning-to-rank$, planteando la recomendación de preguntas como un problema de emparejamiento y teniendo en cuenta los votos de los diferentes expertos que han respondido una pregunta por parte de otros usuarios.

[136] considera, por otro lado, los documentos recuperados en la parte alta del ranking dada una consulta. Cada uno de estos documentos se representa como una mezcla de modelos de lenguaje de los diferentes expertos (como complemento al modelo de lenguaje global) puesto que asume que cada documento se puede asociar a un conjunto de varios expertos. Entonces,

el algoritmo EM se utiliza para estimar las probabilidades de los términos de la consulta dado un experto. En [40, 41], se aplican métodos de búsqueda de expertos (sin usar modelado de temáticas) en un contexto político. Se utilizan aproximaciones basadas en perfiles (monolítico) y aproximaciones basadas en documentos [41] considerando también subperfiles basados en la participación de los MPs en diversas comisiones parlamentarias [40].

CAPÍTULO 7

CONCLUSIONES GENERALES

7.1 Observaciones finales

En esta Tesis Doctoral se han abordado diferentes aproximaciones para mejorar el rendimiento de un Sistema de Recomendación y Filtrado en un ámbito parlamentario. Para ello se ha considerado la aplicación de técnicas de Aprendizaje Automático tanto para obtener representaciones alternativas a los perfiles de los MPs como para sustituir los métodos de indexación clásicos en Recuperación de Información por modelos de clasificación. El objetivo principal de este sistema era el de filtrar documentos con información para obtener aquellos MPs a los que les podría resultar relevante y, por otro lado, recomendar aquellos MPs que cumplan con ciertos requisitos de información establecidos por los usuarios. Para este propósito se han llevado a cabo una serie de experimentos usando como colección las transcripciones literales de los MPs en la octava legislatura del Parlamento de Andalucía. Como resultados de estos experimentos se han obtenido una serie de rankings de MPs (o de subperfiles de MPs) con el propósito de ser evaluados y comparados para obtener qué aproximación se comporta mejor de cara al rendimiento del sistema.

En el Capítulo 3 se lleva a cabo una primera aproximación al problema, construyendo para ello un Sistema de Recomendación y Filtrado desde dos enfoques distintos. El primer enfoque consiste en construir el sistema mediante un conjunto de clasificadores binarios, uno por cada MP, donde las instancias positivas son las intervenciones del propio MP y las instancias negativas son el resto de intervenciones de la colección, y de esta forma las consultas se clasifican siendo más relevantes para aquellos MP en los que se obtenga un mayor porcentaje de pertenencia a la clase positiva. El segundo enfoque se corresponde con el uso de técnicas clásicas de Recuperación de Información para indexar los perfiles de los MPs. Estos perfiles, a su vez, se construyen de dos formas distintas y opuestas entre sí; por un lado, uniendo todas las intervenciones de un mismo MP en un único documento siendo este su perfil y, por otro lado, asignando cada intervención a un documento distinto de modo que el MP tendrá tantos subperfiles como intervenciones.

Observando los resultados obtenidos a partir del proceso de experimentación donde se han utilizado las medidas clásicas para evaluar un Sistema de Recuperación y Filtrado, *precision*, *recall* y *F-measure*, se han obtenido ciertas conclusiones sobre el rendimiento del sistema que cabría destacar. Aunque ambos enfoques se comportan de una manera muy distinta con respecto a las medidas de *precision* y *recall* y están muy condicionadas a los valores que se elijan para el umbral de relevancia, si nos centramos en la *F-measure* tanto en su versión macro-averaged como para micro-averaged y en la medida de evaluación de rankings NDCG@10 no se puede establecer que exista ningún enfoque que sea mejor que el otro, de hecho, de cara al rendimiento del sistema, los resultados son muy similares. Ciertamente, la forma en la que se ha evaluado la aproximación basada en Aprendizaje Automático no es la más justa, puesto que considerar que todas las instancias ajenas a un MP en cuestión pertenecen al conjunto de entrenamiento negativo no es muy realista. Esto lleva al desarrollo de la investigación por los caminos que se relatan en los dos siguientes capítulos a este donde, por un lado se plantea utilizar técnicas de Positive Unlabeled Learning para conseguir mejores modelos de clasificación y por otro lado, encontrar formas alternativas de representación de los perfiles de los MPs para las aproximaciones basadas en Recuperación de Información.

En el Capítulo 4, con el objetivo de obtener modelos de clasificación mejores y más realistas de los MPs y poder ser competitivos contra los enfoques basales de Recuperación de Información se ha abordado el problema tomando en consideración la manera de construir el conjunto de entrenamiento de un MP de cara a construir un clasificador. Cómo primer acercamiento al problema, en el Capítulo 3 se construyen los clasificadores usando las propias intervenciones de un diputado como entrenamiento positivo y el resto de intervenciones como entrenamiento negativo, pues bien, eso no es del todo correcto si se aplica de forma práctica. Puede darse que ciertas intervenciones de MPs traten sobre las mismas temáticas que le son relevantes a un MP en cuestión y, aún así, en este punto se están considerando como entrenamiento negativo. Esto genera en el clasificador mucho ruido, en el sentido que puede estar considerando a la vez como negativo y positivo los mismos elementos del discurso de un MP. Para evitar esta casuística se propone, no solo la aplicación de técnicas de Positive Unlabeled Learning, sino el diseño de una nueva técnica de este tipo que sea capaz de adaptarse mejor al caso de estudio.

Efectivamente, en todos los experimentos que se han llevado a cabo, se demuestra que la hipótesis de que considerar el conjunto de intervenciones ajenas a un MP como entrenamiento negativo, contribuía negativamente a la construcción de los modelos de clasificación. En la aproximación de PUL basada en K-means se obtienen los mejores resultados de la experimentación, siendo estos mejores incluso que los obtenidos con las aproximaciones base de Recuperación de Información. Por tanto, se puede confirmar que la aplicación de técnicas de Positive Unlabeled Learning en este tipo de problemas ayuda a obtener un mejor rendimiento en un Sistema de Recomendación y Filtrado construido mediante la utilización de conceptos del área del Aprendizaje Automático. Por otro lado, cabe destacar que, tal y como se ha dicho anteriormente, las

medidas de *precision* y *recall* son muy dependientes del umbral de relevancia que se escoja para determinar el corte que define qué es relevante y qué no en el ranking. Por tanto, para lidiar con este problema, se han evaluado distintas formas de establecer el valor de umbral óptimo para cada diputado, siendo el criterio de utilizar el conjunto de entrenamiento completo del MP para validar el modelo y determinar así el umbral la mejor forma de abordar este problema, incluso bajo el riesgo de sobreajuste, se consiguen mejorar los resultados que se obtienen en el caso base donde se establece el valor de umbral a 0.5.

