Representaciones multinivel para el Filtrado de Información

por

Adrian Fonseca Bruzón

Tesis sometida como requisito parcial para
obtener el grado de

DOCTOR EN CIENCIAS CON ESPECIALIDAD EN EL ÁREA DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Tonantzintla, Puebla, México
febrero, 2019

Dirigida por:

Dr. Aurelio López López
Coordinación de Ciencias Computacionales
INAOE, México

Dr. José Eladio Medina Pagola
Vicerrectoría Primera
Universidad de las Ciencias Informáticas (UCI), Cuba

© Coordinación de Ciencias Computacionales

INAOE 2019

Luis Enrique Erro 1
Sta. Ma. Tonantzintla
72840, Puebla, México

Resumen

Los métodos de Filtrado Adaptativo de Información tienen la finalidad de permitir a los usuarios concentrarse en los documentos que le son de interés, sin necesidad de realizar una exploración exhaustiva de toda la información que continuamente es generada. Estos métodos, a diferencia de los enfoques de filtrado tradicionales, permiten a los usuarios proporcionar al sistema retroalimentación concerniente a su funcionamiento, lo cual posibilita que él mismo se ajuste con el transcurrir del tiempo a las necesidades cambiantes de los usuarios y el flujo de información. Estos métodos emplean un perfil de usuario para representar las necesidades de información de los usuarios.

Por otro lado, los humanos solemos organizar la información en los documentos de forma lógica e intencionada. Esta organización, a la cual llamaremos estructura textual, puede estar compuesta por secciones, capítulos, párrafos, u oraciones; según sea el tipo de documento. Esta estructura facilita la comprensión del contenido que en ellos deseamos transmitir. Sin embargo, esta estructura en la cual solemos codificar el contenido semántico de la información no suele ser aprovechada por los métodos de filtrado para la construcción del perfil de los usuarios.

Este trabajo constituye una primera aproximación orientada a llenar ese vacío en la tarea del filtrado. En él proponemos dos tipos diferentes de representación en las cuales es tomada en consideración la estructura textual de los documentos. La primera de ellas basada en los conjuntos de términos frecuentes y la segunda en el Indexado Aleatorio.

Adicionalmente, en este trabajo se proponen métodos para la obtención de estas representaciones tomando en consideración el desbalance presente entre los documentos que satisfacen la necesidad de información de los usuarios, así como al problema del Inicio en Frío (contar con muy poca información) en la construcción inicial del perfil de usuario.

Los experimentos realizados permitieron valorar el impacto de las representaciones empleadas en la tarea de filtrado adaptativo de documentos.

Abstract

Document Filtering has the purpose of allowing users to concentrate on the documents that are of interest to them, without having to carry out an exhaustive exploration of all the information that is continuously generated. One variation of the typical document filtering systems is commonly referred to as Adaptive Document Filtering. This variation allows users to provide the system with information about its behavior, which allows it to adjust to the changing in users' needs and information stream. These methods employ a user profile for representing the users' information needs.

On the other hand, humans tend to organize information in documents in a logical and intentional way. This organization, which we will call textual structure, can be composed of sections, chapters, paragraphs, or sentences; according to the type of document. This structure facilitates the understanding of the content that we want to transmit in them. However, this structure, in which we usually encode the semantic content of the information, is not usually exploited by the filtering methods for the construction of the user profile.

This work constitutes a first approximation aiming at filling that gap in the filtering task. We propose two different types of representation in which the textual structure of the documents is taken into account. The first of them based on sets of frequent terms and the second one on Random Indexing.

Additionally, we propose methods for obtaining these representations taking into

consideration the presence of imbalance between the documents that satisfy the information needs of the users, as well as the Cold Start problem (having scarce information) during the initial construction of the user profile.

The experiments carried out allow us to assess the impact on the filtering task of the proposed representations and the methods for obtaining them.

Agradecimientos

Con este trabajo se llega al final de una etapa, la cual ha estado llena de grandes sacrificios, emociones y esfuerzo. Una etapa que no he transitado solo, más bien ha sido todo lo contrario. Durante todo este tiempo he contado con la guía, ayuda, apoyo y amistad de muchas personas, a las cuales llega el momento de dedicarle unas pocas palabras de agradecimiento.

Primeramente debo comenzar por mis asesores, los cuales me han acompañado con mucha paciencia y perseverancia, siempre dispuestos a brindar consejos e ideas para perfeccionar mi formación. Igualmente, creo que es oportuno agradecer a los sinodales de este trabajo, que igualmente me han acompañado durante este período, y los cuales con sus comentarios oportunos han logrado que éste sea un mejor trabajo. Finalmente, debo agradecer al Laboratorio del Tecnologías del Lenguaje, el INAOE, el Conacyt y en general al Gobierno de México por el apoyo material brindado sin el cual todo este esfuerzo no habría sido posible.

Este período no solo me ha permitido crecer en mi formación profesional, sino que me ha permitido a la vez cultivar nuevas amistades, las cuales constituyen un gran tesoro y que espero poder continuar cultivando en el futuro. Las nuevas amistades, de conjunto con las anteriores, han sido una piedra angular en el apoyo y los consejos en los momentos oportunos. Muchas gracias a todos porque creo que he podido recibir de ustedes mucho más de lo que he podido yo otorgarles a cambio.

Sin lugar a dudas, un agradecimiento especial debo dedicarlo a mi familia, porque ellos han sido una fuente de apoyo invaluable para poder llegar a puerto seguro el desarrollo de la investigación de doctorado. Muchas gracias.

E n cuatro años hay muchas personas que influyen en mayor o menor medida en la vida de una persona, y claro está, yo no soy la excepción. Pero resulta imposible realizar una enumeración exhaustiva de todas y cada una de las personas que me han ayudado o inspirado para poder lograr este sueño, por lo que he evitado mencionar

nombres para no pecar de dejar fuera a nadie. Por suerte, tengo la certeza de que todos los que han colaborado no necesitan que se les mencione de forma explícita en estas líneas, porque no hacen las cosas esperando una retribución a cambio, y aunque puedo decirles muchas cosas creo que todo se resume en mi más sincero agradecimiento.

Gracias a todos,
Adrian Fonseca Bruzón.
Tonantzintla, Puebla, México.
febrero, 2019.

Índice general

Índice de figuras

Listado de Tablas

Capítulo 1

Introducción

Cada día la cantidad de información que se encuentra disponible en Internet crece a ritmos gigantescos. Este crecimiento acelerado en la cantidad de información que es generada diariamente trae consigo que a los usuarios les resulte complicado mantenerse actualizados con respecto a aquellos temas que les son de interés. Incluso, este auge en la generación de documentos puede ocasionar que los usuarios corran el riesgo de quedar abrumados por este flujo creciente de información disponible [Albakour et al., 2013][Hanani et al., 2001a].

Para dar respuesta a esta problemática se han propuesto en la literatura varias soluciones que facilitan el acceso a la información que satisface a nuestras necesidades de información, sin tener que inspeccionar de forma manual todos los documentos. Dos de estas soluciones son la Recuperación de Información y el Filtrado de Información. Estas dos tareas son muy similares en el sentido de que ambas tienen la finalidad de proveer al usuario con información útil con respecto a una determinada necesidad de información. Sin embargo, existen grandes diferencias entre ellas. El Filtrado de Información está dirigido a satisfacer necesidades de información que perduran durante un período largo de tiempo, mientras que la Recuperación de Información se centra en consultas de corta duración, las cuales son con frecuencia descartadas una vez terminada la sesión de búsqueda.

En la literatura podemos encontrar diferentes tareas asociadas al Filtrado de Información, orientadas a atender las necesidades de información a largo plazo de los usuarios. Una de ellas son los Sistemas de Recomendación [Ricci et al., 2015]. En ellos la presentación de la información hacia los usuarios viene dada en forma de sugerencias.

Estos sistemas han alcanzado una gran importancia con el desarrollo vertiginoso que ha tenido el comercio electrónico en los últimos años. En los Sistemas de Recomendación, la información disponible del usuario es almacenada por el sistema para poder realizar con posterioridad la selección de las sugerencias que se le harán. Estos sistemas presentan sugerencias personalizadas a los usuarios mediante el aprendizaje de sus intereses a partir de la interacción con el sistema.

Otra tarea relacionada al Filtrado de Información es el monitoreo de un flujo de documentos, para seleccionar aquellos que se ajustan a las necesidades de información de los usuarios. A diferencia de los Sistemas de Recomendación, en el filtrado de flujos de información, ésta no es almacenada para posteriormente tomar la decisión de mostrar o no los documentos al usuario. Por el contrario, cada documento del flujo es analizado y en ese momento se determina si se ajusta o no a las necesidades de información de los usuarios. El filtrado de flujos de información es aplicado en varias actividades. Por ejemplo, un analista de información que debe monitorear flujos de noticias relacionados con determinadas temáticas para elaborar informes o resúmenes a directivos o personas encargadas de la toma de desiciones; el desarrollo de herramientas para el filtrado de correo basura en un servidor de correos; o la generación de alertas para investigadores sobre nuevos artículos científicos relacionados con sus intereses de investigación.

Cuando se monitorean flujos de documentos, el objetivo de un sistema de Filtrado de Información (FI) es clasificar los documentos provenientes del flujo de información, en Relevantes o No relevantes, de acuerdo con el interés de un usuario en particular [Lanquillon and Renz, 1999]. Estos sistemas de FI se dirigen a necesidades de información relativamente estables a largo plazo, aunque usualmente permiten que estos intereses puedan modificarse de forma gradual en el tiempo [Jones and Brown, 2003]. En tal caso se les conoce como sistemas de Filtrado de Información Adaptativos [Zhang, 2009].

El monitoreo continuo del flujo de información por parte de un sistema de FI facilita

a los usuarios concentrarse solamente en los documentos que les son ofrecidos por el sistema y mantenerse actualizado con los temas que les son de interés, sin necesidad de realizar una exploración exhaustiva de la información disponible, o de aquella que se está generando continuamente.

Uno de los componentes más importantes de un sistema de filtrado consiste en la representación del interés del usuario, usualmente denominada perfil. Este perfil se mantiene durante el tiempo que dura la necesidad o interés de información. En el caso del Filtrado Adaptativo, este perfil de usuario puede ser refinado si se cuenta con retroalimentación explícita o implícita por parte del usuario.

De forma general podemos encontrar dos enfoques diferentes a la tarea de FI: el filtrado basado en contenido y el colaborativo [Zhang et al., 2014][Sharma, 2018]. Estos enfoques difieren en la forma en que son representados y comparados los perfiles de usuarios y los documentos. En el filtrado basado en contenido, tanto el perfil del usuario como los documentos son representados empleando características extraídas de los propios documentos. En el caso de aquellos que siguen el enfoque colaborativo, los elementos son caracterizados por la puntuación que reciben por parte de usuarios con intereses similares. En este caso, el perfil del usuario es construido mediante la valoración que emite éste sobre los documentos y su semejanza con lo expresado por otros usuarios con intereses similares. En los Sistemas de Recomendación, el enfoque colaborativo ha sido ampliamente estudiado, aunque encontramos también enfoques que siguen el filtrado basado en contenido, e incluso estudios más recientes siguen una estrategia híbrida. Por el contrario, cuando el FI es empleado al monitoreo de flujos el enfoque seguido es el basado en contenido dado que las necesidades de información de los usuarios suelen variar mucho de unos a otros. En este trabajo nos centraremos en los sistemas de filtrado basados en contenido empleados para monitorear flujos de documentos.

1.1 Motivación

Monitorear un flujo de documentos con la finalidad de seleccionar aquellos que satisfacen las necesidades de información de los usuarios, posibilitando descartar aquellos cuyo contenido no es de interés, es un elemento de mucha utilidad en nuestros días. La posibilidad de mantenerse actualizado sobre determinadas temáticas, sin la necesidad de realizar una exploración exhaustiva o permanente de la información que se genera continuamente, es una funcionalidad que puede incorporar un importante valor agregado en los sistemas que lo incorporen. Esta característica resulta aún más valiosa si además el sistema es capaz de actualizar la información del perfil del usuario con respecto a los intereses de información de los usuarios.

Modelar correctamente las necesidades de información de los usuarios en un sistema de filtrado de información resulta todo un reto. Las personas son muy diferentes entre sí, y de igual manera son así de diferentes las necesidades de información de un usuario a otro. Entre los intereses de información encontramos usuarios interesados en temas generales, por ejemplo en béisbol, actividades culturales, etc. Otros, por el contrario, pueden estar interesados en contenidos mucho más específicos, como pueden ser los accidentes de tránsito en los que se encuentran involucrados autobuses escolares o incidentes aéreos provocados por pasajeros con mala conducta; incluso podemos encontrar intereses que pudieran considerarse raros en determinados contextos como puede ser la guerra bacteriológica. Por demás, un usuario particular puede estar interesado en varios tópicos diferentes y no todos son de la misma naturaleza, unos pueden ser generales y otros muy específicos. Este comportamiento tan diferente entre los intereses de un usuario a otro es, sin lugar a dudas, un reto para el diseño de aplicaciones que implementan el filtrado de información.

Adicionalmente a los problemas antes mencionados, en la modelación de un sistema

de filtrado adaptativo de documentos hay que tomar en consideración además otros fenómenos presentes, como son el Inicio en Frío y el desbalance en los documentos que satisfacen las necesidades de información de los usuarios.

Cada vez que un usuario plantea una nueva necesidad de información, el sistema tiene que hacer frente a un problema de escasez de información, conocido como Inicio en Frío. Cuando el usuario tiene una nueva necesidad de información provee al sistema con una consulta que expresa su necesidad de información, y quizás algunos pocos documentos que se relacionan en cierta medida con la necesidad expresada. Realizar una correcta modelación de la necesidad de información del usuario con tan poca información es un proceso muy complejo. Cuando se cuenta con tan poca información se corre el riesgo de crear un sistema que abrume al usuario con mucha información que no es de su interés; o, por el contrario, se pierdan muchos documentos potencionalmente relevantes para el usuario, pero por la poca información no se logra realizar un emparejamiento acertado entre la información almacenada en el perfil y el contenido de los documentos.

Otro fenómeno al que debe prestársele atención cuando se diseña un sistema de filtrado de información está relacionado con el desbalance en los documentos disponibles para la modelación de la necesidad de información del usuario. Los tópicos de interés de un usuario pueden estar compuestos a su vez por varios subtópicos. El origen de estos subtópicos puede estar relacionado con diversas características propias del tópico de interés, como pueden ser: especificaciones dentro de la temática general, nuevos eventos o aristas diferentes en el tópico de interés, etc. Estos subtópicos no necesariamente tienen que encontrarse igualmente representados ni en la información suministrada por el usuario para la modelación de la necesidad de información, ni en el flujo de documentos. Por ejemplo, en la aeronáutica civil serán más abundantes las noticias relacionadas con las aerolíneas, su comportamiento, servicios y rutas; luego, probablemente, alguna

información sobre accidentes e incidentes y finalmente encontraremos con una menor frecuencia documentos relacionados con leyes y regulaciones sobre el funcionamiento del sector, por citar un ejemplo. Por lo que es de esperar que un usuario interesado en esta temática sea capaz de suministrar una mayor retroalimentación relacionada con las subtemáticas más frecuentes y mucha menos con aquellas menos frecuentes, sin que ello implique que estas subtemáticas sean menos importantes o posean una menor relevancia para el usuario. Un sistema que no tome en consideración este fenómeno puede correr el riesgo de solamente brindar al usuario los documentos de interés más representados en el perfil e ignorar aquellas subtemáticas que sean menos frecuentes. Esta situación es aún más complicada si tomamos en consideración que el desbalance en la información puede tener un comportamiento no homogéneo a lo largo del tiempo. Puede que un subtópico tenga períodos de tiempos en los cuales la cantidad de información relacionada con un subtópico sea abundante, mientras que en otros apenas se mencione e incluso siquiera sea abordado. Por ejemplo, ante un accidente aéreo, la cantidad de noticias e información que se genera en los días siguientes al suceso aumenta considerablemente pero, con el paso del tiempo, el interés disminuye y con ello las noticias y documentos que lo abordan. Con el caso de las regulaciones relacionadas con la aeronáutica civil, la situación puede resultar mucho más compleja, pues puede pasar mucho tiempo sin que se tenga información de algún cambio en las mismas, sin que por ello sea un subtópico que no sea de interés para el usuario.

La mayoría de los trabajos reportados en la literatura que han atacado el problema del filtrado de información hacen uso del tradicional modelo de Bolsa de Palabras para la representación de los documentos, pese a que éste modelo no logra capturar las relaciones semánticas existentes entre los términos de un documento, comúnmente referida como semántica latente [Becker and Kuropka, 2003]. La Bolsa de Palabras solamente toma en consideración si las palabras conocidas aparecen o no en el documento, sin tomar

en cuenta dónde ellas ocurren. Por ejemplo, consideremos las oraciones siguientes:

- En el desarrollo de la web se sustenta el comercio electrónico y el correo mundial.
- El comercio mundial se sustenta en el correo electrónico y el desarrollo de la web.

En ellas podemos notar que, a pesar de contener las mismas palabras, su semántica difiere. La primera de éstas oraciones se centra en el comercio electrónico, mientras que la segunda no. Este tipo de diferencias no pueden ser distinguidas cuando es empleado el modelo de Bolsa de Palabras, ya que ambas oraciones se reducen al mismo conjunto de palabras.

El lenguaje natural es un gran reto para las Ciencias de la Computación. Por un lado, las palabras son ambiguas; es decir, una palabra puede tener diversos significados; por otro lado, varias palabras pueden ser empleadas para referirse a un mismo concepto.

Algunas formas de representación para los documentos han sido reportadas en la literatura que, a diferencia de la Bolsa de Palabras, no suponen que los términos presentes en los documentos son independiente entre sí. Entre ellas podemos encontrar el Indexado de Semántica Latente (LSI) [Deerwester et al., 1990], el Indexado Probabilístico de Semántica Latente (PLSI) [Hofmann, 1999], o la Asignación Latente de Dirichlet (LDA) [Blei et al., 2003]. Sin embargo, estos métodos son computacionalmente costosos, o requieren de tener completamente en memoria la matriz de frecuencias términos-documentos. En un contexto como el Filtrado Adaptativo de documentos, en los cuales se van procesando nuevos documentos a cada momento, con frecuencia ocurre que éstos aporten nuevos términos y su incorporación al perfil resulta costosa. Estas limitaciones reducen su aplicabilidad en la tarea que nos ocupa, donde ocurren actualizaciones frecuentes en la información disponible.

Otras formas de representación reportadas en la literatura son el Indexado Aleatorio y las representaciones basadas en patrones. El Indexado Aleatorio [Sahlgren, 2005]

puede constituir una alternativa viable, dado que este método es computacionalmente menos costoso y no requiere del acceso en memoria de toda la matriz de frecuencias términos-documentos.

Las representaciones basadas en patrones frecuentes, como es el caso del PTM (Pattern Taxonomy Mining) [Wu et al., 2004], tienen la ventaja de que para extraer las relaciones existentes entre los términos no requieren de información disponible más allá de la almacenada en los documentos existentes para la construcción del perfil. Por esta razon, estos métodos resultan más atractivos para ser empleados en un ambiente en línea.

No obstante, las representaciones anteriores, aún cuando no suponen que los términos son independientes entre sí, no toman en consideración el hecho de que los términos en los documentos mantienen diferentes relaciones en función del nivel de contexto en el cual se relacionan. Por ejemplo, si tomamos en consideración los términos *información* y *sistema*. La presencia de estos términos en un documento puede estar relacionada con la información de un sistema, un sistema de información, o incluso una relación mucho más general en la cual información y sistema no están relacionados de forma directa. Sin embargo, a medida que el contexto en el cual están relacionados los términos es más específico, como párrafo u oración, la relación entre los términos se vuelve menos general y gana mayor indicio el hecho de que estamos hablando de la información de un sistema o de un sistema de información. Por ejemplo, si consideramos la oración *La información privada fue suministrada por el sistema de impuestos del Ministerio*, en éste contexto los términos hacen referencia a la información almacenada en un sistema. Por último, si consideramos como contexto los sintagmas nominales, entonces con casi total seguridad estaremos ante un documento que se relaciona con los sistemas de información.

Pese a que el contexto en el que los términos se encuentran relacionados es una

información relevante para determinar de forma más efectiva si el contenido de un documento se ajusta a la información almacenada en el perfil del usuario, hasta el momento, no conocemos de alguna representación que tome en cuenta esta información.

Como hemos visto, el tomar en consideración diferentes niveles de contexto durante el proceso de construcción del perfil del usuario, y del procesamiento de los documentos, puede favorecer a los resultados alcanzados por los algoritmos de filtrado de información.

Es por ello que consideramos pertinente desarrollar una representación para los documentos y el perfil del usuario que no presuponga la independencia entre los términos y a su vez tome en consideración diferentes niveles de granularidad en los contextos entre los términos presentes en el documento.

La tarea de Filtrado Adaptativo impone grandes retos, entre ellos se encuentra la poca homogeneidad en la distribución de las muestras a lo largo del tiempo, el desbalance presente entre los diferentes subtópicos de interés que pueden surgir, y la consecuente posible escasez de información para la correcta modelación del perfil de los usuarios. Esta última situación puede mantenerse incluso durante todo el funcionamiento del sistema, cuando el interés del usuario es muy específico y se tienen pocos documentos que satisfagan su necesidad de información. El tomar en consideración estos elementos durante el diseño de un sistema de filtrado adaptativo puede resultar conveniente si se desea mejorar la efectividad y la experiencia de los usuarios.

1.2 Objetivos

El **objetivo general** de esta investigación es el diseño de una representación del perfil del usuario en la tarea del Filtrado Adaptativo de Documentos, la cual no suponga la existencia de independencia entre los términos, y se obtengan resultados con una eficacia similar o superior a los reportados en la literatura.

Nuestros objetivos específicos son:

1. Proponer una representación del perfil del usuario acorde a la tarea del Filtrado Adaptativo de Documentos, en la cual se consideren los nuevos términos que aparezcan en el flujo de información; así como diferentes contextos, determinados a partir de la estructura de los documentos.

2. Desarrollar un procedimiento para la obtención de la representación propuesta en el objetivo anterior.

3. Elaborar una solución para la tarea de Filtrado Adaptativo de Documentos empleando la representación del perfil propuesta.

4. Proponer una estrategia que permita disminuir el impacto del Inicio Frío (escasez inicial de información) en la construcción de la representación y con ello obtener mayor eficacia en las etapas iniciales del proceso de filtrado.

5. Diseñar una estrategia que permita disminuir el impacto del desbalance entre los diferentes subtópicos que conforman el interés del usuario.

1.3 Contribuciones

Las principales contribuciones de esta investigacin doctoral son las siguientes:

1. Representación del perfil del usuario empleando relaciones multinivel entre términos con Indexado Aleatorio.

2. Algoritmo de Filtrado Adaptativo de Documentos empleando representaciones multinivel basadas en Indexado Aleatorio y relaciones entre términos.

3. Estrategia para combatir el Inicio en Frío (escasez de información) en el funcionamiento del algoritmo de Filtrado y en la construcción de la representación.

4. Estrategia para disminuir el impacto del desbalance entre los diferentes subtópicos presentes en el interés del usuario.

5. Método que incorpora los elementos anteriores y que mejora la eficacia de representaciones previas en la tarea del filtrado adaptativo de documentos.

1.4 Organización del informe de tesis

En el capítulo 1 se presentó la introducción al trabajo, en el mismo se describió la motivación del trabajo desarrollado y se plantearon los objetivos de la tesis.

En el capítulo 2 se describen los conceptos y técnicas fundamentales relacionados con el desarrollo de esta tesis y que facilitan la comprensión de la misma.

En el capítulo 3 se abordan los diferentes enfoques que se han usado para resolver la problemática ligada al filtrado adaptativo de documentos. Se hace un análisis de los trabajos más importantes relacionados con nuestra propuesta. Se presenta un resumen de las propuestas hechas por los investigadores en el área y se realiza una diferenciación con nuestra propuesta.

En el capítulo 4 se presenta una representación del perfil del usuario basado en relaciones multinivel entre términos. También se describen los algoritmos para la extracción de las relaciones multinivel a partir de los documentos disponibles para la construcción del perfil. Se presentan además estrategias para mitigar el efecto del Inicio en Frío y el desbalance entre los subtópicos que satisfacen la necesidad de información del usuario.

En el capítulo 5 se muestra una extensión del Indexado Aleatorio en la cual se toma en consideración varios niveles de granularidad para la construcción de los vectores de contexto. Además, se muestra la aplicación del Indexado Aleatorio Multinivel a la tarea del Filtrado Adaptativo Multinivel.

En el capítulo 6, se propone la combinación de las representaciones multinivel de

términos con el Indexado Aleatorio multinivel. Se presenta la aplicación de ésta a la tarea del Filtrado Adaptativo. Finalmente, se presentan experimentos que comparan los resultados alcanzados con nuestras propuestas con respecto a los resultados reportados en el estado del arte.

En el capítulo 7 se hace un resumen del trabajo y se presentan las conclusiones, las contribuciones y el trabajo futuro propuesto como posible continuación de nuestro trabajo de investigación.

Capítulo 2

Fundamentos Teóricos

En este capítulo se presentan conceptos importantes para la comprensión de los capítulos subsecuentes. En primer lugar, en la sección 2.1, se explican las características fundamentales de un sistema de filtrado adaptativo de documentos. En la sección 2.3 se aborda el tema de la representación de los documentos evaluados. La sección 2.4 aborda la temática de las representaciones basadas en patrones y en la 2.5 se hace referencia a la definición de un clasificador. Finalmente se presentan los elementos básicos de las colecciones de documentos (sección 2.6) y las medidas de evaluación (sección 2.7) empleadas.

2.1 Filtrado Adaptativo de Documentos

Bajo el nombre de Filtrado de Información encontramos una variedad de procesos que tienen la finalidad de recuperar información para el usuario que la necesite [Belkin and Croft, 1992], presentándole solamente aquella información que le es relevante [Hanani et al., 2001b]. Los métodos de filtrado usualmente involucran un flujo de documentos de entrada, ya sean medios de difusión de noticias, u otras fuentes como pueden ser correos electrónicos o páginas web [Belkin and Croft, 1992].

Los sistemas de filtrado de información son diseñados y entrenados para satisfacer las necesidades de información de un usuario, o grupo de usuarios. Una vez entrenados, estos métodos comienzan a monitorear el flujo de documentos con la finalidad de diferenciar aquellos que se ajustan a los intereses del usuario de los que no.

Estos métodos suponen que tanto la naturaleza de la información como los intereses

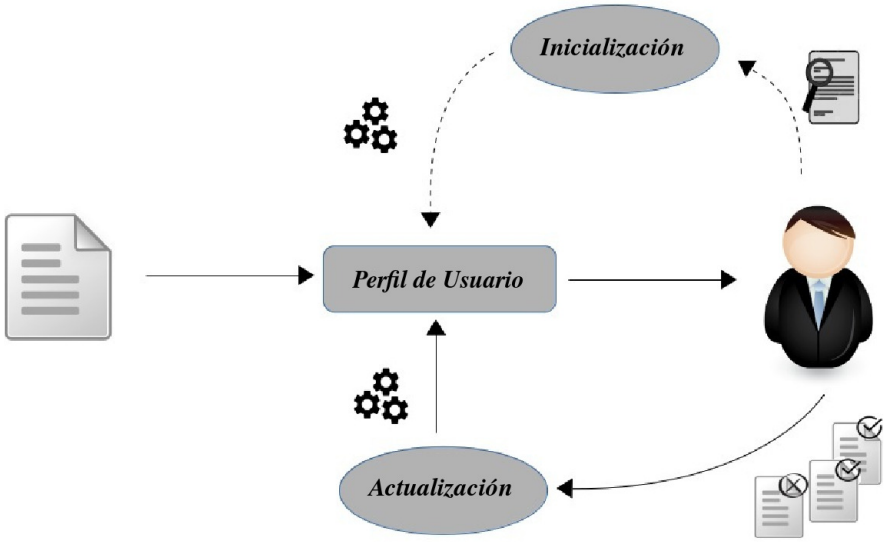


Figura 2.1: Estructura general de un sistema de Filtrado Adaptativo de Documentos.

de los usuarios se mantienen inamovibles con el tiempo, cuando la verdad es que las fuentes de información, y los intereses de los usuarios, pueden variar con el transcurrir del tiempo. Esto conlleva que un sistema, construido en un momento dado, con el paso del tiempo pueda quedar obsoleto y se requiera de la construcción de un nuevo sistema para adaptarlo a las nuevas necesidades.

Para dar solución a este problema fueron presentados una variedad de algoritmos de filtrado capaces de interactuar con el usuario y adaptar su comportamiento a las necesidades cambiantes de los usuarios. Estos algoritmos son conocidos como algoritmos de Filtrado Adaptativo de Documentos.

Al igual que un sistema de filtrado tradicional, uno adaptativo monitorea de forma continua un flujo de documentos textuales con el fin de separar para un usuario aquellos que se ajustan a sus intereses o necesidades de información, ignorando el resto. En la figura 2.1 se muestra el esquema de funcionamiento general de este tipo de sistemas. Aún cuando un sistema de Filtrado Adaptativo puede manipular varios usuarios a la vez, en la figura se muestra solamente uno con el fin de facilitar su comprensión.

Cada vez que se tiene una nueva necesidad de información, se debe proveer al sistema con la especificación del nuevo requerimiento. Esta especificación puede ser expresada por medio de una simple consulta, una breve descripción o algunos pocos documentos que reflejen el nuevo interés. Con esta información se crea un nuevo *perfil*, en el cual se modela el nuevo interés y se prepara al sistema para que comience la selección de los documentos que se ajusten a la necesidad de información especificada. Dado que los intereses de un usuario pueden ser completamente diferentes entre sí, cada nueva necesidad de información de un usuario suele ser tratada de forma independiente, creando un perfil propio para su gestión.

A medida que arriban nuevos documentos en el flujo de información, éstos son analizados con el fin de determinar si se ajustan o no a la necesidad de información especificada. Si el sistema determina que el documento se ajusta al perfil, éste es seleccionado para ser mostrado al usuario. En caso contrario, el documento es descartado. Es importante notar que el usuario tendrá acceso únicamente a aquellos documentos que el sistema considera que se ajustan a la necesidad de información del usuario especificada en el perfil. Aquellos documentos que son descartados por el sistema no son mostrados al usuario, aún cuando éstos pudieran haber sido relevantes para él.

A diferencia de un sistema de filtrado tradicional, en uno adaptativo el usuario tiene la posibilidad de retroalimentar al sistema indicando cuándo un documento recuperado se ajusta realmente a su necesidad de información y cuándo no. Estos documentos provistos al sistema en forma de retroalimentación permiten enriquecer la información empleada para la modelación de la necesidad de información. La finalidad de este proceso es que con el paso del tiempo, y en la medida que el usuario provee al sistema con nuevas muestras, éste sea capaz de identificar en el futuro documentos similares a los realmente relevantes, descartando aquellos con contenido similar a los que ya se conocen que no son de interés. La posibilidad de obtener retroalimentación por parte del

usuario puede ser explotada por los sistemas con el fin de explorar posibles subtópicos de interés para el usuario, de los cuales no se tiene hasta el momento información en el sistema; sin embargo, debe alcanzarse un equilibrio en esta exploración, la cual debe ser realizada sin abrumar al usuario con *falsas alarmas*.

Varias han sido las investigaciones realizadas en el área del filtrado. En este sentido, las conferencias TREC (Text REtrieval Conference[1]) son posiblemente el mejor espacio conocido para la evaluación y comparación de métodos de filtrado. Las conferencias TREC son patrocinadas por el NIST (*National Institute of Standards and Technology*[2]) y la DARPA (*Defense Advanced Research Projects Agency*[3]). Las conferencias TREC consisten en una serie de áreas de interés, en las que se definen un conjunto de tareas de Minería de Textos. La línea de Filtrado es una de ellas, en la cual la tarea de investigación principal es el filtrado adaptativo. Esta tarea está diseñada para modelar el proceso de filtrado de información partiendo desde la construcción inicial del perfil. En esta tarea, las necesidades de información del usuario se plantean estables en el tiempo. Los sistemas para la construcción del perfil inicial disponen de una especificación de la necesidad de información en forma de consulta (al estilo de las empleadas en los buscadores web como el de Google[4]); adicionalmente puede estar disponible una muy breve descripción de la necesidad de información, un número muy reducido de muestras relevantes (2 ó 3) y no se cuenta con muestras de documentos no relevantes. Cuando un nuevo documento se tiene que analizar, el sistema debe tomar la decisión de entregar o no el documento al usuario. Si el documento es seleccionado para ser entregado, inmediatamente se cuenta con la información de si el mismo era realmente relevante o no para este perfil, con el objetivo de simular el proceso en que el sistema es retroalimentado explícitamente por parte del usuario. Si el documento no fue considerado para entregar

[1] http://trec.nist.gov
[2] https://www.nist.gov/
[3] https://www.darpa.mil/
[4] https://www.google.com

al usuario, el sistema nunca tiene disponible información sobre cuál era el estado real del documento para el perfil. Una vez que el sistema toma la decisión de entregar o no un documento al usuario, la misma no puede ser modificada. Estas condiciones son un tanto estrictas y no siempre se ajustan a un entorno real donde se pueda desempeñar un sistema, no obstante son simples, razonables y permiten la comparación del desempeño de los diferentes sistemas. Este planteamiento ha inspirado tareas similares como la relacionada con Microblogs en la TREC y la tarea INFILE en la conferencia CLEF[5].

2.2 Estructura de los documentos

El contenido de los documentos no es una colección aleatoria de palabras. Por el contrario, cuando escribimos, seleccionamos las palabras que nos permitan transmitir las ideas que deseamos expresar. Inclusive la mera selección de las palabras no es suficiente para que sea comprendido el contenido en un documento. Estas palabras deben respetar la gramática del idioma en el cual se elabora el documento [Ventura and Marañón, 2013].

Un documento suele estar compuesto por varias ideas alrededor de su tema central, las cuales son desarrolladas a lo largo de su contenido. Cuando escribimos un documento, para desarrollar las ideas, solemos estructurar la información con la intensión de facilitar que las personas que lo leen puedan comprender con mayor facilidad el mensaje que deseamos compartir.

La contrucción más empleada para expresar una idea es la oración. Sin embargo, las ideas suelen ser complejas y una oración no es suficiente para que en ella se pueda desarrollar completamente. Por ello debemos emplear varias oraciones para expresar o sustentar la misma. Estas oraciones las agrupamos en párrafos.

Documentos complejos suelen estar compuestos por varios párrafos, los cuales pue-

[5]http://www.clef-initiative.eu/

den ser agrupados en estructuras más generales, las cuales a su vez pueden ser agrupadas en otras aún más generales, obteniéndose de esta forma una estructura jerárquica en la información contenida en un documento.

Diferentes documentos suelen tener una estructura diferente, de acuerdo al tipo de documento que se trate. Por ejemplo, los artículos científicos se suelen estructurar en secciones y subsecciones, los libros en capítulos, etc. Cualquiera sea el tipo de documento, casi todos se estructuran en párrafos y oraciones en los cuales se recogen las ideas más básicas.

De esta forma, un documento está compuesto por una idea general (o hilo conductor), la cual a medida que se desciende a estructuras más específicas se va particularizando, hasta llegar a una oración [Hogan, 2012].

En el presente trabajo, a esta organización que le damos a los documentos cuando los escribimos le denominaremos la *estructura del documento*.

2.3 Representación de documentos

Para la representación de los documentos es muy común que sea utilizado el modelo vectorial [Salton, 1989]. El mismo está basado en la idea de representar cada documento de la colección por un vector de características. Cada documento de la colección d_j es transformado en un vector:

$$d_j =< w_1^j, w_2^j, ..., w_{|T|}^j > \tag{2.1}$$

donde T representa el conjunto total de características que aparecen al menos una vez en la colección de documentos. T suele ser denominado como *vocabulario*. w_k^j representa la importancia de la característica t_k en el documento d_j.

La representación más simple para los documentos dentro del modelo vectorial es

conocido como Bolsa de Palabras. Esta representación consiste en transformar los documentos en vectores en los cuales cada característica representa un término diferente. Si un término no aparece en un documento su peso es 0. Normalmente los términos muy comunes y los poco frecuentes se eliminan y las palabras se reducen a su forma canónica. Esta representación de los textos excluye cualquier estructura gramatical o sintática de los textos [Amine et al., 2009].

Los términos que forman parte de un documento no son seleccionados al azar. Es decir, cuando tratamos un tema empleamos la terminología asociada al mismo, la cual consiste de términos particulares que se encuentran interrelacionados, y son en consecuencia interdependientes. Esta terminología viene expresada por diferentes conceptos como pueden ser partes o componentes que integran el tema general, profesiones asociadas, personajes, organizaciones, sinónimos, entre muchas otras. Las relaciones que se establecen entre los términos han sido estudiadas desde el punto de vista matemático en varios trabajos, mostrando la existencia de una dependencia estadística en la ocurrencia de los términos asociados a una misma terminologías [Baayen, 1996] [Katz, 1996].

Aún cuando el modelo vectorial ha sido ampliamente empleado en diversas tareas de Minería de Textos, es conocido que este modelo presenta el inconveniente de suponer que existe una independencia estadística entre las diferentes palabras que componen un documento, lo cual usualmente no se cumple [Harish et al., 2010]. Además, este modelo es incapaz de manejar los problemas de polisemia y sinonimia presentes en el lenguaje.

En la literatura hay reportados otros métodos para la representación de los documentos, como es el caso del LSI (*Latent Semantic Indexing*) [Deerwester et al., 1990], LDA (*Latent Dirichlet Allocation*) [Blei et al., 2003], PLSI (*Probabilistic Latent Semantic Indexing*) [Hofmann, 1999], Word2Vec [Mikolov et al., 2013], entre otros. Todos ellos tienen en común el hecho de no suponer que los términos que forman un documento son independientes entre sí. Estos algoritmos son costosos computacionalmente,

lo cual es una fuerte limitante para ser aplicados a problemas en los cuales se requiere de una actualización constante del espacio representación. Una posible alternativa a las representaciones antes mencionadas es el Indexado Aleatorio, la cual es una representación que no supone independencia entre los términos y es computacionalmente menos costosa que otras.

2.3.1 Indexado Aleatorio

El Indexado Aleatorio [Sahlgren, 2005] es un método para la representación de los documentos en el cual no se supone que los términos que forman un documento son independientes entre sí, y puede ser construido de forma incremental mediante los siguientes dos pasos:

1. A cada contexto (por ejemplo: un documento u oración) en la colección le es asignada una representación única generada de forma aleatoria, denominada Vector Índice. Estos vectores índices son dispersos, de una elevada dimensionalidad (k), y están compuestos por un pequeño número de elementos con valor 1 y -1, distribuidos aleatoriamente, con el resto de los elementos del vector definidos a 0.

2. Posteriormente, se genera un Vector de Contexto para cada uno de los términos. Para ello se recorre el texto y cada vez que aparece un término en el contexto se adiciona su vector índice al vector de contexto asociado al mismo.

La representación de un contexto, por ejemplo un documento u oración, se obtiene combinando los vectores de contextos de los términos que él contenga.

En el Indexado Aleatorio se han empleado básicamente dos contextos fundamentales: documentos y términos. Cuando el contexto es un documento, se le asigna un vector índice a cada documento y los vectores de contexto se obtienen combinando los vectores índices de cada uno de los documentos en los cuales está presente el término.

Por ejemplo, supongamos que tenemos los documentos d_1 y d_2,

$$d_1 = [t_1, t_2, t_3, t_4, t_5, t_6, t_7] \text{ y } d_2 = [t_8, t_9, t_4, t_5, t_7]$$

Al documento d_1 se le asignaría un vector índice I_{d_1}, y al documento d_2 el vector índice I_{d_2}. Luego, el vector de contexto (C_{t_4}) del término t_4 se obtendría como $C_{t_4} = I_{d_1} + I_{d_2}$.

La segunda aproximación más difundida es considerar los términos como contexto. En este caso se asigna un vector índice diferente a cada término, y su vector de contexto es construído combinando los vectores de índices de los términos con los que coocurre en una ventana a su alrededor.

A diferencia del ejemplo anterior, en el que los vectores índices son asignados a los documentos, en esta aproximación son asignados a cada uno de los términos. De esta forma, considerando una ventana de tamaño 1 alrededor del término t_4, el vector de contexto C_{t_4} sería calculado como: $C_{t_4} = I_{t_3} + 2 * I_{t_5} + I_{t_9}$, donde I_{t_3}, I_{t_5} y I_{t_9} son los vectores índices asociados a los términos t_3, t_5 y t_9 respectivamente.

Esta representación ha sido empleada en diversas tareas de la Minería de Textos. Entre ellas encontramos el Reconocimiento de Nombres de Entidades [Jonnalagadda et al., 2010], la desambiguación del sentido de las palabras [Moen et al., 2013], expansión de consultas en la Recuperación de Información [Sahlgren et al., 2002; Sahlgren and Karlgren, 2001], construcción de resúmenes [Hassel and Sjöbergh, 2006], la recuperación de imágenes e información [Rekabsaz et al., 2017][Kejriwal and Szekely, 2017] y la recomendación de citas [Tan et al., 2018]. El Indexado Aleatorio posibilita lidiar con la sinonimia en el lenguaje, no obstante esta técnica no permite atender el problema de la polisemia.

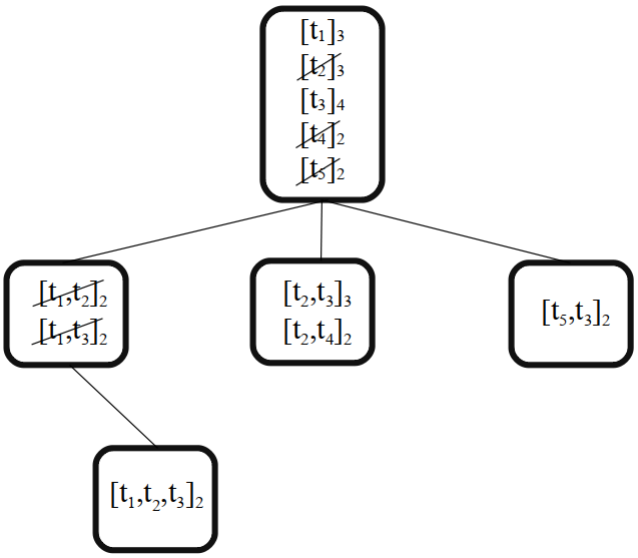
2.4 Representación basada en patrones

Muchas técnicas han sido presentadas con el fin de atacar diferentes tareas de la Minería de Datos en general. Entre ellas encontramos la minería de patrones frecuentes y las secuencias frecuentes. Aún cuando éstas han sido empleadas con éxito en varias tareas, su uso en la Minería de Textos, y en particular en el Filtrado de Información, no ha sido explorado con igual intensidad. Un patrón frecuente, también conocido como conjunto frecuente, es un conjunto de elementos cuya frecuencia de aparición en una colección o base de datos es superior a un valor umbral fijado [Lee et al., 2017]. La frecuencia de aparición de un patrón es conocida como soporte. Una secuencia frecuente es un patrón frecuente en el cual se mantiene el orden de los elementos, y éstos aparecen contiguos.

Uno de los modelos más recientes basados en patrones frecuentes en el Filtrado de Información es conocido por las siglas *PTM* (*Pattern Taxonomy Model*)[Wu et al., 2004; Zhong et al., 2012; Gao et al., 2015; Wai and Aung, 2017; Changala and Rao, 2018]. Este modelo divide el documento en párrafos. El método extrae secuencias cuyos valores de soporte superan el umbral prefijado. Por ejemplo, suponiendo que contamos con un documento compuesto por 4 párrafos de la siguiente forma:

Párrafo	Secuencia
1	$[t_1, t_2, t_3, t_4]$
2	$[t_2, t_4, t_5, t_3]$
3	$[t_3, t_6, t_1]$
4	$[t_5, t_1, t_2, t_7, t_3]$

considerando un valor de soporte de 2, el algoritmo extrairías las secuencias:



Las secuencias con un soporte inferior a 2 fueron omitidas. Las que se muestran tachadas son secuencias descartadas por existir otra de mayor longitud con igual valor de soporte.

Una vez extraídos los patrones, el perfil es representado por un vector centroide, en el cual se concentran los patrones encontrados. La importancia de un patrón en el centroide es calculada como la relación entre los documentos Relevantes en los cuales éste aparece sobre el total de documentos disponibles para la construcción del perfil que lo contienen. La similaridad entre un documento candidato y el centroide es calculada como la suma de los pesos en el centroide de cada uno de los patrones que contiene el documento.

2.5 Clasificadores

Uno de los elementos fundamentales en un sistema de Filtrado de Información es el clasificador, el cual es el responsable de analizar cada documento y determinar si debe ser seleccionado o no para ser mostrado al usuario.

En general, un *clasificador* ψ [Sun and Zhou, 2011] es una función que selecciona

una etiqueta de clase (de entre un grupo de etiquetas predefinidas) para una instancia descrita a partir de un conjunto de características (atributos). En nuestro caso, un clasificador puede ser definido formalmente como:

$$\psi : d \longrightarrow \{Relevante, No\ Relevante\} \tag{2.2}$$

d representa a un documento del flujo que debe ser analizado para determinar si es un candidato para ser mostrado o no al usuario.