En el Capítulo 5, a raíz de las conclusiones extraídas en el Capítulo 3, se pretende encontrar una representación alternativa de los perfiles de los MPs mediante la aplicación de técnicas de agrupamiento. De esta forma se obtienen una serie de subperfiles por cada MP, los cuales se conforman a partir de los documentos, intervenciones en este caso, que se han asignado de forma automática al mismo grupo. El objetivo de este procedimiento es el de intentar potenciar las ventajas de las aproximaciones base y minimizar a su vez las problemáticas derivadas de unas representaciones de los perfiles tan básicas. Para ello se han llevado a cabo diversos experimentos utilizando técnicas clásicas de detección automática de grupos de manera no supervisada como son los algoritmos basados en centroides K-means y K-medoids (PAM) o los algoritmos de clustering jerárquico de tipo divisivo (DIANA) y aglomerativo (AGNES). Además se ha llevado a cabo un agrupamiento de los documentos de una forma menos ortodoxa utilizando para ello el algoritmo LDA y agrupando los documentos en torno a la temática más probable de estos y, por otro lado, aplicando un mapa auto-organizativo (SOM) para reducir la dimensionalidad del espacio donde están definidos los documentos generando así una retícula de neuronas bidimensional, donde cada neurona engloba a una serie de documentos con características similares, realizando el agrupamiento, por tanto, sobre las neuronas. Por otro lado, se han establecido, con el objetivo de ser comparados, diversos criterios de establecer el número de clusters en los que se pueden agrupar las intervenciones de un MP, basándose para ello en las propias características del mismo. Dicho esto, el objetivo de este capítulo es el estudiar si una representación de los perfiles basada en clustering de documentos y, a su vez, intermedia con respecto a las dos aproximaciones basales ayuda a mejorar el rendimiento de un Sistema de Recomendación y Filtrado.

Tras el proceso de experimentación que se ha llevado a cabo, se puede afirmar que la aplicación de técnicas de agrupamiento no supervisado para construir subperfiles de MP es, de forma general, una buena opción que incluso supera el rendimiento de aproximaciones donde se conoce el número de grupos de manera externa como es el caso de las comisiones parlamentarias. La gran mayoría de experimentos que se han llevado a cabo superan, en lo referente al rendimiento general del sistema, a los resultados obtenidos con los casos base con diferencias estadísticamente significativas. Por otro lado, utilizando el sistema en su versión de filtrado de documentos, cabe destacar que el sistema se comporta claramente mejor enfocando la construcción de perfiles de manera global, es decir, realizando el agrupamiento sobre la colección completa de intervenciones y luego agrupar juntas las intervenciones de un mismo diputado que han sido asignadas a cada

grupo. Sin embargo, utilizando el sistema en su versión de recomendación de expertos, no se puede definir claramente qué tipo de clustering arroja mejores resultados aún cuando las primeras cuatro alternativas en el ranking de rendimiento hacen referencia al enfoque de construcción de perfiles local, es decir, llevar a cabo el clustering únicamente sobre las intervenciones de un MP y agrupar juntas las que estén asignadas a un mismo cluster. Este comportamiento está relacionado de forma directa con el tamaño de los clusters que cada aproximación genera y el hecho de que en un problema de filtrado el tamaño de la consulta, en este caso el documento completo, es generalmente mayor. Poniendo el foco ahora sobre los distintos criterios que se han considerado para establecer el número adecuado de grupos, para el filtrado de documentos el mejor criterio es el de establecer un número alto de grupos, en este caso $\sqrt{n/2}$, y para la recomendación, sin embargo, mn/t funciona mejor. Finalmente, con respecto a la elección del algoritmo de clustering en sí, aún pudiéndose apreciar que la decisión en este caso no es crítica, puesto que no se han encontrado diferencias estadísticamente significativas entre los algoritmo a la cabeza del ranking de rendimiento, cabe destacar que los métodos jerárquicos, en particular el aglomerativo, funcionan bastante mejor para problemas de filtrado, mientras que LDA, SOM y los métodos basados en centroides obtienen un mejor rendimiento para el problema de recomendación.

En el Capítulo 6 ha quedado patente que la construcción de subperfiles basados en términos en el contexto de un sistema de recomendación de expertos en un ámbito parlamentario, utilizando la técnica LDA para extraer temáticas latentes y distribuir así los términos en grupos más homogéneos, es un método que, no solo ayuda a mejorar el rendimiento del sistema, sino que contribuye a generar subperfiles más descriptivos en lo que se refiere a un subperfil en cuestión y más heterogéneos en lo que se refiere a la comparación entre subperfiles de un mismo MP. Por otro lado, las diversas estrategias de distribución de los términos de cada temática en un documento dado, ya sean usando medidas de distancia o medidas de similitud, ayuda a que los subperfiles que se obtienen no sean demasiado difusos y evitando que un documento se fragmente hasta el punto de que la información en él pueda perder el contexto. Para la distribución de términos se han estudiado diversas estrategias pero, en definitiva, todas ellas se pueden englobar en cinco qué, ordenadas por cómo de restrictivas son a la hora de distribuir los términos son Euclídea, Dice, Sorensen, Coseno y Overlap. De esta forma, el objetivo es construir tantos subperfiles como temáticas se hayan seleccionado.

Los experimentos realizados demuestran que la aplicación de las diversas estrategias de distribución de los términos, a excepción de la estrategia Overlap, resultan de gran utilidad para construir perfiles basado en términos una vez se ha aplicado el algoritmo LDA. En efecto, la decisión de aplicar cualquiera de estas estrategias supera a los resultados obtenidos en los casos base. También es cierto que con anterioridad es necesario escoger un número adecuado de temáticas a extraer de la colección documental y para ello se han probado con tres valores distintos resultando ser $k = \sqrt{n/2}$ la mejor opción. Dicho esto, a la vista de los resultados obtenidos en los experimentos que se han llevado a cabo en el Capítulo 6 si se tuviera que seleccionar un

método que fuese mejor que el resto, se optaría por la aproximación donde el número de temáticas a extraer de la colección sería, tal y como se ha comentado $k = \sqrt{n/2}$ y la estrategia de distribución de los términos de los documentos para la construcción de los subperfiles de cada MP sería la estrategia Sorensen, puesto que con ella los perfiles de los MP que se obtienen son notablemente más homogeneos.