Algoritmo de Rocchio

Uno de los algoritmos más explorados en la tarea del Filtrado es el algoritmo de Rocchio [Allan, 1996], el cual es bien conocido y surgió en la Recuperación de Información, fundamentalmente al tomar en cuenta la retroalimentación de relevancia. La idea del algoritmo es simple, una vez proporcionada y ejecutada una consulta, el usuario examina los documentos recuperados y debe determinar cuáles de ellos le resultan relevantes y cuáles no. Una vez que se tiene la información del usuario, el sistema genera de forma automática una nueva consulta en la cual los pesos de los términos son recalculados y aplicando diferentes coeficientes a los pesos de la consulta inicial, tomando en cuenta los documentos señalados como Relevantes R y los señalados como No Relevantes NR [Figuerola et al., 2004]. Para esto, el algoritmo emplea la expresión siguiente:

$$Q' = \alpha Q + \beta \frac{\sum\limits_{x_i \in R} x_i}{|R|} - \gamma \frac{\sum\limits_{x_j \in NR} x_j}{|NR|} \tag{2.3}$$

En la misma, Q es la consulta inicial y Q' la consulta modificada. Los coeficientes α, β y γ son parámetros del algoritmo y controlan cuál es la influencia de cada una de las componentes en la nueva consulta.

El algoritmo de Rocchio condensa en un único vector la consulta Q empleada para

representar el perfil, así como los documentos Relevantes R y los No Relevantes NR. Este algoritmo ha sido empleado en sistemas de Filtrado de Información en varias ocasiones. En el mismo, un documento del flujo de documentos es seleccionado para ser presentado al usuario si la semejanza entre la consulta modificada y el vector del documento supera un umbral μ especificado como parámetro. En resumen, un documento es seleccionado para ser presentado al usuario si:

$$Q^T \times d \geqslant \mu \tag{2.4}$$

2.6 Colecciones Experimentales

Para evaluar nuestra propuesta empleamos dos colecciones de documentos que han sido utilizadas en investigaciones previas: Reuters 21578 y RCv1.

Colección Reuters 21578

La Reuters-21578[6] es una colección de documentos recolectada y etiquetada por *Carnegie Group, Inc.* y *Reuters, Ltd.* Esta colección está compuesta por noticias que aparecieron en el rotativo Reuters durante el año 1987, y se encuentra disponible desde 1996. A pesar de tener varios años, esta colección continúa al día de hoy siendo muy utilizada para investigaciones en varias tareas de Minería de Textos.

Varios subconjuntos han sido creados por los investigadores a partir de la colección original, de ellos los más conocidos son:

- el conjunto de las 10 categorías con el mayor número de muestras positivas.
- el conjunto de las 90 categorías con al menos un documento en el conjunto de entrenamiento y uno en el conjunto de prueba.

[6]http://www.daviddlewis.com/resources/testcollections/reuters21578/

- el conjunto de las 115 categorías con al menos un documento en el conjunto de entrenamiento.

En particular, para nuestros experimentos empleamos el primero de estos conjuntos, es decir el conjunto de las 10 categorías. En los resultados mostrados nos referiremos a este conjunto de documentos como REU10, el cual está compuesto por más de 7900 documentos y contiene más de 32000 términos diferentes.

En la colección, los documentos son procesados de acuerdo a su fecha de publicación, y empleamos para la creación del perfil inicial solamente la descripción de las clases (compuestos solamente por 1 ó 2 palabras).

Colección RCv1

Una de las colecciones de documentos más populares existentes hoy en día es la *Reuters Corpus Volume 1* (RCv1) [Lewis et al., 2004], la cual está compuesta por más de 800 000 artículos noticiosos recopilados entre Agosto de 1996 y Agosto del año siguiente. Esta colección contiene más de 1.1 millones de términos diferentes.

En particular, de esta colección empleamos el subconjunto utilizado en las conferencias TREC11. Para dicha conferencia, 100 tópicos fueron creados, de ellos los primeros 50 fueron creados por los asesores de la NIST, y los otros 50 fueron creados de forma artificial mediante la intersección de tópicos originales de la RCv1. Usualmente los primeros 50 tópicos son los empleados por los investigadores en sus trabajos. Cuando se reportan resultados, nos referiremos a esta colección como TREC.

La RCv1 está compuesta tanto por noticias, como por compendios de titulares de diversas agencias de prensa. Los tópicos creados por los asesores son muy diferentes unos de otros en cuanto a la cantidad de documentos positivos presentes en el flujo de noticias, así como en la distribución de éstos a lo largo del flujo de noticias.

Otro elemento que debe notarse de esta colección es que, en el caso de los documentos

que son compendios de titulares, éstos están etiquetados como pertenecientes a un tópico si al menos uno de los titulares es relevante para el mismo, aunque no se especifica cuál o cuáles son los titulares relevantes. Por ejemplo, consideremos el documento con identificador 84873, recolectado el día 30 de Septiembre de 1996. Este documento es un ejemplo de un compendio de titulares, entre los cuales se encuentran los siguientes:

- The presidents of the Baltic states in a joint statement vowed to make efforts to upgrade their defence capabilities to NATO standards and get entry into the alliance as soon as possible after U.S. Defence Secretary William Perry on Friday said the Balts were not yet ready for NATO.

- President Lennart Meri said at the opening of the monument to the Estonia ferry disaster in Tallinn that he would give his full support to finding out the reasons for the disaster.

- The fire at the Oru peat field in northeastern Estonia may spread from 150 hectares to a 500 hectare area after winds picked up over the weekend although a downpour of rain is expected to help the rescue services.

- The former chairman of the European Parliament Egon Klepsch said at a round-table meeting in Tallinn that no prognosis could be made on European Union expansion.

Este documento se encuentra etiquetado como relevante al tópico R103, el cual está dedicado al *"Hundimiento de Ferrys"*. Al analizar el documento podemos notar que solamente el segundo de los titulares es el que se encuentra relacionado con el tópico de interés.

2.7 Medidas de Evaluación

Para evaluar la efectividad de nuestra propuesta empleamos como medidas de evaluación la $T11SU$ y la tradicional F_1.

La medida $T11SU$ [Soboroff et al., 2012] consiste en una normalización de la utilidad lineal definida en la TREC como $T11U$. La medida $T11U$ asigna dos puntos por cada documento relevante recuperado (TP) y recibe una penalización de un punto por cada documento no relevante recuperado (FP). Esta medida es expresada como:

$$T11U = 2 * TP - FP \tag{2.5}$$

La normalización de la medida es realizada basándose en su máximo teórico, y escalada contra el valor de -0.5.

La medida $T11SU$ es calculada mediante la expresión:

$$T11SU = \frac{max\,(T11NU,\ MinU) - MinU}{1 - MinU} \tag{2.6}$$

Donde $MinU = -0{,}5$ y

$$T11NU = \frac{T11U}{MaxU} \tag{2.7}$$

Siendo

$$MaxU = 2 * (TP + FN) \tag{2.8}$$

$MaxU$ representa la máxima utilidad que puede ser alcanzada por un sistema. En la expresión 2.8, FN representa la cantidad de documentos relevantes para el usuario que el sistema es incapaz de seleccionar.

Nótese que la medida $T11SU$ alcanza el valor de 0.33 cuando el sistema se abstiene de seleccionar documentos para ser mostrados al usuario. Como los organizadores de la

TREC notaron, un sistema que tiene este valor de calidad no es útil para el usuario, pero tampoco le hace malgastar el tiempo con falsos documentos positivos.

Adicionalmente a la medida $T11SU$, empleamos en nuestros experimentos la medida F_1, tradicionalmente empleada en varias tareas de Minería de Textos. La medida F_1 es una media harmónica entre Precisión (P) y Relevancia o Recuerdo (R) y es calculada mediante la expresión:

$$F_1 = \frac{2 * P * R}{P + R} \tag{2.9}$$

Siendo la precisión la razón entre los documentos relevantes recuperados para el usuario con respecto al total de documentos recuperados, y el recuerdo la razón entre los documentos recuperados en relación al total de documentos que deberían haberse recuperados. Esto es:

$$P = \frac{TP}{TP + FN} \tag{2.10}$$

$$R = \frac{TP}{TP + FP} \tag{2.11}$$

Por ejemplo, un sistema que selecciona para el usuario erróneamente el doble de documentos No Relevante por cada documento Relevante seleccionado, suponiendo que todos los documentos Relevantes son presentados al usuario, con la medida T11SU obtendría un valor de 0.33 y con la medida F_1 alcanzaría un valor 0.5.

Capítulo 3

Trabajos Relacionados

Las aproximaciones dadas al problema del Filtrado de Información han sido varias. En un primer grupo encontramos trabajos que emplean algoritmos inicialmente propuestos para la tarea de la Recuperación de Información. Otras propuestas han empleado algoritmos de Categorización de Textos, enfoques evolutivos y modelos basados en patrones. En el presente capítulo se presenta una revisión de las propuestas más representativas en cada uno de los enfoques reportados. En la sección 3.1 se presentan los trabajos que siguen el empleo de algoritmos de la tarea de Recuperación de Información aplicados al Filtrado Adaptativo de Documentos. Seguidamente, en la sección 3.2, se presenta un resumen de las aproximaciones más relevantes entre los sistemas que emplean clasificadores para determinar si un nuevo documento debe ser seleccionado o no para ser mostrado al usuario. En la sección 3.3 se presenta un resumen de los métodos basados en patrones aplicados al filtrado de información. Finalmente, en la sección 3.4 se muestran las conclusiones del capítulo, en la cual se muestran las principales deficiencias de los métodos existentes en el estado el arte y se muestran las diferencias fundamentales con la propuestas del presente trabajo.

3.1 Trabajos basados en métodos de Recuperación de Información

Entre los trabajos que siguen ésta línea suele emplearse el tradicional modelo de Bolsa de Palabras para la representación de los documentos. De forma general, el proceso para determinar si un nuevo documento del flujo debe ser seleccionado o no para ser presentado al usuario se limita a asignar una puntuación al documento en función de

cuán similar es su contenido con respecto a la información almacenada en el perfil. Si esta puntuación supera un valor umbral prefijado, el mismo es seleccionado, en caso contrario el documento es descartado. La representación del perfil suele estar compuesta por un único vector en el cual se condensa el interés del usuario. El algoritmo que con mayor frecuencia es empleado entre los trabajos que siguen esta línea es el algoritmo de Rocchio. Las principales diferencias entre los trabajos que siguen este enfoque radican en la forma en que es establecido el umbral para determinar si un documento debe ser seleccionado y la selección de los términos, y su relevancia, que integran el vector representación del interés del usuario.

Uno de los sistemas que siguen este enfoque es CLARIT, desarrollado por la compañía alemana CLARITECH [Zhai et al., 1998]. Este sistema se basa en el algoritmo de Rocchio, empleando el modelo de Bolsa de Palabras, y la importancia de los términos es medida de acuerdo al esquema de pesado TF-IDF. Un elemento característico de este sistema es que el vector empleado para representar el perfil del usuario está compuesto por solo k términos. El valor de k crece a medida que aumenta el número de documentos Relevantes mediante la expresión $k = 10 + 10 * log\,(R + 1)$, donde R representa la cantidad de muestras relevantes disponibles. El umbral empleado para seleccionar los documentos es fijado y actualizado de forma automática.

Otro sistema que emplea el algoritmo de Rocchio en la tarea de filtrado es el YFilter, desarrollado en *Carnegie Mellon University* [Zhang and Callan, 2000, 2001; Collins-Thompson et al., 2002]. A diferencia de la propuesta anterior, este sistema emplea una variante del esquema de pesado de términos *Okapi tf*, en la cual las palabras muy frecuentes y las raras son penalizadas. Para determinar la puntuación que se otorga a cada nuevo documento, se emplea la expresión BM25 tf.idf [Robertson et al., 1996]. El sistema actualiza de forma dinámica el umbral usado para determinar si un documento es entregado o no, para ello cada vez que se recibe retroalimentación negativa este

valor es incrementado, si no éste es decrementado gradualmente tomando en cuenta el desempeño del sistema y la cantidad de documentos seleccionados para mostrar al usuario.

En la conferencia TREC-11, el sistema que obtuvo los mejores resultados fue presentado por la Academia de Ciencias de China [Xu et al., 2002]. Este sistema emplea el modelo de Bolsa de Palabras, utilizando como términos las raices (stems) de las palabras no vacías (non stop-words). Los documentos son pesados considerando el esquema de pesado TF.IDF y la puntuación otorgada a los documentos se realiza empleando la medida del coseno. El sistema introduce el concepto de documentos pseudo-negativos, que son aquellos que obtienen un valor de semejanza inferior a un umbral dado k, para los cuales no se cuenta con retroalimentación por parte del usuario. El algoritmo de Rocchio en esta propuesta es modificado para tomar en consideración los documentos pseudo-negativos.

Trabajos recientes continúan explorando el uso del algoritmo de Rocchio en la tarea del filtrado. [Berardi et al., 2015] exploran la utilización del algoritmo en el procesamiento de Tweets, enriqueciendo el contenido de estos mediante una estrategia de detectar las entidades contenidas en el texto del tweet con las páginas almacenadas en Wikipedia, previo proceso de desambiaguación de la entidad seleccionada.

Zhang en 2004 presenta un sistema en el cual se combina la regresión logística con el algoritmo de Rocchio. En esta propuesta primeramente es empleado el algoritmo de Rocchio, y la consulta resultante es empleada como vector inicial para la regresión logística. La idea de combinar un modelo de regresión logística con un algoritmo diseñado para la recuperación de información ha sido recientemente retomada en [Han et al., 2016]. En este caso se emplea un modelo de lenguaje construido a partir de un repositorio auxiliar, y el sistema fue aplicado al procesamiento de tweets. El uso de modelos de lenguaje fue también explorado por [Zagheli et al., 2017a], quienes propo-

nen el uso de dos modelos de lenguajes, uno para los documentos relevantes y otro para los no relevantes y emplean la diferencias entre los valores asignados por la divergencia de Kullback-Leibler para asignar la puntuación para cada nuevo documento del flujo. Ideas similares han sido exploradas en los siguientes trabajos: [Fan et al., 2016; Rahmatizadeh Zagheli et al., 2017; Zagheli et al., 2017b; Zamani and Shakery, 2018]

Este tipo de algoritmos presentan el inconveniente de requerir de un umbral μ, para el cual no siempre se logra obtener un valor adecuado cuando se tienen pocas muestras y que a la vez sea útil en entornos con muchas más muestras. Además, no toman en consideración la estructura de los documentos ni las diferentes relaciones que se establecen entre los términos cuando se toman en cuenta diferentes niveles de contextos. No suelen presentar una estrategia para combatir el problema de la escasez de información, ni toman en consideración el hecho de que los subtópicos que forman la necesidad de información se pueden encontrar distribuidos de forma no homogénea en el flujo y entre los documentos disponibles para la construcción del perfil.

3.2 Trabajos que emplean métodos de Categorización de Textos

El uso de clasificadores empleados en la tarea de Categorización de Textos es otra gran vertiente de propuestas para seleccionar los documentos que deben ser seleccionados para ser mostrados a los usuarios. En este enfoque se trata el problema del filtrado como una tarea de Categorización de Textos, en la cual se emplea un clasificador binario el cual para cada nuevo documento el clasificador debe asignar una etiqueta de entre dos clases posibles, Relevante y No Relevante, en caso de que el clasificador asigne al documento la etiqueta de Relevante, éste es mostrado al usuario, en caso contrario el documento es descartado y el usuario no tiene acceso a él.

El clasificador Winnow [Littlestone, 1988] fue la opción del sistema de la Universidad de Fudan [Wu et al., 2002] para la competición TREC 11. En este sistema se emplea el modelo de Bolsa de Palabras, y los términos son pesados empleando la expresión:

$$w\left(t_i, d\right) = 1 + log\left(TF\left(t_i, d\right) * \frac{avdl}{dl}\right) \tag{3.1}$$

En la expresión, $w\left(t_i, d\right)$ es el valor del peso del término t_i en el documento d, y $TF\left(t_i, d\right)$ indica el valor de la frecuencia de t_i en d. El valor dl representa el número promedio de términos diferentes en un documento y $avdl$ es la longitud promedio de los documentos. El sistema no considera todos los términos para la construcción del perfil de usuario, sino solamente aquellos que considera importantes. La importancia de cada término se calcula empleando la información mutua, según la expresión:

$$logIM\left(t_i, P\right) = log\left(\frac{Prob\left(t_i|P\right)}{Prob\left(t_i\right)}\right) \tag{3.2}$$

El sistema elimina aquellos términos para los cuales el valor de $logIM\left(t_i, P\right)$ es inferior a 0.3.

Otros algoritmos de categorización de textos que han sido empleados en sistemas de filtrado de documentos son los clasificadores basados en vecindad. Tal es el caso de Ault y Yang 2000; 2001 que exploraron el uso del clasificador k-NN. Para representar los documentos emplearon el modelo de Bolsa de Palabras y los términos fueron pesados empleando una variante del esquema de pesado Okapi, mediante el uso de la expresión:

$$w\left(t, d\right) = \frac{TF(t, d)}{0{,}5 + 1{,}5\frac{len(d)}{avdl} + TF\left(t, d\right)} * \frac{log\left(0{,}5 + N - df\left(t\right)\right)}{0{,}5 + n\left(t\right)} \tag{3.3}$$

en la cual, N representa la cantidad de documentos en el conjunto de entrenamiento, $n(t)$ es la número de documentos que contienen al término t, $len(d)$ el total de términos del documento luego de representar a cada término por su raíz (stem) y eliminar las

palabras vacías (stopwords). Un elemento característico de esta propuesta es que no emplea una clasificación binaria, sino que todos los documentos de los perfiles son agrupados en un solo conjunto de entrenamiento. El sistema emplea un sistema de puntuación para cada documento en cada perfil P mediante la expresión:

$$p\left(d, P\right) = \sum_{x \in V \cap P} sem\left(d, x\right) \tag{3.4}$$

con V representando el conjunto de entrenamiento y $sem\left(d, x\right)$ la semejanza del coseno entre x y d. El documento es seleccionado como relevante para el perfil P si $p\left(d, P\right)$ supera un determinado umbral.

La idea de emplear el clasificador k-NN fue retomada en [Qamar et al., 2010]. En este trabajo los autores siguieron un esquema similar al propuesto por Ault y Yang. En [Cossu et al., 2015], nuevamente fue explorado el uso del k-NN en la tarea del Filtrado de Información.

Otro clasificador cuyo uso que ha sido explorado en la tarea del filtrado son las Máquinas de Vectores de Soporte (*SVM*, por sus siglas en inglés) [Cancedda et al., 2003; Srikanth et al., 2002; McNamee et al., 2002; Montejo-Ráez et al., 2010]

Otras aproximaciones han explorado el uso de otros clasificadores y enfoques, como es el caso del uso de los algoritmos evolutivos [Nanas and de Roeck, 2010; Horváth and de Carvalho, 2017; Mohd Azmi et al., 2017]. Estudios alternativos han explorado el uso de facetas [Zhang and Zhang, 2010; Zhang et al., 2014], o modelos Bayesianos [Zhang and Zhang, 2014].

3.3 Representación basada en patrones

Muchas técnicas han sido presentadas con el fin de atacar diferentes tareas de la Minería de Datos en general. Entre ellas encontramos la minería de patrones frecuentes

y las secuencias frecuentes. Aún cuando estas han sido empleadas con éxito en varias tareas, su uso en la Minería de Textos, y en particular en el Filtrado de Información, no han sido explorados con igual intensidad.

Uno de los modelos más recientes basados en patrones frecuentes en el Filtrado de Información es conocido por las siglas *PTM* (*Pattern Taxonomy Model*) [Wu et al., 2004]. Este modelo divide el documento en párrafos. El método extrae secuencias cuyos valores de soporte superen un umbral prefijado. Las secuencias que son consideradas no útiles, por aparecer en el documento como parte de otra de mayor longitud, son eliminadas. Una vez extraídas secuencias frecuentes en cada uno de los documentos del conjunto de entrenamiento, el perfil del usuario es construído condensando las secuencias en un único vector. La importancia de cada secuencia P en el centroide empleado para representar el perfil es calculado mediante la razón siguiente:

$$p\left(P\right) = \frac{|\{d_a | d_a \in R,\, P\, en\, d_a\}|}{|\{d_b | d_b \in R \cup NR,\, P\, en\, d_b\}|} \tag{3.5}$$

donde d_a y d_b denotan documentos. R y NR los conjuntos Relevantes y No relevantes respectivamente disponibles para la construcción del perfil. Una vez construído el perfil del usuario, para determinar si un nuevo documento debe seleccionarse para ser mostrado al usuario se adicionan los pesos de los patrones P que aparecen en el nuevo documento y se compara con un valor de umbral dado.

En [Algarni et al., 2008], se presentó el algoritmo *APTM* (*Adaptive Pattern Taxonomy Mining*), el cual consiste en adaptaciones al método PTM para que se ajuste a entornos adaptativos. Los cambios se centraron fundamentalmente en el uso de las nuevas muestras de entrenamiento para ajustar los umbrales y la función empleada para medir la importancia de los patrones extraídos.

Con la finalidad de tomar en consideración la existencia de muestras negativas a

la hora de extraer los patrones, Li et. al. 2009; 2011 propusieron el método N-PTM. La estrategia de los autores en estos trabajos consistió en modificar los pesos de los términos y patrones tomando en consideración la ocurrencia de los mismos entre los documentos positivos y los negativos.

Los patrones extraídos pueden ser empleados directamente como características para representar los documentos; o, por el contrario, pueden ser empleados para reducir el espacio de representación, o cambiar el pesado de los términos en los documentos [Li et al., 2015].

Los métodos pueden ser combinados con métodos tradicionalmente aplicados al Filtrado de Información. Tal es el caso de la propuesta presentada en [Li et al., 2012], donde se combina el PTM con otros métodos, con el fin de descartar de manera rápida documentos no relevantes, y en un segundo momento emplear un método basado en patrones solamente en aquellos documentos candidatos.

Propuestas recientes han combinado el uso del PTM con LDA. En [Gao et al., 2015; Wai and Aung, 2017], los patrones son generados a partir de las palabras obtenidas en la representación de los tópicos obtenidos mediante el LDA.

3.4 Discusión

Como hemos visto, varias han sido las aproximaciones reportadas en la literatura con respecto al Filtrado de Información, entre ellas encontramos trabajos que emplean algoritmos de la Recuperación de Información (RI), otros que utilizan métodos de Categorización de Textos (CT), etc. Además en este capítulo se presentaron antecedentes de métodos que emplean patrones en la tarea del Filtrado de Información.

Métodos que emplean algoritmos de RI

Los métodos que emplean técnicas de RI condensan en un único vector la representación del perfil del usuario. Esto trae como consecuencia que los subtópicos menos representados en el perfil puedan verse afectados por la mayor presencia de los subtópicos mejor representados.

Este tipo de algoritmos presentan el inconveniente de requerir de un umbral μ, para determinar si un documentos debe ser seleccionado o no para ser mostrado al usuario, para el cual no siempre se logra obtener un valor adecuado cuando se tienen pocas muestras y que a la vez sea útil en entornos con muchas más muestras.

Por otra parte, estos algoritmos funcionan mejor para satisfacer necesidades de información generales, en las cuales los documentos que satisfacen a los usuarios desarrollan el tema a lo largo de su contenido. Pero cuando la necesidad de información es más específica, su desempeño se ve comprometido. Por ejemplo, si un usuario está interesado en lo que dice un candidato presidencial en una campaña electoral con respecto a la educación pública, éste tipo de algoritmos no son capaces de descartar documentos en los cuales el candidato se refiera a la seguridad pública y la educación privada sin hacer mención a la educación pública.

Métodos basados en CT

Los métodos de CT se ven seriamente afectados por la poca información disponible para la construcción inicial del perfil. La poca disponibilidad de documentos conlleva a que a los clasificadores les resulte muy complicado realizar una correcta modelación de la frontera de decisión entre los documentos que satisfacen la necesidad de información de aquellos que no la satisfacen.

Los trabajos presentados no suelen tomar en consideración el fenómeno del posible

desbalance entre los diferentes subtópicos de interés, y los documentos que abordan cada uno de estos subtópicos.

Al igual que los métodos anteriores, las propuestas que emplean algoritmos de CT son incapaces de filtrar adecuadamente los documentos que se ajustan a necesidades de información más específicas como el del ejemplo anterior.

Métodos basados en Patrones

Los métodos que exploran el uso de técnicas basadas en patrones en la tarea del filtrado de información suelen emplear umbrales para determinar los patrones que deben ser extraídos, o para determinar si un nuevo documento debe ser seleccionado o no para ser mostrado al usuario. El problema con el uso de umbrales en los entornos como el Filtrado Adaptativo es que la disponibilidad de información no es la misma con el paso del tiempo.

Los umbrales que se emplean cuando la información disponible es muy reducida pueden no ser adecuados para cuando aumenta la cantidad de información disponible. Incluso, el problema es mucho más complejo si se toma en consideración la no homogeneidad de los subtópicos de interés entre el conjunto de entrenamiento, en los cuales podemos tener subtópicos muy bien representados y otros con muy pocas muestras de entrenamiento disponible, sin que por ello pueda suponerse de antemano que son menos importante para el usuario.

Ninguno de los modelos toma en consideración el hecho de que los términos guardan entre sí diferentes relaciones en función del nivel de granularidad del contexto en el cual ellos se relacionen.

Aspectos comunes

Para representar los documentos se suele emplear el modelo de Bolsa de Palabras. Sin embargo, como ya fue mencionado en el capítulo anterior, este modelo solamente toma en consideración la ocurrencia o no de los términos en los documentos, sin tomar en consideración cómo se relacionan estos en los documentos. Algunas de las propuestas reportadas en la literatura emplean Modelos de Lenguaje u otras técnicas como la LDA, construidos o no a partir de fuentes de datos externas, los cuales permiten aliviar en cierta medida la falta de información para el inicio del funcionamiento del sistema. En un entorno dinámico, donde nuevos documentos arriban con el paso del tiempo, es común que surjan nuevos términos que resulten de interés para los usuarios (por ejemplo: nuevos nombres de empresas, organizaciones, medicamentos o términos acuñados por la jerga popular o del registro popular) los cuales no son tomados en consideración por no encontrarse en el modelo. La actualización de nuevos términos a estos modelos requiere de que los mismos sean nuevamente construido considerando documentos que aborden los términos nuevos. Este proceso resulta costoso, lo cual puede ser un impedimento importante para la aplicación práctica de estos modelos en soluciones reales.

Las propuestas presentadas no toman en consideración la no homogeneidad en la distribución de los documentos entre los diferentes subtópicos en los cuales puede dividirse el interés de información del usuario. Debido a ello, los subtópicos pocos representados se ven en desventaja en la construcción del perfil del usuario, lo cual conlleva a que los documentos que abordan estos subtópicos en el flujo de documentos puedan no seleccionarse.

Por último, ningunas de las propuestas existentes toma en consideración el hecho de que los términos en un documento pueden guardar una relación diferente de acuerdo al contexto que se tome en consideración.

Nuestra propuesta

En nuestra propuesta para la representación de los documentos se conserva la estructura dada por los usuarios a los documentos. Esta representación permite tomar en consideración los diferentes niveles de granularidad de los contextos en los cuales los términos pueden relacionarse. En el presente trabajo se representan dos modelos en los cuales los diferentes niveles de granularidad en los contextos son tomados en cuenta: las relaciones de términos multinivel y el Indexado Aleatorio Multinivel.

El Indexado Aleatorio, a diferencia de otros modelos empleados en la tarea del filtrado, no presupone que los términos en los documentos son independientes entre sí, y además al ser incremental, permite adicionar nuevos términos al modelo sin que sea preciso el procesamiento de todos los documentos nuevamente, lo cual representa una gran ventaja para ser empleado en el Filtrado Adaptativo.

La necesidad de información de un usuario puede estar compuesta por varios subtópicos más específicos, y éstos pueden no encontrarse igualmente representados en los documentos disponibles para la construcción del perfil del usuario. Por ello podemos tener subtópicos de los cuales contamos con varios documentos de muestras que los abordan y otros subtópicos que son pobremente tratados en los documentos disponibles, trayendo como consecuencia la existencia de un desbalance en la cantidad de información disponible de cada uno de ellos. Este problema del desbalance entre subtópicos que forman la necesidad de información del usuario, a diferencia de propuestas anteriores, en nuestra propuesta es tomado en consideración durante el proceso de obtención de las relaciones de términos multinivel.

Para atacar el problema del Inicio en Frío en nuestra propuesta se favorecen las relaciones más simples por encima de aquellas más complejas y específicas, además se emplean algunos recursos externos, en particular la Wikipedia, la base de nombres geográficos GeoNames y la colección de nombres de entidades JRC.

Capítulo 4

Relaciones Multinivel para el Filtrado Adaptativo

En el presente capítulo proponemos un método para la extracción de relaciones considerando múltiples niveles de granularidad, así como su aplicación en el problema del Filtrado Adaptativo. El contenido en el presente capítulo es estructurado en la siguiente forma: primeramente se presenta un análisis de los métodos basados en conjuntos frecuentes aplicados a la tarea del Filtrado. Seguidamente, se introducen las relaciones de términos multinivel y un procedimiento para su extracción. A continuación, un análisis del problema del Inicio en Frío y algunas estrategias para aminorarlo. Luego, se muestran el marco experimental, los resultados obtenidos y el análisis de los mismos. Finalmente, se muestran las conclusiones del capítulo.

4.1 Análisis de los métodos existentes

Como se detalló en la sección 3.3, han sido varios los trabajos basados en Patrones Frecuentes aplicados a la tarea de Filtrado. Sin embargo, la mayoría de los trabajos existentes no han sido debidamente adaptados a la tarea del Filtrado Adaptativo. Pocos de estos trabajos emplean la retroalimentación del usuario cuando éste indica que uno de los documentos seleccionados no es de su interés (retroalimentación negativa) para actualizar la información del perfil, lo cual los pone en desventaja para su aplicación efectiva en el filtrado.

Los métodos propuestos basan su funcionamiento en la suposición de que se contará con la existencia de un número significativo de muestras que permitan extraer patrones

representativos, de acuerdo a valores de soporte fijados. Sin embargo, esta suposición pocas veces es cierta, dado que en los inicios de especificación de una nueva necesidad de información del usuario es probable que no se cuente con una cantidad suficiente de documentos, e incluso, es común que se deba comenzar el proceso de filtrado con tan solo una consulta clásica, como las empleadas en los sistema de Recuperación de Información, es decir con unos pocos términos iniciales.

Otro elemento que no es tomado en consideración en los métodos previamente propuestos está relacionado con el hecho de que los sistemas probablemente no dispondrán de muestras de todos los subtópicos que forman la necesidad de información del usuario, sino que los mismos aparecerán de a poco, a medida que se va procesando el flujo de información. Al emplear un umbral para medir la significancia de los patrones extraídos se corre el riesgo que aquellos subtópicos que están poco representados en el perfil, ya sea porque es un subtópico "raro", o porque recién está surgiendo en el flujo, pueden quedar sin representación en el perfil.

Por último, ninguno de los métodos existentes toma en consideración la estructura de los documentos. Este punto puede ser de vital importancia dado que no tiene la misma significación que dos términos co-ocurran frecuentemente en los documentos, con independencia del lugar en el que éstos aparezcan, a decir que co-ocurren en la misma oración o frase. Por ejemplo, suponiendo que tenemos un usuario interesado en las propuestas de un candidato político con relación a la educación pública, y estamos procesando noticias que se relacionan con las campañas políticas para un proceso electoral. Supongamos, además, que consideramos los términos: *candidado*, *escuela* y *pública*. Si consideramos solamente la ocurrencia de éstos términos en un documento, podemos reconocer documentos en los que se traten escuelas privadas y cuentas públicas, sin que estos se ajusten a las necesidades expresadas por el usuario. Al considerar varios niveles de contexto, podemos exigir que los términos ocurran en el documento,

pero además que *escuela* y *pública* ocurran en el mismo sintagma nominal, con lo cual se podría tener una modelación más adecuada al interés expresado por el usuario.

En el presente capítulo se presenta una propuesta que toma en consideración todos los elementos anteriormente señalados.

4.2 Relaciones Multinivel de términos

El Filtrado Adaptativo se caracteriza, entre otras cosas, porque al principio del proceso de filtrado se cuenta con muy poca información. Sin embargo, la extracción de información cuando estamos en un escenario con tan poca información es un verdadero reto. La información extraída a partir de tan poca información no es confiable. Por otro lado, la construcción del perfil inicial es un proceso crucial. Una construcción errónea del perfil inicial puede conllevar a que se apunte a tópicos diametralmente opuestos a los deseados y no se logre seleccionar alguna información útil para el usuario.

Los conjuntos frecuentes capturan ciertas relaciones de co-ocurrencia que se establecen entre los términos de un documento. Estas relaciones pueden incluir, entre otras, la sinonimia (dado que los humanos solemos emplear sinónimos en los textos para evitar repeticiones de términos), la meronimia, la hiperonimia (hiponimia), e incluso algunas otras de tipo temático, como pudieran ser autos con ventas, enfermeras con médicos, etc. Estas relaciones en ciertos temas menos específicos pueden ser suficientes para diferenciar aquellos documentos de interés de los que no lo son. Por otro lado, éstos pueden ser insuficientes para representar necesidades de información más específicas.

El problema con los conjuntos frecuentes es que, por un lado, si consideramos conjuntos que involucren varios términos, podemos obtener patrones muy específicos, los cuales probablemente obtengan una alta precisión pero, en dependencia del tópico de interés, tendrán muy poca capacidad de generalización. Por otro lado, emplear conjun-

tos de pocos elementos tienen el efecto contrario, probablemente logremos obtener un alto valor de recuperación (recuerdo) a costa de abrumar al usuario con un alto número de documentos falsos positivos. Ambas situaciones son indeseables y es necesario alcanzar un equilibrio que permita obtener la mayor cantidad de documentos de interés para el usuario, sin sobrecargarlo con documentos que no satisfacen sus necesidades de información. Este elemento es particularmente importante en las primeras etapas del proceso de filtrado debido a que el volumen de información disponible para la construcción del perfil es escaso y las soluciones tradicionales basadas en umbrales sobre los valores de soporte, así como los conjuntos de términos extraídos, no son confiables.

Los seres humanos somos todos diferentes unos de otros; por lo tanto, nuestras necesidades de información son igualmente muy diferentes entre sí. Mientras unos usuarios están interesados en tópicos generales, como es el caso del fútbol, la educación o la política, otros se interesarán en tópicos mucho más específicos, como es el caso de accidentes fatales que involucran autobuses escolares. De esta forma, para usuarios interesados en tópicos generales como es el caso de la educación, el empleo de conjuntos frecuentes que involucren los términos más representativos del tema, como son profesores, maestros, estudiantes, libros, etc. probablemente será suficiente para identificar los documentos que abordan la temática de interés. En el segundo ejemplo, el interés es mucho más específico y el uso de conjuntos frecuentes a nivel de documentos no son suficientes para expresar las necesidades de información del usuario. En el caso de necesidades de información más específicas se requiere de la extracción de relaciones entre los términos que involucren el uso de contextos más específicos. Por ejemplo, la relación de co-ocurrencia entre los términos *autobús* y *urbano* debe ser extraída en la misma oración, o incluso en el mismo sintagma nominal, para que pueda ser empleada para representar la necesidad de información del usuario.

En un entorno de filtrado adaptativo, no es posible conocer de antemano cuán

general o específico es el interés del usuario para decidir cuál es el o los contextos adecuados en los cuales se deben extraer las relaciones entre los términos, para una correcta representación de la necesidad de información del usuario. En las propuestas existentes se suele fijar con anterioridad un nivel de contexto (el más empleado es el de documentos), con lo que se pierden las relaciones que se pueden establecer entre los términos en otros contextos y que pueden ayudar a una mejor definición de la necesidad de información. Por ejemplo, si se selecciona el nivel de documentos para la búsqueda de las relaciones de co-ocurrencia entre los términos, probablemente no será posible procesar de forma correcta los tópicos más específicos, como es el caso del ejemplo dado en el párrafo anterior. Por el otro lado, el seleccionar contextos más específicos tampoco es garantía de una correcta modelación de la necesidad de información del usuario.

Una posible solución a este problema puede ser dejar que sea el propio sistema el que seleccione de forma automática los niveles de contexto que deben considerarse para la correcta modelación del interés del usuario a través de la información recibida por parte de éste mediante del mecanismo de la retroalimentación. Presentamos a continuación una extensión a los conjuntos frecuentes en la cual se extraen las relaciones de co-ocurrencia entre los términos tomando en consideración la estructuración de los documentos en contextos de diferentes niveles de granularidad, como pueden ser los párrafos o las oraciones.

Las relaciones de co-ocurrencia extraídas a nivel de documentos son menos específicas. A medida que el contexto empleado para la extracción de las relaciones se vuelve más circunscrito, por ejemplo de documento a párrafo, las relaciones extraídas son cada vez más específicas; y por ende, con ellas se podrán recuperar menos documentos, pero con una mayor certeza en la decisión tomada por el clasificador.

Una relación multinivel de términos es un conjunto de elementos x_i que ocurren juntos en un nivel de contexto dado. Los elementos x_i pueden ser bien términos u otras

relaciones extraídas con un nivel de granularidad más específico.

Para la representación de las relaciones emplearemos la siguiente notación:

$$\Phi_C\left(x_1, x_2, \ldots, x_k\right)$$

Donde C es el contexto utilizado para relacionar los elementos x_i. Por ejemplo, C puede tomar valores como D(documentos), P(párrafos), O(oraciones) o N(sintagmas nominales).

Algunos ejemplos de relaciones son:

- $\Phi_S\left(auto, vendedor\right)$
- $\Phi_D\left(marketing, publicidad\right)$
- $\Phi_S\left(cosecha, \Phi_N\left(abono, orgánico\right)\right)$

Retornando al ejemplo del tópico específico relacionado con los accidentes en los cuales están involucrados autobuses escolares, de éste pudieran ser extraídas relaciones como las siguientes:

$\Phi_D\left(\Phi_N\left(autobús, escolar\right), accidente\right)$ - $\Phi_D\left(choque, muerto, \Phi_N\left(autobús, escolar\right)\right)$

En cierta forma, las relaciones extraídas pueden ser empleadas no solo para clasificar los documentos en Relevantes para el usuario o no; sino que, además, se pueden utilizar para explicar al usuario el por qué un determinado documento es seleccionado.

A continuación se presenta el proceso de extracción de las relaciones multinivel para representar el perfil del usuario. En la sección 4.3 se explica cómo se realiza el proceso de filtrado de los nuevos documentos una vez extraídas las relaciones que se emplean para representar el interés de información del usuario.

4.3 Proceso de Filtrado

Durante el proceso del filtrado de los documentos podemos identificar dos etapas fundamentales: la construcción inicial del perfil del usuario y el procesamiento del flujo de documentos. En nuestro algoritmo, el perfil del usuario está compuesto por un conjunto de relaciones, al que denominaremos STR. Estas relaciones son interpretadas como reglas en la forma:

$$\Phi_{\Omega_i}(r_1, r_2, \ldots, r_k) \Longrightarrow Positivo \tag{4.1}$$

Luego, para cada documento d_i del flujo, d_i es seleccionado como Relevante para el usuario si y solo si:

$$\exists R \in STR,\ d_i\ satisface\ R \tag{4.2}$$

Con la finalidad de realizar el proceso de emparejamiento entre una relación R y los documentos d_i del flujo, mantenemos la estructura jerárquica de los documentos. En la Figura 4.1 se presenta un ejemplo de esta estructura empleada para representar los documentos, considerando 4 niveles Documento(D), Párrafo(P), Oración(S) y Sintagma Nominal(N).

El proceso de emparejamiento entre un documento y una relación Φ se reduce a buscar un nodo en el árbol en el cual se satisfagan todos los componentes de Φ. Este proceso se realiza mediante un recorrido en el árbol primero en profundidad para ir buscando coincidencia parciales entre los elementos de Φ. El proceso termina cuando se encuentra un nodo a partir del cual existe un camino desde él hasta cada uno de los elementos que componen Φ.

Para ganar en eficiencia en el proceso de emparejamiento, se guarda una relación

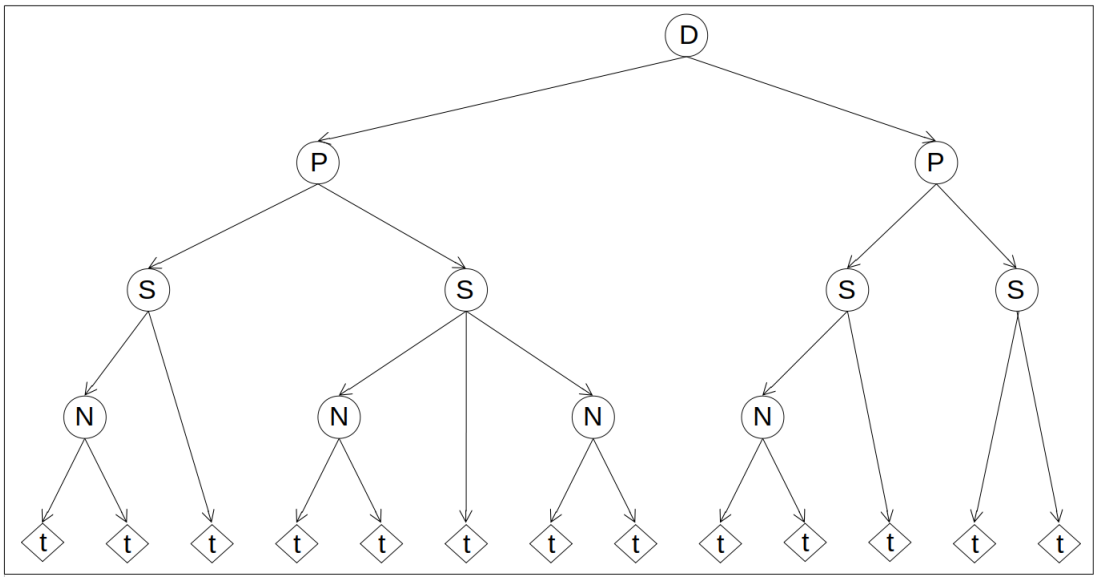


Figura 4.1: Ejemplo de la Estructura empleada en la Representación de los Documentos.

en cada nodo del árbol de todos los términos que se encuentran en los nodos hojas pertenecientes a su sub-árbol correspondiente.

4.4 Método para extraer las relaciones

Para el usuario, expresar de forma concisa, precisa y que abarque completamente todas las vertientes de un tópico por medio de una consulta puede ser extremadamente complicado. Con el método propuesto se intenta disminuir los efectos negativos de extraer relaciones muy específicas en las primeras etapas del proceso de filtrado, y que terminen en un sobreajuste del clasificador. Para ello, el proceso de extracción de las relaciones da prioridad a las relaciones más generales sobre las más específicas. Entendemos que una relación es más general que otra en la medida que se emplea un contexto más general en la relación e involucra una menor cantidad de elementos. El empleo de contextos más generales en las relaciones y que involucren una menor cantidad de términos posibilita que sea menos restrictivo el emparejamiento de las relaciones en

Consulta: Accidentes de autobuses escolares
Relevante: Un autobús escolar se impactó contra un camión esta mañana en la Avenida 41. En el accidente murieron 4 personas, y 15 permanecen hospitalizadas.

$$\downarrow$$

$$\Phi_D\,(escolar)\;\;\Phi_D\,(autobús)\;\;\Phi_D\,(accidente)$$

No Relevante: El gobernador del estado invirtió $500 000 MXN en la compra de 6 autobuses. Estos autobuses serán destinados a las 2 escuelas con mayor matrícula.

$$\downarrow$$

$$\Phi_N\,(autobús, escolar)\;\;\Phi_D\,(accidente)$$

No Relevante: Un autobús urbano colisionó contra el muro del anfiteatro escolar en la 39 cuando una de sus gomas anteriores explotó. No se reportaron lesionados en el accidente.

$$\downarrow$$

$$\Phi_D\,(accidente, \Phi_N\,(autobús, escolar))$$

Figura 4.2: Ejemplo de la evolución del conjunto de relaciones de acuerdo a las retroalimentaciones recibidas por parte del usuario.

los documentos, lo que implica que se entreguen una mayor cantidad de documentos para el usuario.

Durante la extracción de las relaciones entre los términos se comienza con el uso de relaciones que involucran un solo término en el contexto Documentos. A estas relaciones se le van adicionando términos, o se van explorando contextos más específicos, en la medida en que éstas no logran diferenciar los documentos Relevantes de los No Relevantes almacenados en el conjunto de entrenamiento especificando el perfil del usuario. De esta forma se seleccionan relaciones que permiten diferenciar el interés del usuario del resto de los documentos sin afectar abruptamente la capacidad de generalización de las relaciones extraídas.

En la Figura 4.2, presentamos un ejemplo en el cual se muestra como el conjunto de relaciones extraídas puede evolucionar para ajustarse a la necesidad de información

de un usuario expresada por los nuevos documentos adicionados al perfil por medio del proceso de retroalimentación. En el ejemplo se parte de una consulta y una muestra de un fragmento de texto suministrado como muestra Relevante. A partir de esta información se muestran 3 posibles relaciones que pueden ser extraídas para representar el perfil del usuario. Estas relaciones emplean el nivel de contexto más general, es decir nivel documento, y están compuestas por un único término, dado que son las relaciones más simples que permiten representar el perfil. Con estas relaciones se puede seleccionar el fragmento de texto que le sigue dado que menciona a autobús, pero el mismo es No Relevante. Para lograr diferenciar la información Relevantes del No Relevante se combinan las relaciones $\Phi_D\,(escolar)$ y $\Phi_D\,(autobús)$ en una sola relación $\Phi_N\,(autobús, escolar)$. En esta nueva relación no solo se combinan los términos, si no que además es preciso emplear un contexto más específico para que pueda diferenciarse al fragmento No Relevante. El siguiente fragmento No Relevante sería seleccionado porque satisface la segunda relación, por ello se precisa de combinar las relaciones para que la relación diferencie correctamente la información Relevante de los fragmentos No Relevantes.