7.2 Trabajo futuro

A la vista de los resultados generales obtenidos en esta línea de investigación se puede confirmar que la aplicación de técnicas de Aprendizaje Automático, tanto para mejorar la construcción y representación de perfiles de usuario, como para mejorar el diseño de un Sistema de Recuperación de Información, son una correcta aproximación para este tipo de problemas. A nivel más específico, en relación con el uso de clasificadores y técnicas de PUL (Capítulo 4), se propone explorar más detenidamente el empleo de técnicas de balanceo de clases en los conjuntos de datos, así como el estudio de formas alternativas para establecer un umbral específico para los diferentes MPs. También plantear el uso de técnicas de selección de características (términos en este caso de estudio) para mejorar el rendimiento de los clasificadores. En relación con el empleo de métodos de clustering para generar subperfiles más homogéneos (Capítulo 5), explorar el uso de otros métodos distintos a los ya considerados, como DBSCAN, OPTICS, Affinity Propagation, etc. Con respecto al uso del algoritmo LDA para generar subperfiles basados en términos (Capítulo 6), se plantea la distribución de los términos de los documentos entre subperfiles a nivel de párrafos completos en lugar de términos de manera individual, para intentar así capturar mejor las temáticas esenciales tratadas en cada documento. Por otro lado, de forma más global, se considera incluir la dimensión temporal en los perfiles o subperfiles, por ejemplo, considerando, entre otros, subperfiles a corto y largo plazo, usando versiones temporales de los métodos básicos ya considerados (clustering y LDA) y, por último, se propone también exportar el conocimiento extraído a partir de experimentar en el ámbito parlamentario a otros ámbitos más extensos y complejos, con colecciones más grandes y conceptos de usuario distintos, con el objetivo de estudiar la escalabilidad de los métodos que se describen en esta Tesis Doctoral y determinar como deben variarse para ser aplicados en contextos más ambiciosos.

7.3 Lista de publicaciones

L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, L. Redondo-Expósito. **Comparing machine learning and information retrieval-based approaches for filtering documents in a parliamentary setting**. Lecture Notes in Artificial Intelligence 10564, Scalable Uncertainty Management, Moral S., Pivert O., Sánchez D., Marín N. (Eds.), 64-77, 2017. 11th International Conference on Scalable Uncertainty Management (SUM 2017). (ISBN 978-3-319-67582-4)

Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Luis Redondo-Expósito. **Positive unlabeled learning for building recommender systems in a parliamentary setting**. Information Sciences, Volumes 433-434, 221-232, 2018. (ISSN 0020-0255)

L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, L. Redondo-Exposito. **Selecting relevance thresholds to improve a recommender system in a parliamentary setting**. Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, 186-193, 2018. (ISBN 978-989-758-330-8)

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**. Knowledge-Based Systems, Volume 190, 105337, 2020. (ISSN 0950-7051)

Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, Luis Redondo-Expósito. **LDA-based terms profiles for expert finding in a political setting**. Transactions on Intelligent Systems and Technology, 2020 (enviado).

BIBLIOGRAFÍA

[1] J.-W. AHN, P. BRUSILOVSKY, J. GRADY, D. HE, AND S. Y. SYN, *Open user profiles for adaptive news systems: Help or harm?*, in Proceedings of the 16th International Conference on World Wide Web, WWW 07, New York, NY, USA, 2007, ACM, pp. 11–20.

[2] M. AL-TAIE, S. KADRY, AND A. OBASA, *Understanding expert finding systems: domains and techniques*, Social Network Analysis and Mining, 8 (2018).

[3] E. ALPAYDIN, *Introduction to Machine Learning*, The MIT Press, 2nd ed., 2010.

[4] B. AMINI, R. IBRAHIM, M. S. OTHMAN, AND A. SELAMAT, *Capturing scholar's knowledge from heterogeneous resources for profiling in recommender systems*, Expert Syst. Appl., 41 (2014), pp. 7945–7957.

[5] J. D. ANDERSON AND J. PÉREZ-CARBALLO, *The nature of indexing: how humans and machines analyze messages and texts for retrieval. part i: Research, and the nature of human indexing*, Information Processing & Management, 37 (2001), pp. 231 – 254.

[6] C. M. AU YEUNG, N. GIBBINS, AND N. SHADBOLT, *Multiple interests of users in collaborative tagging systems*, in Weaving Services and People on the World Wide Web, R. Baeza-Yates and I. King, eds., Springer, 2009, pp. 255–274.

[7] R. A. BAEZA-YATES AND B. RIBEIRO-NETO, *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[8] R. C. BALABANTARAY, C. SARMA, AND M. JHA, *Document clustering using k-means and k-medoids*, CoRR, abs/1502.07938 (2015).

[9] K. BALOG, L. AZZOPARDI, AND M. DE RIJKE, *Formal models for expert finding in enterprise corpora*, in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 06, New York, NY, USA, 2006, ACM, pp. 43–50.

[10] K. BALOG, Y. FANG, M. DE RIJKE, P. SERDYUKOV, AND L. SI, *Expertise retrieval*, Found. Trends Inf. Retr., 6 (2012), pp. 127–256.

[11] S. BAYATI, *Security expert recommender in software engineering*, in Proceedings of the 38th International Conference on Software Engineering Companion, ICSE 16, New York, NY, USA, 2016, ACM, pp. 719–721.

[12] N. J. BELKIN AND W. B. CROFT, *Information filtering and information retrieval: Two sides of the same coin?*, Commun. ACM, 35 (1992), pp. 29–38.

[13] D. BILLSUS, M. J. PAZZANI, AND J. CHEN, *A learning agent for wireless news access*, in Proceedings of the 5th International Conference on Intelligent User Interfaces, IUI 00, New York, NY, USA, 2000, ACM, pp. 33–36.

[14] D. M. BLEI, A. Y. NG, AND M. I. JORDAN, *Latent dirichlet allocation*, J. Mach. Learn. Res., 3 (2003), pp. 993–1022.

[15] J. BOBADILLA, F. ORTEGA, A. HERNANDO, AND A. GUTIÉRREZ, *Recommender systems survey*, Know.-Based Syst., 46 (2013), pp. 109–132.

[16] C. BOURAS AND V. TSOGKAS, *Improving news articles recommendations via user clustering*, International Journal of Machine Learning and Cybernetics, 8 (2017), pp. 223–237.