Para la extracción del conjunto de relaciones se formuló un primer algoritmo, cuyos pasos son resumidos en el Algoritmo 1, el cual recibe como entrada los documentos disponibles para la construcción del perfil, así como el conjunto de términos presentes en los documentos Relevantes o Positivos, los tipos de relaciones o contextos a considerar y el valor umbral de soporte mínimo a considerar para extraer las relaciones. El método devuelve como salida el conjunto de relaciones entre los términos que será empleada para el proceso de emparejamiento por parte del clasificador para determinar si los documentos del flujo deben ser presentados al usuario o no.

En el algoritmo, $DF\,(R, S)$ representa una función que recibe como parámetros una relación R y un conjunto de documentos S, y devuelve como resultado el subconjunto

de documentos de S en los cuales encontramos la relación R. Es decir:

$$DF\left(R,S\right)=\{d_i \mid d_i \in S \wedge R \Mapsto d_i\} \tag{4.3}$$

Además, $Sop\left(R\right)$ es una función que devuelve el valor de soporte de la relación R, Q es una cola de prioridades para mantener las relaciones candidatas, considerando como prioridad el número de documentos positivos en los cuales encontramos las relaciones candidatas. *last* es una función que dada una relación X devuelve el índice, dentro de la lista de términos T, del último término adicionado a X.

Γ representa a una función que, dada una relación R, un nivel de contexto Ω_i y un término t, devuelve nuevas relaciones a explorar. Para determinar las nuevas relaciones a explorar, se emplean los operadores σ y ρ, definidos en la forma siguiente:

$$\sigma\left(R,S\right)=\Phi_{\Omega_i}\left(r_1,r_2,\ldots,r_k,s_1,s_2,\ldots,s_n\right) \tag{4.4}$$

$$\rho\left(R,S\right)=\Phi_{\Omega_R}\left(r_i|r_i \notin S\right) \tag{4.5}$$

Donde $S=\{s_1,s_2,\ldots,s_n\}$ es un conjunto de elementos.

σ es un operador de expansión, el cual permite adicionar nuevos elementos a considerar en la nueva relación. Por el contrario, ρ es de reducción, y permite sustituir algunos elementos por otros.

La función Γ es definida en la forma:

$$\Gamma\left(\Omega_i,R,x\right)=\begin{cases}\{\sigma\left(R,\{x\}\right)\} & \Omega_R=\Omega_i \\ \Psi\left(R,H_1,x,\Omega_i\right)\bigcup\Delta\left(R,H_2,x,\Omega_i\right) & \Omega_R\neq\Omega_i\end{cases} \tag{4.6}$$

Siendo, $H_1=\{r_i|\, r_i\, es\, un\, término\}$ y $H_2=\{r_i\}\setminus H_1$. Ψ y Δ son funciones definidas

Algoritmo 1: Extracción de las relaciones de términos

Entrada: P: Documentos Relevantes (Positivos)
N: Documentos No Relevantes (Negativos)
T: Términos presentes en P
Ω: Arreglo de Tipos de Relaciones
μ: Soporte mínimo para seleccionar las relaciones

Salida: STR: Conjunto de relaciones

```
1  STR = {} ;
2  Q ← ∅ /* Cola de Prioridades Vacía */
3  foreach t in T do
4      if Sop(Φ_{Ω_1}(t)) ≥ μ then
5          if DF(Φ_{Ω_1}(t), N) ≠ ∅ then
6              Encolar(Q, Φ_{Ω_1}(t)) ;
7          else
8              Adicionar Φ_{Ω_1}(t) a STR ;
9  while |Q| ≠ 0 do
10     X = Desencolar(Q) ;
11     /* X es una relación en la forma Φ_{Ω_j}(x_1, x_2, ..., x_n) */
12     k = last(X) ;
13     for s = k + 1 to |T| do
14         pushed = False;
15         for i = j to |Ω| do
16             R = Γ(Ω_i, X, t_s);
17             foreach R' in R do
18                 if Sop(R') ≥ μ then
19                     if DF(R', N) ≠ ∅ then
20                         Encolar(Q, R') ;
21                     else
22                         Adicionar R' a STR;
23                         pushed = True;
24             if pushed == True then
25                 break;
```

como:

$$\Psi(R, C, x, \Omega_i) = \bigcup_{c_k \in \{\wp(C) \setminus \emptyset\}} \{\sigma\left(\rho\left(R, c_k\right), \Phi_{\Omega_i}\left(t_j | t_j \in c_k, x\right)\right)\} \tag{4.7}$$

$$\Delta(R, C, x, \Omega_i) = \bigcup_{c_k \in \{\wp(C) \setminus \emptyset\}} \bigcup_{V \in \Xi(c_k, x, \Omega_i)} \{\sigma\left(\rho\left(R, c_k\right), V\right)\} \tag{4.8}$$

Siendo $\wp\left(C\right)$ el conjunto potencia de C y Ξ una función definida como:

$$\Xi\left(C, x, \Omega_i\right) = \Gamma\left(C_1, x, \Omega_i\right) \times \ldots \times \Gamma\left(C_k, x, \Omega_i\right) \tag{4.9}$$

Cuando el contexto de la relación R coincide con el contexto objetivo ($\Omega_R = \Omega_i$), la función Γ devuelve un conjunto compuesto por una única relación la cual coincide con la relación R a la cual se le adiciona el elemento x. Por ejemplo:

$$\Gamma\left(D, \Phi_D\left(\textit{autobús}\right), \textit{escolar}\right) = \left\{\Phi_D\left(\textit{autobús}, \textit{escolar}\right)\right\}.$$

En otro caso, cuando no coincide el nivel de contexto deseado con el contexto de la relación R, debe adicionarse el elemento x a la relación R en el contexto solicitado. Para ello se emplean las funciones Ψ y Δ. La función Ψ genera las siguientes relaciones a explorar cuando se adiciona el nuevo término a los elementos simples. Por otra parte, Δ es la encargada de generar las nuevas relaciones cuando se adiciona el nuevo término a los elementos restantes. A continuación se ilustra cómo operan las funciones Ψ y Δ. Sea la relación $R = \Phi_D\left(x_1, x_2, \Phi_P\left(x_3, x_4\right), \Phi_P\left(x_3, x_5\right)\right)$, a la cual deseamos adicionar el término z en el nivel $\Omega_i = P$. Dado que el nivel de contexto Ω_i no coincide con el de la relación R, se recurre a las funciones Ψ y Δ. Para facilitar la comprensión del ejemplo, denotaremos como $G_1 = \Phi_P\left(x_3, x_4\right)$ y $G_2 = \Phi_P\left(x_3, x_5\right)$. Así, la relación R queda expresada de la forma $R = \Phi_D\left(x_1, x_2, G_1, G_2\right)$. Los conjuntos H_1 y H_2 quedan

definidos como: $H_1 = \{x_1, x_2\}$ y $H_2 = \{G_1, G_2\}$. La llamada a la función Ψ quedaría como: $\Psi\left(R, \{x_1, x_2\}, z, \Omega_i\right)$. Dado que

$$\wp\left(\{x_1, x_2\}\right) = \{\{x_1\}, \{x_2\}, \{x_1, x_2\}, \emptyset\}$$

Siendo $\Phi_D\left(x_2, G_1, G_2, \Phi_P\left(x_1, z\right)\right)$ la primera de las relaciones que se generaría a partir de la función Ψ.

El comportamiento de la función Δ es relativamente similar al de la función Ψ, en este caso la llamada a la función quedaría como: $\Delta\left(R, \{G_1, G_2\}, z, \Omega_i\right)$. Para esta función la primera de la relaciones generadas por esta función es: $\Phi_D\left(x_2, G_2, G_1^{'}\right)$ con $G_1^{'} = \Phi_P\left(x_3, x_4, z\right)$.

Cuando el contexto Ω_i en el cual se desea adicionar el nuevo término es más específico que los contextos considerados en la relación R se emplea la función Ξ.

Cada vez que es encontrada una relación R que es capaz de diferenciar los documentos positivos de los negativos, el algoritmo no continúa explorando otras relaciones en contextos más específicos con los elementos de R, ni adiciona nuevos elementos a la relación R.

La función Γ es un componente fundamental en el Algoritmo 1. Esta función es la encargada de generar las relaciones que deben explorarse para diferenciar los documentos Relevantes de los No Relevantes. Esta función involucra la computación de relaciones sobre combinaciones en el conjunto potencia de los elementos que forman la relación R. Por ello, esta función tiene en general un costo computacional de orden exponencial. No obstante, si tomamos en consideración que a medida que a una relación se le adicionan más elementos, ésta se vuelve cada vez más específica, y por ende la probabilidad de que ella pueda ser encontrada en nuevos documentos disminuye grandemente. Por lo cual no es preciso generar en nuestro algoritmo relaciones que involucren un valor elevado de

términos. En nuestros experimentos empleamos como máximo relaciones que involucran a lo sumo 5 términos, valor que fue establecido de forma empírica. Al establecer una cota superior a la cantidad de términos que puede contener una relación, el costo de la función Γ se encuentra acotado por un valor fijo que coincide con el número máximo de combinaciones que pueden generarse en relaciones de hasta 5 elementos, al considerarse el nivel de contexto más específico.

En el caso del costo computacional del Algoritmo 1, el caso peor es aquel en el cual un mismo documento se encuentra tanto en el conjunto de muestras Relevantes como entre las No Relevantes, y se emplea un valor de 1 para el parámetro μ. Ante este escenario, en el primer ciclo del algoritmo se tendría que adicionar una relación a la cola Q por cada término presente en T, por lo cual el costo computacional del mismo es $O(|T|)$

Para realizar el análisis en el siguiente ciclo, comenzaremos a partir de la llamada a la función Γ. Esta función es la encargada de expandir las nuevas relaciones a explorar. Como se mencionó con anterioridad, para un caso de uso práctico la llamada a ésta función se limita a relaciones que involucren a no más de 5 términos. Esto conlleva a que en términos de complejidad algorítmica el costo de la misma pueda considerarse $O(1)$. Este hecho lleva a que el tamaño del conjunto de relaciones que se pueden generar se encuentre igualmente acotado por las relaciones que se pueden generar considerando 5 términos con el nivel más de contexto más específico. Por lo tanto el ciclo encargado de adicionar las relaciones a la cola Q se encuentra igualmente acotado y su tiempo de ejecución es $O(1)$. La cantidad de llamadas que se realizan a la función Γ se encuentra condicionada a los diferentes niveles de contexto que se emplean en el algoritmo (ciclo *for* en la línea 15). Sin embargo, los contextos posibles a explorar no varía y es un valor fijo en el algoritmo, lo que nos lleva a que este ciclo también tiene un tiempo de ejecución $O(1)$. Esto nos deja a que el tiempo de ejecución del ciclo for en la línea 13

tiene un tiempo de ejecución $O(|T|)$. El análisis del ciclo principal *while* es un poco más complicado de comprender dado que a simple vista no hay una relación directa entre él y el valor $|T|$. Este ciclo depende de la cantidad de relaciones encoladas Q, que a su vez depende de la función Γ, pues es ésta la encargada de generar la explosión de relaciones nuevas. Como en Γ se trabaja con la generación de subconjuntos de términos, y estamos limitando el tamaño de los mismos a un máximo de 5 términos, la cantidad de relaciones que se adicionan a la cola tiene un orden de

$$O\left(\frac{|T|!}{5!\,(|T|-5)!}\right)$$

lo cual se reduce a que el costo computacional del ciclo, y con ello del algoritmo sea de $O\left(|T|^5\right)$, en el peor de los casos.

Las condiciones en las cuales se da el peor caso en la ejecución del algoritmo son muy poco realistas, debido a que valor μ siendo 1 implica que todos los términos que aparecen en cualquiera de los documentos Relevantes es de interés y caracteriza a la necesidad de información del usuario, lo cual es irreal. Esto se tendría que dar junto con el hecho de que un mismo documento se encuentre entre los documentos que satisfacen la necesidad de información del usuario y a la vez no satisfaga esta necesidad, lo cual es ilógico.

En el algoritmo hay varios elementos que permiten acotar sustancialmente el espacio de búsqueda de la solución. El primero de ellos es el valor del umbral μ, pues términos cuya frecuencia de aparición no supera su valor son descartados. Una relación que con sus elementos actuales puede diferenciar a los documentos Relevantes de los No Relevantes no es expandida. Por último, en el ciclo for de la línea 15 debemos notar que si con el nivel de contexto actual se pueden diferenciar los documentos de interés de los que no lo son, el ciclo se interrumpe y no se exploran niveles de contexto más específicos.

Todos estos combinados permiten reducir el volumen de operaciones requerido para encontrar las relaciones que se emplean para representar el perfil del usuario.

4.5 Términos Frecuentes Globales

El número de relaciones que pueden ser extraídas a partir de un conjunto de documentos es enorme, sin embargo la mayoría de éstos son elementos ruidosos o poco representativos. La extracción de un volumen elevado de relaciones no representativas no solo influye en la eficiencia del sistema, dado que muchas más relaciones deben ser evaluadas para determinar si un documento es relevante o no para el usuario, sino que también influye en la calidad obtenida, pues pueden seleccionarse documentos que no son realmente de interés.

Una posible estrategia para reducir el número de relaciones extraídas puede ser el limitar el espacio de búsqueda de las relaciones solamente a aquellas compuestas por términos que con mayor frecuencia se encuentran en los documentos Relevantes. Sin embargo, esta idea es válida únicamente cuando estos documentos se encuentran distribuidos homogéneamente entre los subtópicos del perfil.

Un tópico usualmente está compuesto por varios subtópicos, los cuales no necesariamente están expresados en la misma cantidad de documentos. Cuando en un tópico algunos subtópicos aparecen en muchos documentos y otros en pocos, y extraemos los términos frecuentes ignorando cómo se encuentran los documentos distribuidos entre los diferentes subtópicos, podemos extraer muchos términos que ocurren frecuentemente en los subtópicos más representados y muy pocos, o incluso ninguno, de los términos que solo ocurren entre los subtópicos poco representados en el perfil.

Los subtópicos que componen el tópico pueden estar desbalanceados por varias razones. Por ejemplo, un subtópico que en el flujo de documentos está emergiendo y

comienza a ser detectado, o porque el subtópico en sí representa casos especiales y excepcionales.

Para ilustrar, si se supone que se está interesado en "*Accidentes aéreos y las regulaciones existentes para evitarlos*" y nuestro conjunto de entrenamiento está compuesto por varias noticias relacionadas con un accidente en particular y otros dos conjuntos de noticias más pequeños, uno integrado por unas pocas noticias sobre regulaciones aeronáuticas y el otro por noticias que hacen referencias a seguridad en los aeropuertos europeos. Si seleccionamos únicamente los términos frecuentes, probablemente seleccionaremos términos relacionados con el accidente, tales como el nombre de la aerolínea involucrada, el lugar del suceso, el tipo de aeronave, número de victimas, etc., y seguramente se descartarían otros términos relevantes relacionados con los otros dos subtópicos pocos representados.

Por otro lado, aún cuando un tópico puede estar compuesto por varios subtópicos, siempre podremos encontrar un conjunto de términos que engloban los elementos fundamentales, o generales, que resumen su contenido y pueden ser vistos como un hilo conductor entre los diferentes subtópicos. Estos términos aparecen frecuentemente entre los diferentes subtópicos que forman el tópico. Por ejemplo, si retomamos el ejemplo relacionado con la Aviación Civil y las regulaciones que la rigen, entre los subtópicos de los cuales se compone ésta temática encontramos los resultados económicos de las mismas, documentos relacionados con accidentes e incidentes, así como aquellos que tratan las regulaciones en el sector. Aunque tenemos varios subtópicos diferentes, existe un conjunto de términos que caracterizan de forma global la temática, independientemente de los subtópicos que la componen. En este caso en particular, tendríamos términos como aerolínea, avión, aeronave, aeropuerto, piloto, etc. A este conjunto de términos les llamaremos "*términos frecuentes globales*" (TFG). Estos términos pueden ser empleados con la finalidad de disminuir el número de relaciones extraídas, evitando aquellas

poco representativas o ruidosas.

Para determinar este conjunto de términos, primeramente realizamos un agrupamiento sobre los documentos Relevantes disponibles para la construcción del perfil. Un algoritmo de agrupamiento estructura los documentos en grupos de forma tal que documentos que son ubicados en un mismo grupo son muy parecidos entre sí, y guardan una menor semejanza con los documentos ubicados en otros grupos. En este trabajo suponemos que cada grupo está asociado con un subtópico diferente. Al conjunto de subtópicos detectados lo denotaremos como $L = \{L_1, L_2, \ldots, L_c\}$.

Una vez estructurados los documentos en subtópicos extraemos los términos TFG como aquellos términos más frecuentes entre los diferentes subtópicos. Para ello empleamos la expresión siguiente:

$$TFG = \left\{ t_i \mid t_i \in T \ \wedge \ \frac{|\{L_k \mid df(t_i, L_k) > 0\}|}{|L|} \geq \gamma \right\} \tag{4.10}$$

La función *df* devuelve como resultado la cantidad de documentos del grupo L_k en los cuales aparece el término t_i. γ es un parámetro a definir.

De acuerdo a la expresión anterior, los TFG son aquellos que aparecen en la mayor cantidad de los diferentes subtópicos que forman el interés del usuario.

Una vez seleccionados los TFG, durante el proceso de extracción de las relaciones entre términos tomamos en consideración solo aquellas que involucran al menos a uno de los TFG.

4.6 Atendiendo el desbalance entre los subtópicos

El uso de un mismo valor de soporte mínimo para seleccionar las relaciones que formarán el perfil del usuario impide que los subtópicos menos representados puedan ser atendidos de forma correcta.

Con la finalidad de tomar en consideración los subtópicos menos representados en el perfil del usuario eliminamos del Algoritmo 1 el empleo de la función de soporte para determinar si una relación formará parte del perfil del usuario. Para ello introducimos el concepto de compatibilidad entre una relación R y un conjunto de relaciones S.

Definición: Una relación candidata R es compatible con respecto a un conjunto de relaciones $S = \{S_1, ..., S_n\}$ si, siendo $X = DF(R, P)$, $G^1 = \{S_j | \, |DF(S_j, P)| > |X|\}$, y $G^2 = \{S_j | \, |DF(S_j, P)| = |X| \wedge |S_j| < |R|\}$, se cumple que:

$$X \setminus \bigcup_{G_i^1} DF\left(G_i^1, P\right) \neq \emptyset \ \wedge X \setminus \bigcup_{G_i^2} DF\left(G_i^2, P\right) \neq \emptyset \tag{4.11}$$

El operador $|R|$, siendo R una relación, devuelve el número de términos diferentes existentes en R. Recordemos que $DF(R, P)$ devuelve los documentos pertenecientes al conjunto P en los cuales se satisface la relación R.

En el Algoritmo 2 detallamos el proceso de extracción de las relaciones, tomando en consideración el concepto de compatibilidad.

El empleo de la definición de compatibilidad, en el Algoritmo 1, garantiza que podamos descartar las relaciones que cubren un conjunto de documentos que puedan ser abarcados por otras relaciones con una mayor presencia entre los documentos Relevantes, o por aquellos con igual presencia pero que involucran un menor número de términos. En la medida que un mayor número de relaciones puedan ser descartadas y no exploradas, más eficiente será el algoritmo 2.

Los algoritmos basados en patrones al ser aplicados a la tarea del filtrado emplean el soporte para descartar las relaciones. Sin embargo, este enfoque en la tarea del Filtrado Adaptativo se vuelve muy complicado por las diversas circunstancias que se presentan en la tarea, tales como:

- Poca información para la construcción del perfil inicial.

Algoritmo 2: Extracción de las relaciones de términos

Entrada: P: Documentos Relevantes (Positivos)
N: Documentos No Relevantes (Negativos)
T: Conjunto de términos presentes en P
Ω: Arreglo de Tipos de Relaciones

Salida: STR: Conjunto de relaciones

```
1  STR = {} ;
2  Q ← ∅ /* Cola de Prioridades Vacía */
3  foreach t in T do
4      Encolar (Q, Φ_Ω1 (t)) ;
5  while |Q| ≠ 0 do
6      X = Desencolar(Q) ;
7      /* X es una relación en la forma Φ_Ωj (x_1, x_2, ..., x_n) */
8      if not Compatible (X, P, STR) then
9          continue;
10     if DF (X, N) == 0 then
11         Adicionar X a STR ;
12     else
13         foreach t in T \ {s|s in X} do
14             for i = j to |Ω| do
15                 R = Γ (Ω_i, X, t);
16                 W = {};
17                 foreach R_k in R do
18                     if DF (R_k, P) ≠ ∅ then
19                         Adicionar R_k a W;
20                 if |W| ≠ 0 then
21                     foreach R_k in W do
22                         Encolar (Q, R_k) ;
```

- Aparición de nuevos subtópicos en el flujo de documentos con el paso del tiempo.

- Presencia no homogénea de documentos entre los diferentes subtópicos.

La escasa información provoca que las estadísticas que se extraen no sean fiables. Por otro lado, el proceso de retroalimentación permite que el número de muestras disponibles crezca con el paso del tiempo, sin embargo, si bien esto es favorable porque se disponen de un mayor número de muestras para modelar el interés del usuario, provoca que los valores de soporte determinados cuando se tienen pocas muestras no sean factibles cuando el número de muestras crece y viceversa. Por último, la coexistencia de subtópicos nuevos con otros anteriores, trae como consecuencia que un mismo valor de soporte deba emplearse para subtópicos pequeños y para otros más grandes.

En nuestra propuesta favorecemos el uso de relaciones tan simples como sea posible, siempre que ellas sean capaces de diferenciar las muestras relevantes de las no relevantes. El empleo de relaciones simples permite ir ajustando la modelación del interés del usuario sin comprometer la capacidad de recuperación (recuerdo) del proceso de filtrado. Por el contrario, si optamos por el empleo de relaciones más complejas, podemos caer en un sobreajuste del modelo, y por consecuencia en una rápida caída de la efectividad del clasificador. Por otro lado, con nuestra definición de Compatibilidad, evitamos el empleo de un umbral para el valor del soporte, y las complicaciones que están asociadas con el ajuste de un umbral.

Nuestro algoritmo, en correspondencia con el flujo de documentos y la retroalimentación recibida por parte del usuario, automáticamente ajusta el nivel de especificidad del contexto empleado, así como el número de términos involucrados para diferenciar los documentos positivos de los negativos, con el fin de obtener una buena recuperación.

4.7 Atendiendo el Inicio en Frío

Como fue mencionado en la sección 2.1, el fenómeno del Inicio en Frío debe tomarse en consideración si se desea ofrecer una buena solución al problema de filtrado. En nuestra propuesta, el Inicio en Frío es tratado desde varias perspectivas. Una de ellas ya fue descrita con anterioridad y está relacionada con la forma en que son seleccionadas las relaciones que formarán el perfil del usuario. El favorecer las relaciones más simples por sobre las que tienen una mayor cantidad de términos posibilita atender el problema del Inicio en Frío pues, al no tenerse muestras suficientes para una correcta modelación del perfil del usuario, el uso de relaciones simples permite una mayor probabilidad de obtener un emparejamiento entre las relaciones y los nuevos documentos del flujo.

Otra de las formas en la que es combatido este problema es mediante la adición de conocimiento del mundo a los documentos. Para ello, empleamos recursos que se encuentran disponibles como es el caso de JRC-Names, Wikipedia y Geonames.

JRC-Names[1][Steinberger et al., 2011] es un recurso multilingüe de nombres de entidades (personas y organizaciones) desarrollado por el Centro de Investigación Conjunta, más conocido por *JRC* (en inglés *Joint Research Centre*). Este recurso es empleado con la finalidad de identificar y unificar nombres de entidades en los documentos procesados.

Otro recurso empleado en nuestro trabajo es la Wikipedia[2]. Ésta es una enciclopedia libre, editada de forma colaborativa y voluntaria por millones de personas en todo el mundo y en varios idiomas. De la Wikipedia, a través del procesamiento de uno de los archivos que contienen toda la información en forma de fichero XML comprimido en formato ZIP, extraímos igualmente nombres de personas y organizaciones así como diversas formas en las cuales pueden referirse a una misma entidad, por medio de un procesamiento inteligente de las páginas de redireccionamiento presentes en ella.

[1]https://ec.europa.eu/jrc/en/language-technologies/jrc-names
[2]https://en.wikipedia.org/

La Wikipedia está estructurada en forma de páginas. Estas páginas pueden dividirse en dos grandes grupos, las de contenidos y las de estructura. Las páginas de contenido se dividen a su vez en una página, principalmente, por cada entidad diferente. Por el contrario, las páginas de estructura son páginas que se emplean para categorizar las páginas de contenido o para asociar diferentes formas de referirse a un mismo concepto o entidad. De la Wikipedia extraímos las páginas asociadas a Personas y Organizaciones, procesando las páginas de categorías, las plantillas y algunos elementos del contenido de las mismas. Una vez extraídas las entidades se analizaron las páginas de redireccionamiento para buscar las diferentes formas de referirse a una misma entidad.

Por ejemplo, del análisis de las organizaciones se puede extraer la entidad *OTAN*, y del análisis de los enlaces se puede extraer que los nombres: *Alianza Atlántica*, *NATO*, y la *Organización del Tratado del Atlántico del Norte* se emplean para referirse a la misma entidad.

La información extraída a partir de la Wikipedia se empleó durante el procesamiento de los documentos para detectar los nombres de entidades presentes en los documentos. De igual manera, si se encuentra en el documento una de las diferentes variantes de referirse a una entidad ésta es sustituida por el nombre de la entidad a la que hace referencia.

A diferencia de los dos anteriores, GeoNames[3], el tercer recurso empleado, es una base de datos con información geográfica, la cual cubre todos los países y millones de nombres de lugares. A través de ella podemos identificar continentes, países, regiones geográficas, etc. Esta información permite procesar de forma más adecuada los perfiles de los usuarios que requieren este tipo de información, por ejemplo: *Turismo en Europa*.

El procesamiento de este recurso permite no solo identificar nombres de lugares, sino que permite obtener una información jerárquica de los continentes, regiones y países.

[3]http://www.geonames.org/

Así, nos permite expandir la información geográfica presente en los documentos. Por ejemplo, si tenemos la oración: "*En Etiopía podemos encontrar más de 10 grupos étnicos*". Empleando GeoNames no solo identificamos a Etiopía, sino que además podemos inferir que estamos hablando de un país del continente africano, inclusive se pueden identificar regiones, como es el caso de que Etiopía forma parte de los países del cuerno africano, por lo que pudiera darse un mejor tratamiento a tópicos como pudiera ser: "*Demografía en los países del cuerno africano*"

El Inicio en Frío es un fenómeno complejo que debe ser tomado en cuenta cuando se construye una propuesta para el Filtrado Adaptativo. Este fenómeno, relacionado con la escasa disponibilidad de documentos para la construcción del perfil, es tratado desde la óptica de dos perspectivas diferentes. La primera de ellas se trata en la construcción de las relaciones empleadas para representar el perfil. La segunda perspectiva consiste en realizar un procesamiento inteligente del contenido de los documentos para adicionarle conocimiento del mundo.

Como se ha mencionado previamente, las relaciones más complejas tiene una menor probabilidad de aparición en nuevos documentos. Una relación se vuelve más compleja a medida que se le adicionan términos o se consideran niveles de contexto más específicos. Los métodos basados en patrones favorecen la extracción de patrones compuestos por varios términos por sobre aquellos que se componen por un menor número. Sin embargo, cuando se extraen los patrones a partir de poca información se pueden extraer patrones compuestos por varios términos, con la consecuente disminución de la probabilidad de encontrar luego dichos patrones en el flujo de documentos. Para evitar esta situación, en nuestro método de extracción se favorecen las relaciones más simples posible, siempre que éstas sean capaces de diferenciar los documentos Relevantes de los No Relevantes. Al extraerse relaciones más simples es posible que aún cuando se tengan pocos documentos para la construcción del perfil, las relaciones extraídas tengan una mayor probabilidad

de ser encontradas en los documentos del flujo de información.

Los humanos cuando leemos un documento conectamos su contenido con el conocimiento previo que poseemos. Sin embargo, durante el procesamiento del flujo de información ese conocimiento adquirido no está disponible. Esta limitante se puede ir disminuyendo en la medida que tenemos un mayor volumen de documentos para la construcción del perfil del usuario. Pero cuando partimos de muy pocos documentos para la construcción del perfil, la información disponible es muy escasa y esto limita sustancialmente las posibilidades de seleccionar para el usuario documentos del flujo de información si para ello se requiere de conocimiento del mundo. Con la finalidad de disminuir esta limitante en nuestra propuesta empleamos recursos externos que nos ayuden a enriquecer la información de los documentos. Este enriquecimiento se realiza durante el procesamiento de los documentos para obtener la representación empleada por nuestra propuesta antes de realizarse el proceso de filtrado. Los recursos JRC-Names y Wikipedia permiten identificar entidades, así como diversas variantes con las cuales suelen referirse a las mismas. Todas estas variantes son asociadas a un mismo token en los documentos.

En el caso de GeoNames, su contenido es empleado en el perfil para adicionar información geográfica en las relaciones que hacen referencia a ella. Por ejemplo, si en una relación se tiene una referencia al continente europeo, en las relaciones se considera además que el documento contenga cualquiera de los integrantes de dicho continente.

4.8 Experimentos

En los experimentos realizados para evaluar nuestra propuesta se utilizaron las colecciones Reuters 21578 y RCv1 (descritas en la sección 2.6) y las medidas T11SU y F_1 (presentadas en la sección 2.7).

Los documentos son procesados empleando la herramienta FreeLing de Padró and Stanilovsky [2012]. Los documentos fueron tokenizados, lematizados y se realizó además un análisis sintáctico superficial. Las palabras carentes de significado semántico (conocidas comúnmente como *Palabras Vacías*) fueron eliminadas. Se empleó el reconocedor de entidades nombradas, de manera conjunta con los recursos JRC-Names y Wikipedia. Para la extracción de los nombres de entidades a partir de la Wikipedia se realizó el procesamiento de un archivo de salvas del contenido en Inglés de esta enciclopedia. Del procesamiento de este archivo se emplearon las páginas de categorías, las de contenido y las plantillas.

En la experimentación se siguió la metodología propuesta en las competiciones TREC, la cual fue abordada en la sección 2.1. Como fue descrito con anterioridad, para la construcción del perfil se dispone de una breve especificación del interés del usuario, expresadas por unos pocos términos en forma de consulta de un buscador web. Adicionalmente, en la colección RCv1 se dispone de una breve descripción de un par de oraciones y dos o tres documentos relevantes de muestra. En la colección Reuters 21578 se parte de la consulta que define el tópico de interés únicamente.

Una vez construído los perfiles de los usuarios se comienza el procesamiento de los documentos respetando la fecha de publicación de los documentos. Cada documento es analizado para determinar si se ajusta o no a la información recogida en el perfil de un usuario. Si se determina que el contenido del documento satisface la necesidad de información de un usuario, éste es presentado al usuario, en caso contrario el usuario nunca tiene acceso a la información del mismo. Cada vez que un documento es seleccionado para ser mostrado al usuario, el sistema obtiene de forma inmediata la retroalimentación del usuario indicando si el mismo es verdaderamente relevante o no para el perfil.

Como fue mencionado con anterioridad las relaciones se limitaron a un máximo de 5 términos. El valor de μ para el valor mínimo de soporte fue fijado de forma tal que

se seleccionara el 20 % de los documentos relevantes disponibles para la construcción del perfil. Para la selección de los Términos Frecuentes Global (TFG) el valor de γ fue fijado en 0.33. Estos valores fueron fijados de forma empírica.

Los experimentos fueron diseñados para medir el impacto de cada uno de los componentes y propuestas realizadas en este trabajo en la tarea del filtrado adaptativo.

4.8.1 Resultados obtenidos

En la presente subsección se muestran los resultados alcanzados en la tarea del Filtrado Adaptativo al emplear las relaciones de términos.

Relaciones de términos multinivel aplicadas al Filtrado Adaptativo

Los primeros experimentos aquí presentados están orientados a mostrar el comportamiento de los algoritmos de filtrado al emplear en la representación del perfil los conjuntos frecuentes compuestos por la mayor cantidad de elementos en oposición a aquellos compuestos por la menor cantidad de elementos. En el Algoritmo 1 se favorecen la relaciones compuestas por pocos elementos por encima de aquellas que se componen por una mayor cantidad de elementos. En la Tabla 4.1 se contrastan los resultados de aplicar el Algoritmo 1 en comparación con los resultados que se obtienen cuando se extraen las relaciones con el mayor número de elementos. En esta tabla son mostrados además los resultados alcanzados cuando además del nivel Documentos, son considerados los niveles Párrafo (P), Oración (S) y Sintagmas Nominales (N).

En la tabla, la fila *Extenso* se refiere a los resultados alcanzados al emplear conjuntos frecuentes con la mayor cantidad de términos y D para los alcanzados al emplear la estrategia de seleccionar los más cortos. En las filas DP, DPS y DPSN se muestran los resultados obtenidos cuando, además del nivel Documento (D), se emplean el resto de

Medida	T11SU		Macro F1	
Niveles	TREC	Reu10	TREC	Reu10
Extenso	0.367	0.357	0.248	0.077
D	0.415	0.484	0.368	0.358
D+P	**0.428**	0.491	0.403	0.369
D+P+S	0.427	**0.494**	**0.405**	0.373
D+P+S+N	0.427	**0.494**	**0.405**	**0.374**

Tabla 4.1: Resultados obtenidos al emplear un valor de soporte para la selección de las relaciones (Algoritmo 1).

los niveles de interés.

La primera conclusión que podemos obtener una vez inspeccionados los resultados presentados en la Tabla 4.1 es que con el uso de conjuntos frecuentes extensos obtenemos resultados inferiores a los obtenidos cuando se emplean conjuntos frecuentes con la menor cantidad de términos requeridos para diferenciar los documentos Relevantes de los No Relevantes. Este comportamiento se debe a que la extracción de relaciones compuestas por más términos conlleva a una mayor restricción impuesta a los documentos del flujo, pues será preciso que contengan una mayor cantidad de términos para satisfacer una determinada relación. Esta situación es particularmente desventajosa cuando se cuenta con muy pocos documentos para la construcción del perfil del usuario pues se extraen relaciones muy específicas que complican el proceso de selección de los documentos del flujo.

La segunda observación que podemos hacer a partir de estos resultados es que se puede observar una mejoría consistente al considerar, además del nivel Documento , las relaciones a nivel de Párrafo. Esta situación se mantiene al tomar en cuenta además los niveles más específicos, aunque en menor medida y menos uniforme en comparación a los obtenidos con el nivel de Párrafo.

Medida	T11SU		Macro F1	
Niveles	TREC	Reu10	TREC	Reu10
D	0.423	0.602	0.497	0.501
D+P	0.444	0.61	0.512	0.51
D+P+S	0.447	**0.616**	0.515	**0.516**
D+P+S+N	**0.448**	**0.616**	**0.516**	**0.516**

Tabla 4.2: Resultados obtenidos al eliminar el uso de un valor de soporte (Algoritmo 2).

Evaluación de estrategia para desbalance entre subtópicos

En la Tabla 4.2 se muestran los resultados alcanzados en el filtrado cuando se incluye en el proceso de extracción de las relaciones la estrategia para combatir el desbalance entre los subtópicos. En la Figura 4.3 se muestra una comparativa entre los resultados alcanzados al emplear los Algoritmos 1 y 2, el primero de ellos usa un valor de soporte para seleccionar las relaciones, mientras que el segundo no emplea este valor.

Como puede notarse en la figura, al modificar la estrategia de selección de relaciones se obtiene una mejora en la calidad en ambas colecciones. Este comportamiento se mantiene con las dos medidas de evaluación. Igualmente se mantiene el comportamiento mediante el cual al emplear niveles de contextos más específicos se obtienen resultados ligeramente superiores.

Evaluación del empleo de TFG

Al emplear conjuntos frecuentes, uno de los problemas que deben afrontarse está relacionado con la cantidad de subconjuntos que deben explorarse, además de que pueden extraerse relaciones ruidosas que afectan la calidad del clasificador. Los resultados obtenidos al emplear los TFG son presentados en la Tabla 4.3.

Del análisis de los resultados presentados en la tabla 4.3 podemos notar que se mantiene la tendencia a obtener una mejoría, aunque leve pero consistente, en la medida

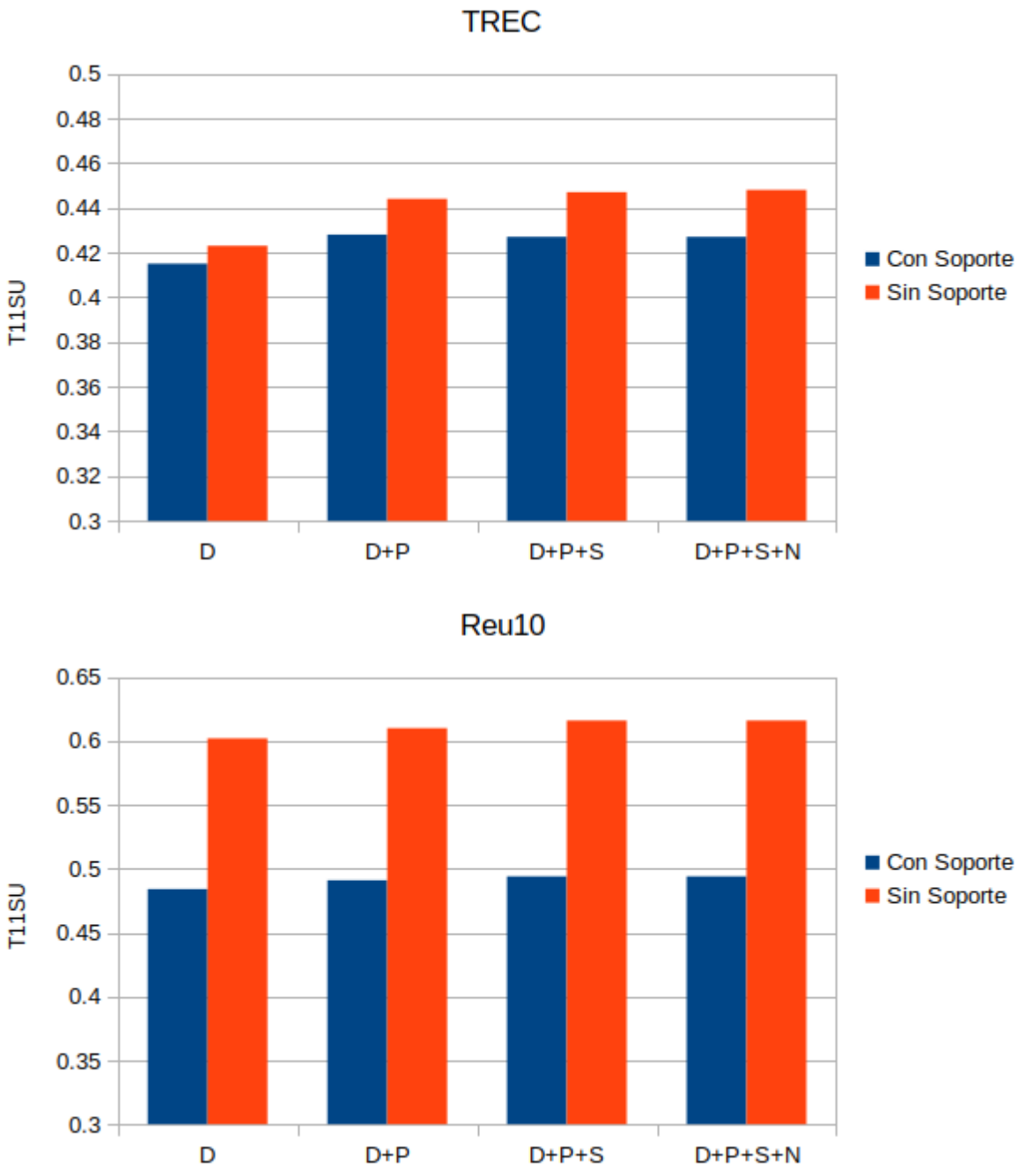


Figura 4.3: Comparación de los resultados obtenidos al emplear soporte para selección de relaciones.

Medida	T11SU		Macro F1	
Niveles	TREC	Reu10	TREC	Reu10
D	0.499	0.544	0.535	0.415
D+P	0.508	0.554	0.538	0.427
D+P+S	0.512	0.555	0.541	0.428
D+P+S+N	**0.513**	**0.558**	**0.545**	**0.430**

Tabla 4.3: Resultados obtenidos al considerar en el Algoritmo 2 el uso de los TFG durante la extracción de las relaciones.

que se consideran niveles más específicos de granularidad.

En la Figura 4.4 se presenta una comparativa de los resultados obtenidos al emplear el uso de los TFG. Del análisis de la figura podemos notar que en la colección TREC se obtiene un aumento en el valor de eficacia obtenida. Por el contrario, si analizamos el comportamiento en la colección de datos Reu10 observamos que los resultados decrecen al emplear los TFG. Esta disminución en la calidad obtenida al emplear los TFG en la colección Reu10 se debe a que, en esta colección, el conocimiento inicial está compuesto únicamente por una consulta, de 1 ó 2 términos solamente, y muchas de las relaciones que se descartan al emplear los TFG servían para seleccionar documentos del flujo, aún cuando los términos que éstas contenían no eran necesariamente los más representativos del tópico, sino todo lo contrario, estaban compuestas por términos que pudieran ser considerados como de uso más general.

Evaluación de las diferentes estrategias para el Inicio en Frío

Uno de los problemas más serios que se tienen en un sistema de Filtrado Adaptativo está relacionado con la escasez de información disponible para la construcción del perfil, fundamentalmente durante las etapas iniciales del proceso.

En la Tabla 4.4, se muestran los resultados alcanzados el considerar cada una de las estrategias. En la tabla, la fila relacionada con el uso de GeoNames no se muestran resultados en la colección Reu10 debido a que en esta colección ningún tópico hace

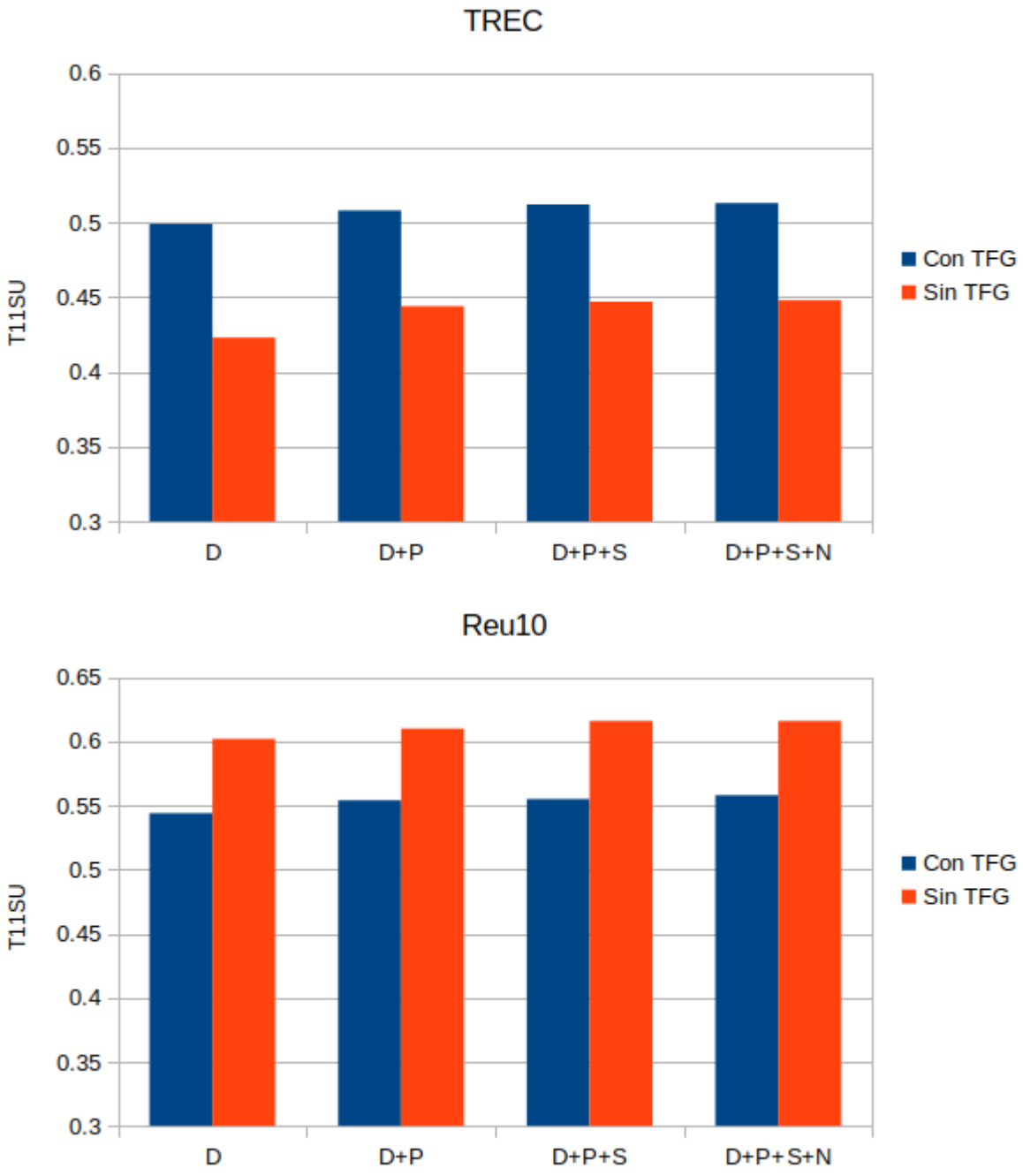


Figura 4.4: Comparación de resultados obtenidos al emplear los TFG.

Medida	T11SU		Macro F1	
Estrategia	TREC	Reu10	TREC	Reu10
Sin Soporte	0.448	**0.616**	0.516	**0.516**
TFG	0.513	0.558	0.545	0.430
Wikipedia	0.512	0.558	0.547	0.430
GeoNames	**0.531**	-	**0.559**	-

Tabla 4.4: Comparativa de las diferentes estrategias.

referencia a locaciones.