[17] A. BOZZON, M. BRAMBILLA, S. CERI, M. SILVESTRI, AND G. VESCI, *Choosing the right crowd: Expert finding in social networks*, in Proceedings of the 16th International Conference on Extending Database Technology, EDBT 13, New York, NY, USA, 2013, ACM, pp. 637–648.

[18] Q. V. BUI, K. SAYADI, S. B. AMOR, AND M. BUI, *Combining latent dirichlet allocation and k-means for documents clustering: Effect of probabilistic based distance measures*, in Intelligent Information and Database Systems, N. T. Nguyen, S. Tojo, L. M. Nguyen, and B. Trawiński, eds., Cham, 2017, Springer International Publishing, pp. 248–257.

[19] R. BURKE, *Knowledge-based recommender systems*, Encyclopedia of library and information systems, 69 (2000).

[20] R. BURKE, *Hybrid recommender systems: Survey and experiments*, User Modeling and User-Adapted Interaction, 12 (2002).

[21] V. BUSH, *As we may think*, Interactions, 3 (1996), pp. 35–46.

[22] B. CALVO, P. LARRAÑAGA, AND J. A. LOZANO, *Learning bayesian classifiers from positive and unlabeled examples*, Pattern Recogn. Lett., 28 (2007), pp. 2375–2384.

[23] F. CAN AND E. A. OZKARAHAN, *Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases*, ACM Trans. Database Syst., 15 (1990), pp. 483–517.

[24] M. E. CELEBI AND K. AYDIN, *Unsupervised Learning Algorithms*, Springer Publishing Company, Incorporated, 1st ed., 2016.

[25] S.-H. CHA, *Comprehensive survey on distance/similarity measures between probability density functions*, International Journal Mathematical Models and Methods in Applied Sciences, 1 (2007), pp. 300–307.

[26] O. CHAPELLE, B. SCHLKOPF, AND A. ZIEN, *Semi-Supervised Learning*, The MIT Press, 1st ed., 2010.

[27] N. V. CHAWLA, K. W. BOWYER, L. O. HALL, AND W. P. KEGELMEYER, *Smote: Synthetic minority over-sampling technique*, J. Artif. Int. Res., 16 (2002), pp. 321–357.

[28] C. CHEN, M. CHEN, AND Y. SUN, *A web document personalization user model and system*, in Proceedings of the Information Retrieval and User Modelling Conference, 2001.

[29] L. CHEN AND K. SYCARA, *Webmate: A personal agent for browsing and searching*, in Proceedings of the Second International Conference on Autonomous Agents, AGENTS 98, New York, NY, USA, 1998, ACM, pp. 132–139.

[30] R. CHI, B. WU, AND L. WANG, *Expert identification based on dynamic lda topic model*, 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), (2018), pp. 881–888.

[31] C. CLEVERDON, *Optimizing convenient online access to bibliographic databases*, Inf. Serv. Use, 4 (1984), pp. 37,Äì–47.

[32] W. W. COHEN, *Learning rules that classify e-mail*, in Proceedings of the 1996 AAAI Spring Symposium on Machine Learning and Information Access, 1996, pp. 18–25.

[33] G. V. CORMACK, C. L. A. CLARKE, AND S. BUETTCHER, *Reciprocal rank fusion outperforms condorcet and individual rank learning methods*, in Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 09, New York, NY, USA, 2009, ACM, pp. 758–759.

[34] M. E. CORTÉS-CEDIEL, I. CANTADOR, AND O. GIL, *Recommender systems for e-governance in smart cities: State of the art and research opportunities*, in Proceedings of the International Workshop on Recommender Systems for Citizens, CitRec 17, New York, NY, USA, 2017, ACM, pp. 7:1–7:6.

[35] N. CRISTIANINI AND J. SHAWE-TAYLOR, *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*, Cambridge University Press, New York, NY, USA, 2000.

[36] B. CROFT, D. METZLER, AND T. STROHMAN, *Search Engines: Information Retrieval in Practice*, Addison-Wesley Publishing Company, USA, 1st ed., 2009.

[37] A. DAUD, J. LI, L. ZHOU, AND F. MUHAMMAD, *Temporal expert finding through generalized time topic modeling*, Know.-Based Syst., 23 (2010), pp. 615–625.

[38] D. L. DAVIES AND D. W. BOULDIN, *A cluster separation measure*, IEEE Trans. Pattern Anal. Mach. Intell., 1 (1979), pp. 224–227.

[39] L. M. DE CAMPOS, J. M. FERNÁNDEZ-LUNA, AND J. F. HUETE, *A lazy approach for filtering parliamentary documents*, in Electronic Government and the Information Systems Perspective, A. Kő and E. Francesconi, eds., Cham, 2015, Springer International Publishing, pp. 364–378.

[40] L. M. DE CAMPOS, J. M. FERNÁNDEZ-LUNA, AND J. F. HUETE, *Committee-based profiles for politician finding*, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 25 (2017), pp. 21–36.

[41] L. M. DE CAMPOS, J. M. FERNÁNDEZ-LUNA, AND J. F. HUETE, *Profile-based recommendation: A case study in a parliamentary context*, Journal of Information Science, 43 (2017), pp. 665–682.

[42] L. M. DE CAMPOS, J. M. FERNÁNDEZ-LUNA, J. F. HUETE, C. J. MARTÍN-DANCAUSA, A. TAGUA-JIMÉNEZ, AND C. TUR-VIGIL, *An integrated system for managing the andalusian parliament's digital library*, Program, 43 (2009), pp. 156–174.

[43] F. DENIS, R. GILLERON, AND F. LETOUZEY, *Learning from positive and unlabeled examples*, Theoretical Computer Science, 348 (2005), pp. 70 – 83.

[44] P. W. FOLTZ AND S. T. DUMAIS, *Personalized information delivery: An analysis of information filtering methods*, Commun. ACM, 35 (1992), pp. 51–60.

[45] N. FUHR AND C. BUCKLEY, *A probabilistic learning approach for document indexing*, ACM Trans. Inf. Syst., 9 (1991), pp. 223–248.

[46] N. FUHR AND U. PFEIFER, *Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumptions*, ACM Trans. Inf. Syst., 12 (1994), pp. 92–115.

[47] G. P. C. FUNG, J. X. YU, H. LU, AND P. S. YU, *Text classification without negative examples revisit*, IEEE Trans. on Knowl. and Data Eng., 18 (2006), pp. 6–20.