Del análisis de los resultados mostrados en la tabla, podemos observar primeramente que, contrario a lo que habíamos supuesto, el uso de la Wikipedia para enriquecer y complementar la información de los documentos del flujo, no mejoró la calidad de los resultados de acuerdo a la métrica específica de filtrado. Segundo, podemos destacar que el empleo de los nombres de lugares, cuando éstos son especificados en el tópico, permite mejorar la calidad obtenida por el clasificador.

Evolución de estrategias con el tiempo

De los algoritmos de filtrado adaptativo se espera que con el paso del tiempo mejoren su eficacia en el proceso de filtrado, por medio del aprovechamiento que puedan hacer de la retroalimentación proveniente por parte del usuario, cuando éste indica que un documento que le es mostrado por el sistema es realmente de su interés o no.

En la Figura 4.5 se muestra el comportamiento a lo largo del tiempo de las diferentes estrategias exploradas en la investigación, en la colección TREC. Del análisis de la figura podemos extraer varias conclusiones interesantes. La más notable resulta ser la marcada diferencia obtenida en la calidad al principio del proceso de filtrado en comparación con la del período final. El segundo elemento que debe notarse es que, aunque discreto, el empleo de relaciones más cortas obtienen mejores resultados a los obtenidos cuando se emplean los conjuntos compuestos por un mayor número de elementos. De igual

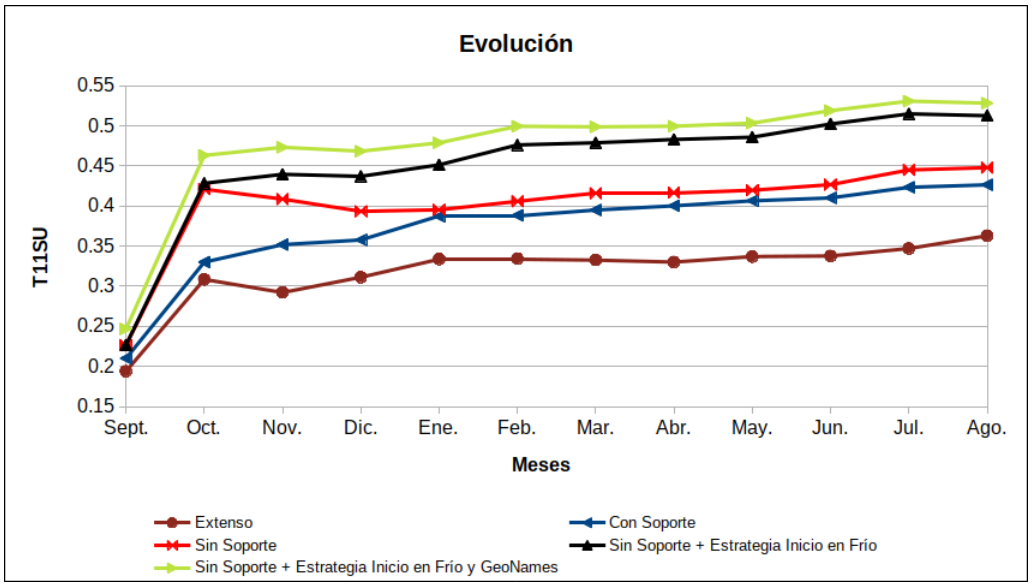


Figura 4.5: Evolución de las diferentes estrategias a lo largo del tiempo en la colección TREC.

manera, los mejores inicio de partida se obtienen cuando no se emplea un valor de soporte para la selección de las relaciones que forman el perfil. Esta última diferencia se vuelve mucho más notable cuando ha transcurrido el primer mes, con lo cual se puede notar que el no utilizar valores de soporte beneficia al inicio del proceso de filtrado.

Debe destacarse además que al emplear las estrategias para combatir el desbalance entre los subtópicos de interés y el inicio en frío propuestas, se obtienen valores de calidad superiores, que se manifiestan desde las primeras etapas del proceso del filtrado. Otro elemento que debe resaltarse es que se puede notar una tendencia en la gráfica a mejorar la calidad obtenida a medida que pasa el tiempo, lo cual es el comportamiento esperado en este tipo de algoritmos.

Por último, debemos destacar como el uso de un recurso como GeoNames para poder dar un mejor tratamiento a los perfiles que restringen su necesidad a una determinada zona geográfica permite mejorar la calidad del proceso en todo el tiempo que dura el proceso de filtrado.

No.	# Docs. Relevantes	# Docs. No Relevantes	Cantidad de Relaciones			
			D	D+P	D+P+S	D+P+S+N
1	15	24	22	20	13	13*
2	14	16	18	9	9*	9*
3	50	21	25	27	25	27
4	67	57	56	54	55	56
5	61	26	24	17	17	18
6	307	40	36	39	39	39*
7	15	42	22	31	30	25

Tabla 4.5: Cantidad de relaciones por niveles considerados.

Cantidad de relaciones por niveles

En la Tabla 4.5 se muestra la cantidad de relaciones que se obtienen en varios conjuntos de documentos extraídos de la colección TREC. En la tabla, se encuentran señalados con un * los valores en los cuales no varió la cantidad de relaciones extraídas, debido a que contextos menos específicos fueron suficiente para diferenciar a los conjuntos de documentos. Como puede observarse en la tabla, cada vez que se considera un nivel de contexto más específico, el número de relaciones se mantiene en valores similares o incluso disminuye. Aunque en esta tabla se muestran solo algunos conjuntos de documentos de muestra, el comportamiento se mantuvo durante toda la experimentación realizada.

4.9 Conclusiones Parciales

El filtrado adaptativo de documentos permite a los usuarios enfocarse en aquella información que le es de interés sin necesidad de inspeccionar todo el volumen de documentos disponibles. Esta tarea entraña varios retos que deben ser atendidos con el fin de mejorar la calidad de los resultados obtenidos.

En el presente capítulo se propuso una extensión a los conjuntos frequentes de térmi-

nos para su aplicación en la tarea del filtrado. Dicha extensión no solo toma en consideración las coocurrencias entre los términos sino que además toma en consideración el nivel de contexto en el que estos términos suelen co-ocurrir.

Se evaluaron además varias estrategias para combatir tanto el desbalance entre los subtópicos que forma el tópico de interés, como la escasez de información disponible en determinadas etapas del proceso. Los resultados alcanzados muestran la factibilidad y validez de las estrategias propuestas para mejorar la calidad de la tarea del filtrado adaptativo.

Capítulo 5

Indexado Aleatorio Multinivel

Las relaciones de términos multinivel capturan relaciones de co-ocurrencia entre los términos que se encuentran en los documentos relevantes disponibles para la construcción del perfil del usuario. Sin embargo, estas relaciones no son capaces de resolver el problema de la sinonimía presente en el lenguaje natural. Por otro lado, los modelos reportados en la literatura, como es el caso de LDA, LSI y word2vec, fueron diseñados para resolver en cierta medida esta limitación. El problema con estas propuestas es que su aplicación en el filtrado adaptativo se ve limitada por el costo que tienen las mismas cuando se desea adicionar nuevo conocimiento al modelo. A diferencia de éstos modelos, el Indexado Aleatorio es menos costoso y su desempeño en la tarea es similar.

En este capítulo comenzamos realizando un análisis del Indexado Aleatorio, posteriormente proponemos una extensión del mismo, a la que denominamos Indexado Aleatorio Multinivel, en la cual tomamos en consideración los diferentes niveles de granularidad pudiéndose relacionar los términos en los documentos. Después se presentan los resultados obtenidos al aplicar el modelo propuesto, en comparación con el modelo actual, en la tarea del Filtrado de Documentos. Por último, presentamos las conclusiones parciales del capítulo.

5.1 Análisis del Indexado Aleatorio

Así como otras propuestas existentes, como es el LSI o LDA, el Indexado Aleatorio no supone que las características presentes en los documentos son independientes entre sí. Pero, a diferencia de las anteriores, su construcción es menos costosa, y para adicionarle

nueva información no se requiere el tener almacenado todos los documentos empleados en su construcción, debido a que este modelo de representación puede ser construido incrementalmente.

En la literatura podemos encontrar variaciones a la formulación inicial del Indexado Aleatorio. En su formulación inicial, los vectores índices son asociados a los contextos, por ejemplo documento u oración; sin embargo una de las variantes presentadas está relacionada con la asignación de los vectores índices no a los contextos, sino a los términos que se encuentran en estos contextos, usualmente definidos como una ventana de términos [Musto, 2010]. Otra variante, conocida como *Reflective Random Indexing* [Cohen et al., 2010], asigna igualmente un vector índice a cada término, luego la representación de los documentos es obtenida como la suma de los vectores índices de los términos que contiene. Seguidamente, son empleados estos documentos para actualizar los vectores de contextos de los términos, y así iterativamente. Higging & Burstain proponen substraer la media de los vectores de contexto antes de obtener la representación final de los documentos [Higgins and Burstein, 2007].

Sin embargo, al igual que muchas otras técnicas de representación de documentos existentes, el Indexado Aleatorio no toma en consideración los diferentes niveles en los cuales los términos de un documento pueden relacionarse. La información que dos términos pueden compartir si ellos suelen co-ocurrir en los mismos documentos no es la misma a si ellos suelen co-ocurrir en las mismas oraciones, e incluso en los mismos sintagmas nominales.

Por ejemplo, si tomamos en consideración las palabras: *Obama*, *presidente*, *ciudad* y *Japón*, la co-ocurrencia de estos términos en los mismos documentos pueden referirse a la visita que realizó el Primer Ministro de Japón al presidente Barack Obama en 2013, al viaje llevado a cabo por el presidente Obama a Japón en 2016, o a la ciudad nombrada Obama, en la prefectura Fukui en Japón. Por el contrario, si *Obama* y *ciudad*

co-ocurren en el mismo sintagma nominal, al igual que *Obama* y *presidente*, así como la co-ocurrencia de éstos en la misma oración con la palabra *Japón*, entonces es más probable que se esté refiriendo a la ciudad Obama, y la publicidad adquirida por la misma a raíz de la victoria electoral alcanzada por el presidente Barack Obama por compartir nombre.

En la definición del Indexado Aleatorio tradicional es preciso especificar un nivel de contexto para el cual se le asignará los vectores índices. El problema con esta aproximación es que debe conocerse con anterioridad cuál es el contexto adecuado, lo cual no siempre es posible, además no permite tomar en consideración varios niveles de contextos simultáneamente.

5.2 Indexado Aleatorio Multinivel

Los documentos tienen usualmente una estructura en la organización de su contenido, por ejemplo: los libros se encuentran divididos en capítulos, secciones, párrafos y oraciones; mientras que los artículos científicos se estructuran en secciones, subsecciones, párrafos y oraciones. Esta organización tiene la finalidad de facilitar la comprensión de los contenidos por los lectores, por lo que la ocurrencias de los términos en cada una de estas estructuras no es por azar, sino que responden a la semántica del contenido que se desea transmitir.

Como mencionamos con anterioridad, en el Indexado Aleatorio es preciso definir qué contexto se va a emplear. Esta situación no es favorable cuando se requiere captar la semántica que se establece entre los diferentes términos en diversos contextos. Por ejemplo, supongamos que tenemos las oraciones siguientes extraídas de un documento que aborda la temática de las energías renovables:

> *Today, new fully electric buses began to circulate in the city.*
>
> ...
>
> *Other investments have been made in the installation of solar panels and wind turbines, to promote the use of clean energies.*

En este ejemplo, si tomamos todo el documento como contexto, no puede diferenciarse la presencia de las palabras *turbine* y *eletric* en un contexto diferente, como puede ser en la oración siguiente: *A new electric turbine was installed in the northwest plant.* Por otro lado, si seleccionamos como contextos las oraciones perderemos la información de que estas palabras aparecen en un mismo documento.

Tomando en consideración esta limitación, y con la finalidad de reflejar mejor la información del contenido de los documentos tomando en cuenta su estructura, y con ello la co-ocurrencia de los términos en diferentes contextos, proponemos almacenar un Vector de Contexto diferente por cada elemento de interés en la estructura de los documentos. A esta propuesta le nombramos Indexado Aleatorio Multinivel (MLRI, por sus siglas en el idioma inglés).

5.2.1 Construcción del índice

Para tomar en consideración los diferentes niveles de contexto en los cuales los términos pueden relacionarse en los documentos, el proceso de construcción de los Vectores de Contexto debe ser modificado.

Primeramente, debemos tomar en consideración que un Vector Índice distinto debe ser asignado a cada contexto diferente dentro del documento que va a ser procesado.

Si retomamos la Figura 4.1, debemos generar un vector índice por cada contexto diferente, los cuales en la figura coinciden con los elementos encerrados en círculos. Una vez generados los vectores índices, podemos obtener los vectores de contexto asociados a cada término. En el caso de la figura antes mencionada, tendríamos 4 vectores de

contexto por cada uno de los términos presentes en el documento, uno para cada nivel D, P, S y N. Generalizando, en nuestra propuesta en vez de un vector de contexto por cada término, ahora se tendrá una colección del vectores de contexto [V_{Ω_1}, V_{Ω_2}, ..., V_{Ω_n}] donde Ω_j representa los diferentes niveles de contexto tomados en consideración.

El proceso de construcción del Indexado Aleatorio Multinivel se resume en el Algoritmo 3.

Algoritmo 3: Construcción del Indexado Aleatorio Multinivel

```
1 foreach nivel Ω_j a considerar do
2     Inicializar el Vector de Contexto V_{Ω_j};
3 foreach t in documento D do
4     foreach nivel Ω_j a considerar do
5         if Un nuevo contexto comienza en t para el nivel Ω_j then
6             Generar un nuevo Vector Índice I_{Ω_j}
7         Adicionar el Vector Índice I_{Ω_j} al Vector de Contexto V_{Ω_j};
```

Para facilitar la comprensión del algoritmo 3 considerese el siguiente ejemplo:

Supongamos que $\Omega = [D, S]$ y D_1 y D_2 son documentos expresados en la forma: $D_1 = [s_1 :< t_1, t_2, t_3 >, s_2 :< t_1, t_4, t_5 >]$, consistente de las oraciones s_1 y s_2, y $D_2 = [s_3 :< t_1, t_5, t_6 >, s_4 :< t_2, t_7, t_8 >]$, compuesto por s_3 y s_4. El algoritmo generaría los vectores índices: I_{D_1}, I_{s_1} e I_{s_2} a partir de D_1 y I_{D_2}, I_{s_3} e I_{s_4} a partir de D_2. Los vectores de contexto para el término t_1 quedarían construídos como $V_D = I_{D_1} + I_{D_2}$ y $V_S = I_{s_1} + I_{s_2} + I_{s_3}$.

El proceso de construcción del Indexado Aleatorio Multinivel, aún cuando se tienen varios vectores de contexto para cada término, mantiene un costo espacial en el mismo orden que el Indexado Aleatorio tradicional, dado que el mismo se verá afectado por un valor constante equivalente a la cantidad de tipos de contextos diferentes que serán considerado.

En cuando al proceso de construcción, es importante notar que se mantienen las

mismas ventajas que ya ofrece la versión tradicional del Indexado Aleatorio:

- Los vectores de contextos pueden ser construidos de forma incremental, sin la necesidad de almacenar todos los documentos para incorporar nueva información al modelo.

- El proceso de construcción no es computacionalmente costoso, dado que el tiempo de ejecución del método es $O\left(k\left|\Omega\right|\left|D\right|\right)$, siendo $|D|$ la cantidad de términos en el documento.

- Los documentos son recorridos una sola vez durante su adición a los diferentes vectores de contexto.

- El tamaño de la representación se mantiene constante aún cuando nuevos documentos son adicionados al índice.

La idea detrás del Indexado Aleatorio Multinivel es que los términos guardan diferentes relaciones con otros términos de acuerdo al nivel de contexto que se emplee. Para ilustrar esta situación se muestran en la Tabla 5.1 los 10 términos más similares a la palabra *telemarketing*. En ésta tabla se han resaltado en negritas las palabras que se encuentran entre las más similares a *telemarketing* desde el nivel Documento. Se han subrayado las que aparecen entre las más similares por primera vez en el nivel Párrafo y con doble subrayado las que aparecen en el nivel Oración.

Podemos observar en la Tabla 5.1 que hay términos que aparecen en todos los contextos, pero hay otros que van surgiendo en la medida que se van considerando contextos más específicos. Podemos notar también que van ganando en especificidad en la medida que se van considerando los niveles más restringidos.

Este comportamiento observado con relación al término *telemarketing* se mantiene cuando se combinan los vectores de contextos de términos diferentes. Por ejemplo, en la

Nivel: D	Nivel: P	Nivel:S	Nivel:N
include	**call**	**call**	**call**
company	**company**	**company**	**company**
make	marketing	marketing	firm
call	**include**	**include**	fraud
year	**service**	mail	abusive
service	sale	telemarketer	sale
work	telephone	**service**	telemarketer
receive	telemarketer	firm	**service**
time	**work**	**work**	phone
business	consumer	sell	campaign

Tabla 5.1: Diez términos más similares en cada nivel para la palabra telemarketing.

Nivel: D	Nivel: P	Nivel:S	Nivel:N
include	**company**	**company**	**company**
call	**call**	**call**	fraud
company	customer*	customer*	financial*
make	**include**	**include**	campaign
business	**provide***	offer*	customer*
year	sell*	**provide***	**call**
provide*	product*	marketing	provider*
base*	**make***	product*	television*
part*	telephone	sell	telephone*
people*	**people***	**consumer***	internet*

Tabla 5.2: Diez términos más similares en cada nivel para telemarketing + service.

Tabla 5.2, se muestran los términos más similares cuando son combinados los términos *telemarketing + service*. En esta tabla se han señalado con * los nuevos términos que surgen en comparación con la Tabla 5.1. Podemos notar que el comportamiento en esta nueva tabla es similar a la anterior. Igualmente cabe destacar que los nuevos términos están en esta ocasión mucho más relacionados con los servicios de telemercadotécnia.

5.2.2 Obtención de la representación de los documentos

Una vez que se cuenta con un modelo de Indexado Aleatorio Multinivel construído, el mismo puede ser empleado para representar los documentos. A diferencia de la re-

presentación de los documentos en el Indexado Aleatorio tradicional, en esta propuesta se tiene un vector diferente por cada nivel, denotados aquí como $[R_{\Omega_1}, R_{\Omega_2}, \ldots, R_{\Omega_n}]$, considerando n niveles.

El proceso de obtención de esta representación es similar al empleado para la obtensión de la representación de los documentos empleando el modelo tradicional. Primeramente, se inicializan los vectores R_{Ω_j} a 0. Luego se recorre el documento y se adicionan los vectores de contexto de cada uno de los términos que lo componen en los diferentes niveles a los vectores R_{Ω_j}. Este procedimiento tiene una complejidad similar a la del proceso de actualización de los vectores de contexto, descrito por el Algoritmo 3.

5.2.3 Aplicación del MLRI en la tarea de Filtrado

La aplicación del Indexado Aleatorio Multinivel a la tarea del Filtrado de Información es similar a lo que sería la aplicación del Indexado Aleatorio tradicional. Sin embargo, dado que en el MLRI se cuenta con varios vectores de contexto en vez de solo uno, si se tienen algunas diferencias. En particular, se debe decidir si los diferentes vectores que son empleados en la representación de los documentos son concatenados uno al lado del otro o no.

Si los vectores son concatenados, se obtendría una representación compuesta por un único vector y la aplicación sería similar del Indexado Aleatorio tradicional. De esta forma el procesamiento es relativamente simple y puede ser empleado casi cualquier algoritmo de clasificación para determinar cuáles documentos del flujo deben ser seleccionados para ser mostrados al usuario. El problema de este enfoque radica en el punto de que con la concatenación de los vectores se pierde la fronteras entre ellos y los clasificadores empleados pueden "*aprender*" erróneamente con información parcial de vectores de contexto de diferente nivel.

Si se decide no concatenar los diferentes vectores que forman la representación de

los documentos, entonces se tendría que manejar de forma independiente el vector asociado a cada nivel de contexto. Al emplearse en la tarea de Filtrado debe tomarse una decisión independiente de acuerdo a los vectores asociados a cada nivel y finalmente debe aplicarse alguna estrategia que permita tomar la decisión final mediante la combinación de las decisiones tomadas para los diferentes niveles. Formalmente se tendría un clasificador para cada nivel, C_{Ω_i}, y finalmente un clasificador (M) encargado de combinar la salida de los C_{Ω_i}, es decir:

$$F\left([R_{\Omega_1}, R_{\Omega_2}, \ldots, R_{\Omega_n}]\right) = M\left(C_{\Omega_1}\left(R_{\Omega_1}\right), C_{\Omega_2}\left(R_{\Omega_2}\right), \ldots, C_{\Omega_n}\left(R_{\Omega_n}\right)\right)$$

La ventaja de esta variante radica en el hecho de que se puede manejar de forma independiente cada nivel de contexto, y con ello se evita que se mezclen de forma errónea la información asociada a los diferentes niveles. Sin embargo, al emplear esta aproximación implica un mayor costo pues se requiere de tomar decisiones independientes en cada nivel y luego combinar éstas para tomar la decisión final.

5.3 Experimentos

En la siguiente sección se presenta el marco experimental empleado para evaluar el comportamiento del Indexado Aleatorio Multinivel en la tarea del Filtrado Adaptativo.

En la tarea de Filtrado Adaptativo, como ha sido mencionado con anterioridad, uno de los problemas fundamentales que deben ser atendido es el problema de la escasez de información. Por el contrario, el Indexado Aleatorio, al igual que otros modelos semánticos, requiere de un volumen de documentos para su construcción con la finalidad de que la información codificada sea de utilidad. Para solventar esta limitación de la escasez de información, fue construido un modelo de Indexado Aleatorio a partir de noticias y documentos provenientes de Wikipedia. En la construcción de estos modelos

fueron empleados vectores de contextos de 750 elementos y los vectores de índices se tomaron con un 10 % de valores ajustados a 1. Igualmente, se consideraron 4 niveles diferentes de contextos: Documentos (D), Párrafos (P), Oraciones (S) y Sintágmas Nominales (N). Los valores para la construcción del modelo de Indexado Aleatorio se determinaron a partir de las pruebas realizadas con modelos construidos con volúmenes de datos menores. En estos modelos se variaron los parámetros de tamaño de los vectores índice entre 10 y 2000 elementos y la cantidad de elementos 1 en los vectores índice de 5, 10 y 15 %. Finalmente se seleccionó el valor de 750 por ser el menor, dado que no hubo mucha variación en los resultados obtenidos al variar este parámetro para los valores superiores a 750. Una vez seleccionado el tamaño de los vectores, se tomó la cantidad de 1 que mejores resultados ofreció.