[48] H. GAN, Y. ZHANG, AND Q. SONG, *Bayesian belief network for positive unlabeled learning with uncertainty*, Pattern Recogn. Lett., 90 (2017), pp. 28–35.

[49] M. GAO, K. LIU, AND Z. WU, *Personalisation in web computing and informatics: Theories, techniques, applications, and future research*, Information Systems Frontiers, 12 (2010), pp. 607–629.

[50] S. GAUCH, M. SPERETTA, A. CHANDRAMOULI, AND A. MICARELLI, *User profiles for personalized information access*, in The Adaptive Web, Methods and Strategies of Web Personalization, P. Brusilovsky, A. Kobsa, and W. Nejdl, eds., vol. 4321 of Lecture Notes in Computer Science, Springer, 2007, pp. 54–89.

[51] M. R. GHORAB, D. ZHOU, A. O'CONNOR, AND V. WADE, *Personalised information retrieval: Survey and classification*, User Modeling and User-Adapted Interaction, 23 (2013), pp. 381–443.

[52] C. A. GOMEZ-URIBE AND N. HUNT, *The netflix recommender system: Algorithms, business value, and innovation*, ACM Trans. Manage. Inf. Syst., 6 (2015), pp. 13:1–13:19.

[53] T. R. L. GRIFFITHS AND M. STEYVERS, *Finding scientific topics.*, Proceedings of the National Academy of Sciences of the United States of America, 101 Suppl 1 (2004), pp. 5228–35.

[54] Z. GUAN, G. MIAO, R. MCLOUGHLIN, X. YAN, AND D. CAI, *Co-occurrence-based diffusion for expert search on the web*, IEEE Trans. on Knowl. and Data Eng., 25 (2013), pp. 1001–1014.

[55] J. A. GULLA, A. D. FIDJESTOL, X. SU, AND H. CASTEJON, *Implicit user profiling in news recommender systems*, in Proceedings of the 10th International Conference on Web Information Systems and Technologies - Volume 1: WEBIST,, INSTICC, ScitePress, 2014, pp. 185–192.

[56] J. GUO, S. XU, S. BAO, AND Y. YU, *Tapping on the potential of q&a community by recommending answer providers*, in Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 08, New York, NY, USA, 2008, ACM, pp. 921–930.

[57] U. HANANI, B. SHAPIRA, AND P. SHOVAL, *Information filtering: Overview of issues, research and systems*, User Modeling and User-Adapted Interaction, 11 (2001), pp. 203–259.

[58] J. HANNON, K. MCCARTHY, M. P. O'MAHONY, AND B. SMYTH, *A multi-faceted user model for twitter*, in User Modeling, Adaptation, and Personalization, J. Masthoff, B. Mobasher, M. C. Desmarais, and R. Nkambou, eds., Berlin, Heidelberg, 2012, Springer Berlin Heidelberg, pp. 303–309.

[59] J. HERNÁNDEZ-GONZÁLEZ, I. N. INZA, AND J. A. LOZANO, *Learning from proportions of positive and unlabeled examples*, International Journal of Intelligent Systems, 32 (2017), pp. 109–133.

[60] G. HUGHES AND R. CROWDER, *Experiences in designing highly adaptable expertise finder systems*, Design Engineering Technical Conferences and Computers and Information in Emgineering, (2008), pp. 451–460.

[61] A. JAISWAL AND P. N. JANWE, *Article: Hierarchical document clustering: A review*, IJCA Proceedings on 2nd National Conference on Information and Communication Technology, NCICT (2011), pp. 37–41.

[62] N. JARDINE AND C. VAN RIJSBERGEN, *The use of hierarchic clustering in information retrieval*, Information Storage and Retrieval, 7 (1971), pp. 217 – 240.

[63] K. JÄRVELIN AND J. KEKÄLÄINEN, *Cumulated gain-based evaluation of ir techniques*, ACM Trans. Inf. Syst., 20 (2002), pp. 422–446.

[64] S. JAYAPRADA, A. ASWANI, AND G. GAYATHRI, *Hierarchical divisive clustering with multi view-point based similarity measure*, in Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013, S. C. Satapathy, S. K. Udgata, and B. N. Biswal, eds., Cham, 2014, Springer International Publishing, pp. 483–491.

[65] A. JENNINGS AND H. HIGUCHI, *A user model neural network for a personal news service*, User Modeling and User-Adapted Interaction, 3 (1993), pp. 1–25.

[66] K. JONES, S. WALKER, AND S. ROBERTSON, *A probabilistic model of information retrieval: development and comparative experiments: Part 2*, Information Processing & Management, 36 (2000), pp. 809–840.

[67] K. S. JONES, S. WALKER, AND S. E. ROBERTSON, *A probabilistic model of information retrieval: Development and comparative experiments*, Inf. Process. Manage., 36 (2000), pp. 779–808.

[68] P. JUNTUNEN, M. LIUKKONEN, M. J. LEHTOLA, AND Y. HILTUNEN, *Cluster analysis by self-organizing maps: An application to the modelling of water quality in a treatment process*, Applied Soft Computing, 13 (2013), pp. 3191–3196.

[69] D. JURAFSKY AND J. H. MARTIN, *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*, Pearson Prentice Hall, Upper Saddle River, N.J., 2009.

[70] P. JURCZYK AND E. AGICHTEIN, *Discovering authorities in question answer communities by using link analysis*, in Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 07, New York, NY, USA, 2007, ACM, pp. 919–922.

[71] M. KARIMZADEHGAN, R. W. WHITE, AND M. RICHARDSON, *Enhancing expert finding using organizational hierarchies*, in Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, ECIR 09, Berlin, Heidelberg, 2009, Springer-Verlag, pp. 177–188.

[72] L. KAUFMAN AND P. J. ROUSSEEUW, *Finding Groups in Data: An Introduction to Cluster Analysis.*, John Wiley, 1990.

[73] J. W. KIM, B. H. LEE, M. J. SHAW, H.-L. CHANG, AND M. NELSON, *Application of decision-tree induction techniques to personalized advertisements on internet storefronts*, Int. J. Electron. Commerce, 5 (2001), pp. 45–62.

[74] T. KOHONEN, M. R. SCHROEDER, AND T. S. HUANG, eds., *Self-Organizing Maps*, Springer-Verlag, Berlin, Heidelberg, 3rd ed., 2001.

[75] H. J. KOOK, *Profiling multiple domains of user interests and using them for personalized web support*, in Proceedings of the 2005 International Conference on Advances in Intelligent Computing - Volume Part II, ICIC 05, Berlin, Heidelberg, 2005, Springer-Verlag, pp. 512–520.

[76] B. LANTZ, *Machine Learning with R*, Packt Publishing, 2013.

[77] W. S. LEE AND B. LIU, *Learning with positive and unlabeled examples using weighted logistic regression*, in Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ICML 03, AAAI Press, 2003, pp. 448–455.

[78] H. LI, S. JIN, AND S. LI, *A hybrid model for experts finding in community question answering*, 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, (2015), pp. 176–185.

[79] C. LIANG, Y. ZHANG, P. SHI, AND Z. HU, *Learning very fast decision tree from uncertain data streams with positive and unlabeled samples*, Inf. Sci., 213 (2012), pp. 50–67.

[80] L. LIN, Z. XU, Y. DING, AND X. LIU, *Finding topic-level experts in scholarly networks*, Scientometrics, 97 (2013), pp. 797–819.

[81] S. LIN, W. HONG, D. WANG, AND T. LI, *A survey on expert finding techniques*, Journal of Intelligent Information Systems, 49 (2017), pp. 255–279.

[82] B. LIU, Y. DAI, X. LI, W. S. LEE, AND P. S. YU, *Building text classifiers using positive and unlabeled examples*, in Proceedings of the Third IEEE International Conference on Data Mining, ICDM 03, Washington, DC, USA, 2003, IEEE Computer Society, pp. 179–186.

[83] B. LIU, W. S. LEE, P. S. YU, AND X. LI, *Partially supervised classification of text documents*, in Proceedings of the Nineteenth International Conference on Machine Learning, ICML 02, San Francisco, CA, USA, 2002, Morgan Kaufmann Publishers Inc., pp. 387–394.

[84] J. LIU, B. LI, B. LIU, AND Q. LI, *Topic-centric candidate priors for expert finding models*, Communications in Computer and Information Science, 391 (2013), pp. 253–262.

[85] M. LIU, Y. LIU, AND Q. YANG, *Predicting best answerers for new questions in community question answering*, in Proceedings of the 11th International Conference on Web-age Information Management, WAIM 10, Berlin, Heidelberg, 2010, Springer-Verlag, pp. 127–138.

[86] X. LIU, W. B. CROFT, AND M. KOLL, *Finding experts in community-based question-answering services*, in Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM 05, New York, NY, USA, 2005, ACM, pp. 315–316.

[87] X. LIU, S. YE, X. LI, Y. LUO, AND Y. RAO, *Zhihurank: A topic-sensitive expert finding algorithm in community question answering websites*, in Advances in Web-Based Learning – ICWL 2015, F. W. Li, R. Klamma, M. Laanpere, J. Zhang, B. F. Manjón, and R. W. Lau, eds., Cham, 2015, Springer International Publishing, pp. 165–173.

[88] Z. LIU, K. LI, AND D. QU, *Knowledge graph based question routing for community question answering*, in Neural Information Processing, D. Liu, S. Xie, Y. Li, D. Zhao, and E.-S. M. El-Alfy, eds., Cham, 2017, Springer International Publishing, pp. 721–730.

[89] S. LLOYD, *Least squares quantization in pcm*, IEEE Trans. Inf. Theor., 28 (2006), pp. 129–137.

[90] S. LOEB, *Architecting personalized delivery of multimedia information*, Commun. ACM, 35 (1992), pp. 39–47.

[91] P. LOPS, M. DE GEMMIS, AND G. SEMERARO, *Content-based recommender systems: State of the art and trends*, in Recommender Systems Handbook, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds., Springer US, Boston, MA, 2011, pp. 73–105.

[92] J. LU, D. WU, M. MAO, W. WANG, AND G. ZHANG, *Recommender system application developments*, Decis. Support Syst., 74 (2015), pp. 12–32.

[93] C. MACDONALD AND I. OUNIS, *Voting techniques for expert search*, Knowl. Inf. Syst., 16 (2008), pp. 259–280.

[94] J. MACQUEEN, *Some methods for classification and analysis of multivariate observations*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and

Probability, Volume 1: Statistics, Berkeley, Calif., 1967, University of California Press, pp. 281–297.

[95] Y. MASUDA, *Information Society as Post-Industrial Society*, World Future Society, 1980.

[96] J. P. MCGOWAN, N. KUSHMERICK, AND B. SMYTH, *Who do you want to be today? web personae for personalised information access*, in Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 02, Berlin, Heidelberg, 2002, Springer-Verlag, pp. 514–517.

[97] P. MCNAMEE AND J. MAYFIELD, *Character n-gram tokenization for european language text retrieval*, Inf. Retr., 7 (2004), pp. 73–97.

[98] M. MCTEAR, Z. CALLEJAS, AND D. GRIOL, *The Conversational Interface: Talking to Smart Devices*, Springer Publishing Company, Incorporated, 1st ed., 2016.

[99] D. MIMNO AND A. MCCALLUM, *Expertise modeling for matching papers with reviewers*, in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 07, New York, NY, USA, 2007, ACM, pp. 500–509.

[100] M. MOHRI, A. ROSTAMIZADEH, AND A. TALWALKAR, *Foundations of Machine Learning*, The MIT Press, 2012.

[101] S. MOMTAZI AND F. NAUMANN, *Topic modeling for expert finding using latent dirichlet allocation*, Wiley Int. Rev. Data Min. and Knowl. Disc., 3 (2013), pp. 346–353.

[102] F. NARDUCCI, C. MUSTO, G. SEMERARO, P. LOPS, AND M. DE GEMMIS, *Exploiting big data for enhanced representations in content-based recommender systems*, in E-Commerce and Web Technologies, C. Huemer and P. Lops, eds., Berlin, Heidelberg, 2013, Springer Berlin Heidelberg, pp. 182–193.

[103] P. T. NGUYEN, K. ECKERT, A. RAGONE, AND T. DI NOIA, *Modification to k-medoids and clara for effective document clustering*, in Foundations of Intelligent Systems, M. Kryszkiewicz, A. Appice, D. Slkezak, H. Rybinski, A. Skowron, and Z. W. Raś, eds., Cham, 2017, Springer International Publishing, pp. 481–491.

[104] N. NIKZAD KHASMAKHI, M. BALAFAR, AND M. R. FEIZI DERAKHSHI, *The state-of-the-art in expert recommendation systems*, Engineering Applications of Artificial Intelligence, 82 (2019), pp. 126–147.

[105] M. NILASHI, M. SALAHSHOUR, O. IBRAHIM, A. MARDANI, M. D. ESFAHANI, AND N. ZAKUAN, *A new method for collaborative filtering recommender systems: The case of yahoo! movies and tripadvisor datasets*, in Journal of soft computing and decision support systems, 2016, pp. 44–46.

[106] M. PACELLA, A. GRIECO, AND M. BLACO, *On the use of self-organizing map for text clustering in engineering change process analysis*, Intell. Neuroscience, 2016 (2016), p. 7.

[107] F. PALAMARA, F. PIGLIONE, AND N. PICCININI, *Self-organizing map and clustering algorithms for the analysis of occupational accident databases*, Safety Science, 49 (2011), pp. 1215–1230.

[108] M. PAVAN AND E. W. D. LUCA, *Semantic-based expert search in textbook research archives*, in Proceedings of the 5th International Workshop on Semantic Digital Archives co-located with 19th International Conference on Theory and Practice of Digital Libraries (TPDL 2015), Poznan, Poland, September 18, 2015., 2015, pp. 18–29.

[109] M. PAZZANI AND D. BILLSUS, *Learning and revising user profiles: The identification of interesting web sites*, Mach. Learn., 27 (1997), pp. 313–331.

[110] M. J. PAZZANI AND D. BILLSUS, *Content-based recommendation systems*, in The Adaptive Web, P. Brusilovsky, A. Kobsa, and W. Nejdl, eds., Springer-Verlag, Berlin, Heidelberg, 2007, pp. 325–341.

[111] D. PETKOVA AND W. B. CROFT, *Hierarchical language models for expert finding in enterprise corpora*, in Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 06, Washington, DC, USA, 2006, IEEE Computer Society, pp. 599–608.

[112] M. C. PLESSIS, G. NIU, AND M. SUGIYAMA, *Class-prior estimation for learning from positive and unlabeled data*, Mach. Learn., 106 (2017), pp. 463–492.

[113] R. K. PON, A. F. CARDENAS, D. BUTTLER, AND T. CRITCHLOW, *Tracking multiple topics for finding interesting articles*, in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 07, New York, NY, USA, 2007, ACM, pp. 560–569.

[114] D. POWERS, *Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation*, Mach. Learn. Technol., 2 (2008).

[115] M. QU, G. QIU, X. HE, C. ZHANG, H. WU, J. BU, AND C. CHEN, *Probabilistic question recommendation for question answering communities*, in Proceedings of the 18th International Conference on World Wide Web, WWW 09, New York, NY, USA, 2009, ACM, pp. 1229–1230.

[116] S. RANI, K. RAJU, AND V. KUMARI, *Expert finding system using latent effort ranking in academic social networks*, International Journal of Information Technology and Computer Science, 7 (2015), pp. 21–27.

[117] F. RIAHI, Z. ZOLAKTAF, M. SHAFIEI, AND E. MILIOS, *Finding expert users in community question answering*, in Proceedings of the 21st International Conference on World Wide Web, WWW 12 Companion, New York, NY, USA, 2012, ACM, pp. 791–798.

[118] F. RIBADAS, L. DE CAMPOS, J. FERNÁNDEZ-LUNA, AND J. HUETE, *Concept profiles for filtering parliamentary documents.*, Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, (2015), pp. 409–416.

[119] F. RICCI, L. ROKACH, B. SHAPIRA, AND P. B. KANTOR, *Recommender Systems Handbook*, Springer-Verlag, Berlin, Heidelberg, 1st ed., 2010.

[120] C. J. V. RIJSBERGEN, *Information Retrieval*, Butterworth-Heinemann, Newton, MA, USA, 2nd ed., 1979.

[121] V. RIJSBERGEN, *A theoretical basis for the use of co-occurence data in information retrieval*, J Doc, 33 (1977).

[122] S. ROBERTSON AND H. ZARAGOZA, *The probabilistic relevance framework: Bm25 and beyond*, Found. Trends Inf. Retr., 3 (2009), pp. 333–389.

[123] S. E. ROBERTSON, *The Probability Ranking Principle in IR*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 281–286.

[124] L. ROKACH AND O. MAIMON, *Clustering methods*, in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, eds., Springer US, Boston, MA, 2005, pp. 321–352.

[125] M. ROSEN-ZVI, T. GRIFFITHS, M. STEYVERS, AND P. SMYTH, *The author-topic model for authors and documents*, in Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI 04, Arlington, Virginia, United States, 2004, AUAI Press, pp. 487–494.

[126] M. ROUX, *A comparative study of divisive and agglomerative hierarchical clustering algorithms*, J. Classif., 35 (2018), pp. 345–366.

[127] S. RUSSELL AND P. NORVIG, *Artificial Intelligence: A Modern Approach*, Prentice Hall Press, USA, 3rd ed., 2009.

[128] J. RYBAK, K. BALOG, AND K. NØRVÅG, *Temporal expertise profiling*, in Advances in Information Retrieval, M. de Rijke, T. Kenter, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, and K. Hofmann, eds., Cham, 2014, Springer International Publishing, pp. 540–546.

[129] F. SAAD, O. MOHAMED, AND R. AL-QUTAISH, *Comparison of hierarchical agglomerative algorithms for clustering medical documents*, International Journal of Software Engineering and Applications, 3 (2013), pp. 1–15.

[130] G. SALTON, *The SMART Retrieval System - Experiments in Automatic Document Processing*, Prentice-Hall, Inc., USA, 1971.

[131] G. SALTON AND C. BUCKLEY, *Term-weighting approaches in automatic text retrieval*, Inf. Process. Manage., 24 (1988), pp. 513–523.

[132] J. SAN PEDRO AND A. KARATZOGLOU, *Question recommendation for collaborative question answering systems with rankslda*, in Proceedings of the 8th ACM Conference on Recommender Systems, RecSys 14, New York, NY, USA, 2014, ACM, pp. 193–200.

[133] J. SCHAFER, J. KONSTAN, AND J. RIEDL, *Recommender systems in e-commerce*, in Proceedings of the 1st ACM Conference on Electronic Commerce, EC 1999, ACM International Conference Proceeding Series, 12 1999, pp. 158–166.

[134] S. SCHIAFFINO AND A. AMANDI, *Intelligent user profiling*, in Artificial Intelligence, M. Bramer, ed., Springer-Verlag, Berlin, Heidelberg, 2009, pp. 193–216.

[135] F. SEBASTIANI, *Machine learning in automated text categorization*, ACM Comput. Surv., 34 (2002), pp. 1–47.

[136] P. SERDYUKOV AND D. HIEMSTRA, *Modeling documents as mixtures of persons for expert finding*, in Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, ECIR 08, Berlin, Heidelberg, 2008, Springer-Verlag, pp. 309–320.

[137] N. SHAH AND S. MAHAJAN, *Document clustering: A detailed review*, International Journal of Applied Information Systems, 4 (2012), pp. 30–38.

[138] J. SHAMIN AND C. NEUHOLD, *'connecting europe': The use of 'new' information and communication technologies within european parliament standing committees.*, The Journal of Legislative Studies 13, (2007), pp. 388–402.

[139] G. L. SOMLO AND A. E. HOWE, *Incremental clustering for profile maintenance in information gathering web agents*, in Proceedings of the Fifth International Conference on Autonomous Agents, AGENTS 01, New York, NY, USA, 2001, ACM, pp. 262–269.

[140] A. STARCZEWSKI AND A. KRZYŻAK, *Performance evaluation of the silhouette index*, in Artificial Intelligence and Soft Computing, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, eds., Cham, 2015, Springer International Publishing, pp. 49–58.

[141] R. SUBHASHINI AND V. S. KUMAR, *A roadmap to integrate document clustering in information retrieval*, Int. J. Inf. Retr. Res., 1 (2011), pp. 31–44.

[142] M. TINGHUAI, Z. JINJUAN, T. MEILI, T. YUAN, A.-D. ABDULLAH, A.-R. MZNAH, AND L. SUNGYOUNG, *Social network and tag sources based augmenting collaborative recommender system*, IEICE Transactions, 98-D (2015), pp. 902–910.

[143] A. TJOA, M. HOFFERER, G. EHRENTRAUT, AND P. UNTERSMEYER, *Applying evolutionary algorithms to the problem of information filtering.*, Proceedings of the 8th International Workshop on Database and Expert Systems Applications, (1997), pp. 450–458.

[144] G. TSOUMAKAS, I. KATAKIS, AND I. VLAHAVAS, *Mining Multi-label Data*, Springer US, Boston, MA, 2010, pp. 667–685.

[145] J. URE, *Lexical density and register differentiation.*, G. Perren, J.L.M. Trim (Eds.), Applications of Linguistics, London: Cambridge University Press, (1971), pp. 443–452.

[146] W. WEI, G. CONG, C. MIAO, F. ZHU, AND G. LI, *Learning to find topic experts in twitter via different relations*, IEEE Transactions on Knowledge and Data Engineering, 28 (2016), pp. 1764–1778.

[147] D. H. WIDYANTORO, J. YIN, M. SEIF, E. NASR, L. YANG, A. ZACCHI, AND J. YEN, *Alipes: A swift messenger in cyberspace*, in Proceedings of AAAI Spring Symposium on Intelligent Agents in Cyberspace, 1999, pp. 62–67.

[148] H. WU, Y. PEI, AND J. YU, *Hidden topic analysis based formal framework for finding experts in metadata corpus*, in 8th IEEE/ACIS International Conference on Computer and Information Science, 01 2009, pp. 369–374.

[149] H. WU, Y. WANG, AND X. CHENG, *Incremental probabilistic latent semantic analysis for automatic question recommendation*, in Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 08, New York, NY, USA, 2008, ACM, pp. 99–106.

[150] F. XU, Z. JI, AND B. WANG, *Dual role model for question recommendation in community question answering*, in Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 12, New York, NY, USA, 2012, ACM, pp. 771–780.

[151] L. YANG, M. QIU, S. GOTTIPATI, F. ZHU, J. JIANG, H. SUN, AND Z. CHEN, *Cqarank: jointly model topics and expertise in community question answering*, in CIKM, 2013, pp. 99–108.

[152] H. YU, J. HAN, AND K. C.-C. CHANG, *Pebl: Positive example based learning for web page classification using svm*, in Proceedings of the Eighth ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, KDD 02, New York, NY, USA, 2002, ACM, pp. 239–248.

[153] H. YU, K. INOUE, K. HARA, AND K. URAHAMA, *A robust k-means for document clustering*, Journal of the Institute of Industrial Applications Engineers, 6 (2018), pp. 60–65.

[154] S. YUAN, Y. ZHANG, J. TANG, AND J. CABOTÁ, *Expert finding in community question answering: A review*, Artificial Intelligence Review, (2018).

[155] S. ZAHRA, M. A. GHAZANFAR, A. KHALID, M. A. AZAM, U. NAEEM, AND A. PRUGEL-BENNETT, *Novel centroid selection approaches for kmeans-clustering based recommender systems*, Inf. Sci., 320 (2015), pp. 156–189.

[156] J. ZAMORA, *Recent Advances in High-Dimensional Clustering for Text Data*, Springer International Publishing, Cham, 2017, pp. 323–337.

[157] H.-J. ZENG, Z. CHEN, AND W.-Y. MA, *A unified framework for clustering heterogeneous web objects*, in Proceedings of the 3rd International Conference on Web Information Systems Engineering, WISE 02, Washington, DC, USA, 2002, IEEE Computer Society, pp. 161–172.

[158] C. ZHAI AND J. LAFFERTY, *A study of smoothing methods for language models applied to ad hoc information retrieval*, in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 01, New York, NY, USA, 2001, Association for Computing Machinery, pp. 334–342.

[159] B. ZHANG AND W. ZUO, *Learning from positive and unlabeled examples: A survey*, in 2008 International Symposiums on Information Processing, May 2008, pp. 650–654.

[160] T. ZHAO, N. BIAN, C. LI, AND M. LI, *Topic-level expert modeling in community question answering*, in SDM, 2013, pp. 776–784.

[161] Y. ZHAO, G. KARYPIS, AND U. FAYYAD, *Hierarchical clustering algorithms for document datasets*, Data Min. Knowl. Discov., 10 (2005), pp. 141–168.

[162] G. ZHOU, J. ZHAO, T. HE, AND W. WU, *An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities*, Know.-Based Syst., 66 (2014), pp. 136–145.

[163] H. ZHU, H. CAO, H. XIONG, E. CHEN, AND J. TIAN, *Towards expert finding by leveraging relevant categories in authority ranking*, in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 11, New York, NY, USA, 2011, ACM, pp. 2221–2224.

[164] J. ZOBEL AND A. MOFFAT, *Inverted files for text search engines*, ACM Comput. Surv., 38 (2006), p. 6.