En los experimentos se empleó un clasificador basado en Centroide. El perfil del usuario se modeló por medio de dos vectores, uno en el cual se adicionan la representación de todos los documentos Relevantes y el otro en el que se adicionan la representación de los No Relevantes. Para determinar si un documento debe seleccionarse para ser mostrado al usuario se calcula la similaridad del mismo con cada uno de los vectores, el clasificador lo selecciona para ser mostrado al usuario si el valor de la similaridad con el centroide construido a partir de los documentos relevantes supera al obtenido con los no relevantes. Si se emplean varios clasificadores, uno por cada nivel, se selecciona un documento para ser mostrado al usuario si los clasificadores por nivel que descartan el documento superan a los que lo seleccionan. En caso de empate, el documento es seleccionado para ser mostrado al usuario con el fin de obtener la retroalimentación.

5.3.1 Resultados

En las Tablas 5.3 se muestran los resultados alcanzados en la tarea del Filtrado Adaptativo al aplicar el Indexado Aleatorio. En la Tabla 5.3a se muestran los resul-

Medida	T11SU		Macro F1	
Niveles	TREC	Reu10	TREC	Reu10
D	0.274	0.219	0.166	0.192
D+P	0.301	0.593	0.234	0.532
D+P+S	0.301	0.625	0.234	0.561
D+P+S+N	**0.372**	**0.650**	**0.272**	**0.615**

(a) Un clasificador con vectores de contexto concatenados.

Medida	T11SU		Macro F1	
Niveles	TREC	Reu10	TREC	Reu10
D	0.274	0.219	0.166	0.192
D+P	0.277	0.654	0.295	0.624
D+P+S	0.345	0.647	0.284	0.594
D+P+S+N	**0.354**	**0.665**	**0.303**	**0.643**

(b) Múltiples clasificadores.

Tabla 5.3: Aplicación del Indexado Aleatorio Multinivel en la tarea de Filtrado Adaptativo

tados alcanzados al concatenar los vectores de contexto de los diferentes niveles en la representación del perfil. Por el contrario, la Tabla 5.3b recoge los resultados alcanzados al emplear un clasificador diferente para cada nivel involucrado.

En las mismas, el nivel D coincide con el valor alcanzado por el modelo de Indexado Aleatorio tradicional. En ambas tablas, los valores asociados a este nivel coinciden. Con texto en negritas se muestran en cada columna el valor máximo alcanzado.

A partir de los resultados mostrados en estas tablas varios elementos pueden ser destacados. Primeramente, podemos notar que en todos los casos, los valores alcanzados al emplear varios niveles de contexto es superior al valor alcanzado al emplear el Indexado Aleatorio en su modelación tradicional.

Igualmente podemos notar que en todos los casos, al emplear niveles de contextos más específicos por lo general se obtienen resultados superiores, obteniéndose los mejores resultados al emplear los 4 niveles de contexto considerados.

Además, debemos notar que los resultados alcanzados al emplear uno o varios cla-

sificadores son similares.

Al analizar los resultados mostrados en las Tablas 5.3a y 5.3b, un elemento que resalta es la marcada diferencia entre los resultados obtenidos por cada una de las colecciones de documentos. Mientras que en Reu10 los resultados obtenidos son superiores a los alcanzados por las relaciones de términos multinivel, en la colección TREC los valores decrecen considerablemente.

Para entender este comportamiento, comenzaremos por analizar los resultados alcanzados en la colección Reu10. El primer aspecto que debemos considerar es el tipo de tópicos que forman esta colección. En ella encontramos solamente temas generales relacionados con temas económicos:

- Previsión de ganancias - Comercio - Maíz - Granos - Crudo (Petróleo)

- Adquisiciones/fusiones - Divisas - Trigo - Navíos - Tasas de Interés

Por otro lado, para la construcción del perfil solamente estaba disponible la consulta que describe el interés, lo que resulta en el mínimo de información para la construcción del perfil. Cuando se tiene tan poca información, la búsqueda de las relaciones falla si los documentos del flujo de información no contienen los términos presentes en la consulta. Sin embargo, con el Indexado Aleatorio no se requiere de la presencia de los términos exactos para que puedan ser seleccionado documentos del flujo que aborden la temática de interés. Esta es la razón por la cual la calidad en la colección Reu10 incluso aumenta en comparación con las relaciones multinivel de términos.

Por el contrario, la colección TREC tiene características diferentes. Muchos de los perfiles en esta colección son de temáticas bastante específicas. Además en ella muchos documentos están etiquetados como pertenecientes a un perfil por la sola mención en su contenido de información relacionada con el perfil, aún cuando el contenido general del mismo no se trate del tema del perfil. El Indexado Aleatorio Multinivel permite solventar el problema de la coincidencia exacta de los términos en los documentos pero

el contenido de los documentos son reducidos a uno o varios vectores, por lo cual las menciones o necesidades puntuales o que involucran determinadas restricciones, por su lugar de origen o la fuente de la información por ejemplo, no pueden ser tratadas adecuadamente con ésta técnica. Por esta razón, los resultados alcanzados en la colección TREC con el Indexado Aleatorio Multinivel son tan discretos.

5.4 Conclusiones parciales

Al emplear el Indexado Aleatorio debemos definir el nivel de contexto que se considera para la construcción del índice, lo cual es una limitación por el hecho de que los términos comparten diferentes relaciones semánticas con otros términos en dependencia del nivel de contexto que se toma en consideración. El MLRI elimina la necesidad de fijar con anterioridad el contexto con el cual se desea trabajar, y posibilita tener en un solo modelo las diferentes relaciones que se establecen entre los términos. Los resultados obtenidos nos muestran la efectividad de la propuesta realizada.

Los experimentos realizados mostraron que el empleo de múltiples niveles de contexto permite obtener una mayor calidad en los resultados del proceso del filtrado. El empleo del MLRI permite solventar el problema de la coincidencia exacta de los términos en los documentos del flujo, hecho que es particularmente importante cuando se parte de tan solo una consulta. Finalmente, el desempeño del MLRI fue efectivo para el filtrado de intereses de información generales, pero no resultó ser igual de efectivo cuando el interés del usuario es un tanto más específico.

Capítulo 6

Combinación de representaciones

En este capítulo se inicia realizando un análisis de las representaciones propuestas. Seguidamente se propone una forma de combinar las mismas. Luego se describen los experimentos realizados y se discuten los resultados obtenidos. Finalmente, se brindan las conclusiones parciales del capítulo.

6.1 Análisis comparativo de las representaciones propuestas

Las relaciones de término multinivel permiten que se puedan establecer restricciones entre las co-ocurrencias de los términos que permiten diferenciar documentos Relevantes de los No Relevantes. Con estas relaciones no solo se puede tomar en consideración que los términos co-ocurran entre los documentos, si no que además se puede especificar en qué contextos deben aparecer estas co-ocurrencias.

Las relaciones de términos multinivel permiten combinar restricciones relacionadas con la aparición de un determinado conjunto de términos para seleccionar un documento a ser mostrado al usuario. Con ellas se puede además establecer cómo deben ubicarse estos términos en la estructura del documento para seleccionar el documento.

Sin embargo, las relaciones de términos no toman en consideración relaciones tales como la sinonimia que existe entre los términos. Por ejemplo, si tenemos la relación:

$$\Phi_D\left(autobús, urbano\right)$$

con ella no podemos seleccionar aquellos documentos que, en vez de emplear el término

autobús, emplean *ómnibus*.

Por otro lado, con el Indexado Aleatorio Multinivel, los términos *autobús* y *ómnibus* estarán fuertemente relacionados, pues sus vectores de contexto serán similares entre sí, basados en el hecho de que estos términos suelen ser empleados en los mismos contextos.

Sin embargo, aún cuando con el Indexado Aleatorio podemos encontrar términos relacionados semánticamente, al emplear este modelo para representar los documentos, no podemos especificar restricciones sobre los términos que en él ocurren. Lo cual es una limitante importante cuando se requiere de reconocer construcciones específicas.

Cada una de estas representaciones tienen sus propias ventajas y limitaciones. Por ello exploramos el uso combinado de ambas representaciones en la construcción del perfil con el fin de maximizar las ventajas obtenidas con ellas.

6.2 Indexado Aleatorio en expansión de información en relaciones

En las relaciones tenemos un conjunto de términos que se encuentran relacionados tomando en consideración uno o varios niveles de contexto. Una vez que tenemos el perfil modelado como un grupo de relaciones que permiten diferenciar los documentos Relevantes de los No Relevantes, cada nuevo documento es seleccionado para ser mostrado al usuario si al menos una de estas relaciones se satisface en él. Durante el proceso para determinar si una relación se satisface, los términos que ella contiene son buscados exactamente en la forma en que se encuentran especificados en la relación. Una posible forma de eliminar esta limitación puede ser no buscar solamente los términos presentes en la relación, sino además considerar aquellos que comparten una mayor relación semántica con ellos. De esta forma no perderíamos la ventaja de poder especificar restricciones sobre los términos en los contextos del documento, sin tener la limitante de

Significado	Términos Relacionados
Empresa dedicada a realizar operaciones financieras con el dinero procedente de sus accionistas y de los depósitos de sus clientes.	banca, dinero, transacciones
Asiento, con respaldo o sin él, en que pueden sentarse dos o más personas.	asiento, grada
En los mares, ríos y lagos navegables, bajo que se prolonga en una gran extensión.	arrecife, bajío, escollo

Tabla 6.1: Algunos significados de *banco*, y posibles términos relacionados.

obligar a que los documentos deban contener exactamente todos los términos presentes en la relación, para que pueda ser seleccionado para ser mostrado al usuario. Sin embargo, debido al fenómeno de la polisemia un mismo término puede tener varios significados diferentes. Por ello, los términos que semánticamente se encuentran relacionados con él pueden no tener relación entre sí. Por ejemplo, tomemos en consideración la palabra *banco*. De acuerdo al diccionario de la Real Academia Española, esta palabra tiene varios significados. En la Tabla 6.1 se muestran tres de ellos.

El empleo del Indexado Aleatorio permite suponer que se está abordando un término sin que deba estar presente el mismo. Por ejemplo, supongamos que contamos con la relación siguiente:

$$\Phi_D\left(transacción, \Phi_N\left(comercio, electrónico\right)\right)$$

relacionada con transacciones realizadas a través del comercio electrónico.

El empleo de un modelo de Indexado Aleatorio posibilitaría disponer del conocimiento necesario como para conocer que los términos: *venta*, *compra*, *pasarela*, *pago*, entre otros, se encuentran relacionados con el término *transacción*, por lo que si identificamos estos términos, puede presumirse que se están abordando de forma implícita sobre las transacciones, aún cuando la misma no se encuentre de forma explícita en un documento. Este enfoque posiblemente permitirá mejorar el recuerdo alcanzado por

las relaciones, aunque esto no signifique necesariamente un aumento en cuanto a la precisión alcanzada por el método.

Nótese que al emplear modelos semántico como el Indexado Aleatorio, se obtienen términos semánticamente relacionados, pero no es posible especificar cuál es el tipo de relación que guardan los términos obtenidos. Por ejemplo, si estamos interesados en *autobuses*, probablemente encontremos términos relacionados a sus componentes, como pueden ser *llantas*, *puertas* y *asientos*. No obstante estos mismos términos los podemos encontrar en otros tipos de vehículos, como pueden ser los *tracto-camiones* y los *vehículos ligeros*.

Existe otro fenómeno que se debe tomar en consideración al intentar expandir la información presente en una relación multinivel de términos para disminuir la limitación de la coincidencia exacta entre sus componentes. El mismo está relacionado con la polisemia de las palabras. Consideremos el caso del término *droga*. Al buscar cuáles son aquellos términos semánticamente relacionados con él, seguramente encontraremos *salud*, *píldora* y *medicamento*, relacionados con uno de los sentidos de esta palabra, pero igualmente encontraremos *tráfico*, *crimen*, *policía* y *cártel*, con otro sentido de la palabra *droga*.

Con la finalidad de disminuir el impacto de este último fenómeno, al intentar buscar en el modelo del MLRI los términos que se encuentran semánticamente relacionados a los especificados en una relación dada, se le adicionan los vectores de contextos de los términos que forman parte de la consulta inicial suministrada por el usuario. Al realizar esta operación, aquellos términos menos relacionados con la temática de interés deberán obtener una semejanza menor y por ende es menos probable que sean seleccionados para expandir la información presente en la relación.

Para la selección de los términos que serán empleados para complementar la información disponible en un relación dada R, primeramente obtenemos una representación de

los términos presentes en ella tomando en cuenta los niveles involucrados. Este proceso se detalla en el Algoritmo 4.

Algoritmo 4: Proceso para enriquecer una relación R a partir de un modelo MLRI.

Entrada:
R: Relación multinivel.
M: Modelo Indexado Aleatorio Multinivel

Salida:
E: Términos a considerar para enriquecer la relación.

```
1 n = Nivel(R) //Nivel de la relación R ;
2 T = Términos(R) //Términos involucrados en R;
3 V = ∑_T M(t_i, n) // Vectores de Contexto de los términos de T en el nivel n;
4 foreach t_k in T do
5     E_{t_k} = {t_i | sem(M(t_i, n), V) ≥ γ_n} ;
6 for x_i in R do
7     if x_i es una relación then
8         Aplicar Algoritmo 4 a x_i ;
```

En el Algoritmo 4, la función *sem* se emplea para comparar los vectores de contexto almacenados en M con la combinación de los vectores de contexto de los términos presentes en R. Los valores γ_n controlan la cantidad de términos que se adicionan a E. El empleo del Algoritmo 4 permite enriquecer las relaciones de términos. Por ejemplo, para la relación $\Phi_D\left(espionaje, industrial\right)$ pudiéramos tener como resultado

$$E_{espionage} = \{espía, agente, secreto, robo, documento, tecnología\}$$

$$E_{industrial} = \{compañía, tecnología, economía, empresa, producción, ventas\}$$

Para el emparejamiento de las relaciones con los documentos del flujo, ahora, además de buscar un término, si éste no se encuentra, se busca que una cantidad predeterminada de aquellos considerados se encuentran en él, si este es el caso se supone que ese término de la relación en el documento se satisface.

6.3 Experimentos

En esta sección se presentan los experimentos realizados para evaluar el comportamiento de los algoritmos de filtrado al emplear la expansión de las relaciones con un modelo de Indexado Aleatorio Multinivel. Igualmente son presentados los resultados alcanzados al comparar los métodos propuestos con algoritmos existentes del estado del arte. En los experimentos se empleó el mismo modelo de Indexado Aleatorio Multinivel usado en los experimentos del capítulo anterior.

Los resultados mostrados en el capítulo son divididos en dos grupos, primeramente son mostrados los resultados alcanzados al combinar en las Relaciones de Términos Multinivel con el uso de las expansiones por medio de un modelo de Indexado Aleatorio Multinivel. Seguidamente son presentadas las comparaciones realizadas entre los métodos aquí presentados con métodos existentes en el estado del arte.

El Indexado Aleatorio basa su funcionamiento en la ocurrencia de elementos semejantes en entornos semejantes. Sin embargo, a medida que empleamos contextos más específicos las probabilidades de co-ocurrencia de los términos en estos contextos disminuye. Por tal motivo no es recomendable emplear un único valor para los valores γ_n. De igual forma, no se debe exigir la misma cantidad de elementos empleados en la exploración en todos los niveles de contexto. Para los experimentos, estos valores se variaron entre 0.1 y 0.5. Finalmente se pudo comprobar que en el nivel N no se puede emplear un valor de γ_N superior a 0.2; para el caso de oraciones y párrafos este parámetro debe ajustarse entre 0.2 y 0.3. En los experimentos reportados se emplearon para estos valores: $\gamma_D = 0.4$, $\gamma_P = 0.25$, $\gamma_S = 0.2$ y $\gamma_N = 0.15$. De igual forma, para buscar los términos en la exploración se consideraron que a nivel D y P se deberían contener 5 términos, 2 para el nivel S y finalmente 1 para el nivel N.

En los experimentos se siguió la misma metodología explicada en la sección 4.8.

Medida	T11SU		Macro F1	
Colecciones	TREC	Reu10	TREC	Reu10
Relaciones	**0.531**	0.558	**0.561**	0.430
MLRI	0.354	**0.665**	0.303	**0.643**
Relaciones + MLRI	0.436	0.65	0.523	0.632

Tabla 6.2: Resultados alcanzados con las diferentes propuestas.

Para las expansiones se empleó el modelo del MLRI construído a partir de la Wikipedia y empleado en la experimentación del capítulo anterior.

6.3.1 Resultados

En la presente subsección se muestran los resultados alcanzados en la tarea del Filtrado Adaptativo al emplear las relaciones de términos combinadas con el Indexado Aleatorio Multinivel y seguidamente la comparativa con métodos existentes del estado del arte.

Evaluación de combinación de representaciones propuestas

En la Tabla 6.2 se presentan los resultados alcanzados con cada una de las principales representaciones exploradas en el presente trabajo.

Como puede notarse en la tabla, si bien hay una mejoría en cuanto a la calidad obtenida con respecto al uso del MLRI, este valor es inferior al emplear la exploración, fundamentalmente en la medida T11SU. Esto se debe principalmente a que en la colección TREC hay varios tópicos compuestos por muy pocos documentos Relevantes a lo largo de todo el flujo de documentos y, dado que al realizar algún tipo de exploración se seleccionan documentos No Relevantes, esto afecta la eficacia en términos de la medida. Con respecto a la colección Reu10, se puede observar un aumento en la calidad al emplear MLRI con respecto a la obtenida con el uso de las relaciones. El objetivo de la

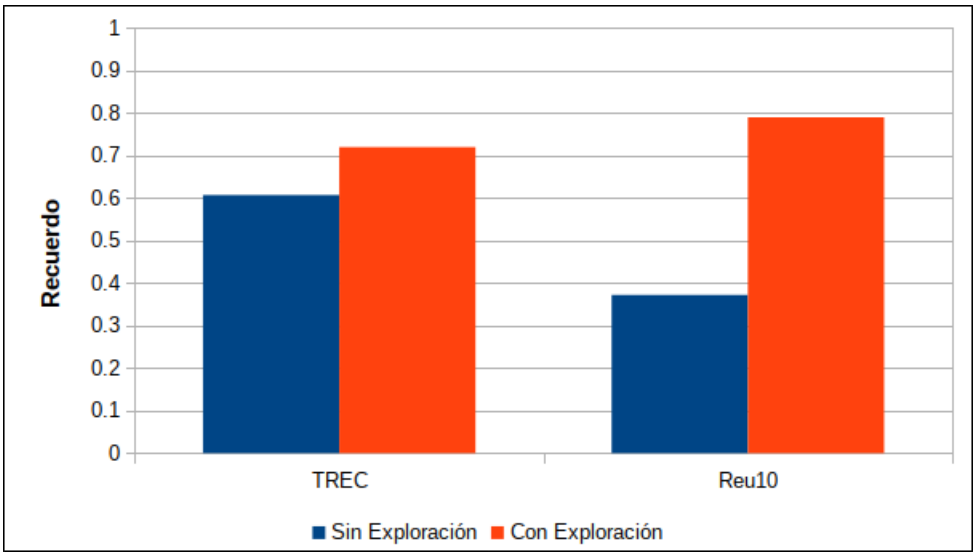


Figura 6.1: Comparación del recuerdo alcanzado al emplear la exploración con MLRI.

exploración es obtener un mayor recuerdo, evitando así que puedan perderse una menor cantidad de documentos positivos, aún cuando esto requiere que sean seleccionados un mayor número de muestras negativas. En la Figura 6.1 se presenta el comportamiento alcanzado en la tarea de filtrado al emplear o no la exploración con el modelo MLRI. Como puede notarse, en ambas colecciones existe un incremento importante en cuanto al valor del recuerdo, siendo en más de un 100 % en el caso de la colección Reu10, lo cual corrobora el hecho de que este tipo de exploración permite que un menor número de documentos Relevantes sean descartados durante el proceso de filtrado, lo cual puede ser muy importante cuando se están procesando documentos sensibles, por ejemplo en un sistema dedicado a filtrar documentos que abordan amenazas terroristas.

Este incremento es más notorio en el caso de la colección Reu10 debido, fundamentalmente, al hecho de que en la modelación realizada en esta colección la cantidad de información disponible para la construcción del perfil es más escasa (solamente se dispone de la consulta). Esto a diferencia de la colección TREC, en la cual se disponía, además de la consulta que define al perfil, con una breve descripción y algunos pocos documentos de muestras.

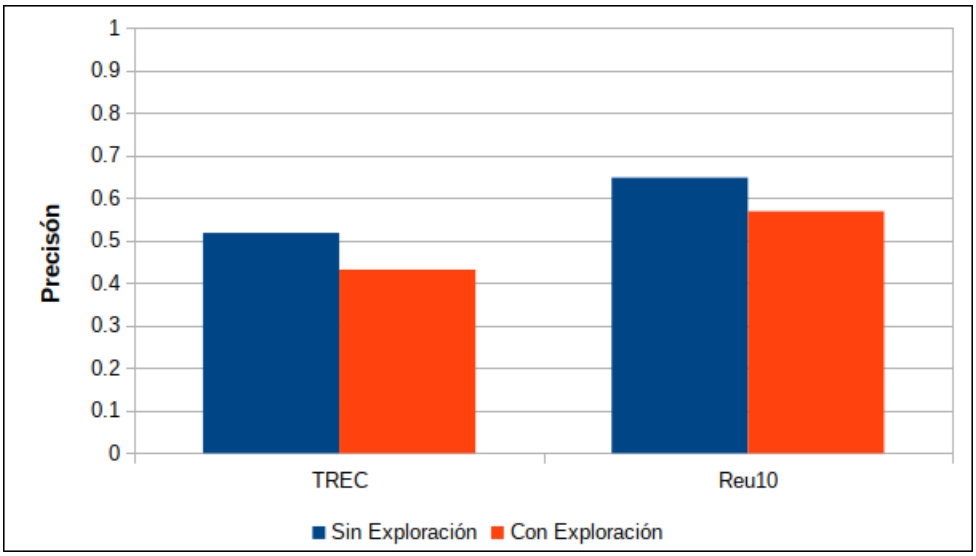


Figura 6.2: Comparación de la precisión alcanzada al emplear la exploración con MLRI.

Como puede observarse en la Figura 6.2, este incremento en el recuerdo al emplear la exploración con MLRI viene unido de una disminución de la precisión alcanzada. Aunque debe notarse que la disminución en términos de precisión es inferior a los valores de incremento en el recuerdo que se alcanza en ambas colecciones. En la Figura 6.3 se presenta el comportamiento del proceso de filtrado en el tiempo. En la gráfica puede notarse que aún cuando al emplear la exploración con MLRI se obtienen un resultado inferior, el comportamiento de ambas gráficas es similar. En ambas se observa una tendencia ascendente.

Comparativa con métodos existentes

En la literatura se han reportado varias estrategias diferentes destinadas a mejorar los resultados alcanzados en la tarea del Filtrado Adaptativo.

En la Tabla 6.3 se muestran los resultados alcanzados por varios de los métodos más recientes reportados en la literatura. Los resultados mostrados fueron extraídos a partir de los valores reportados por los autores en sus trabajos. Dado que no hay homogeneidad

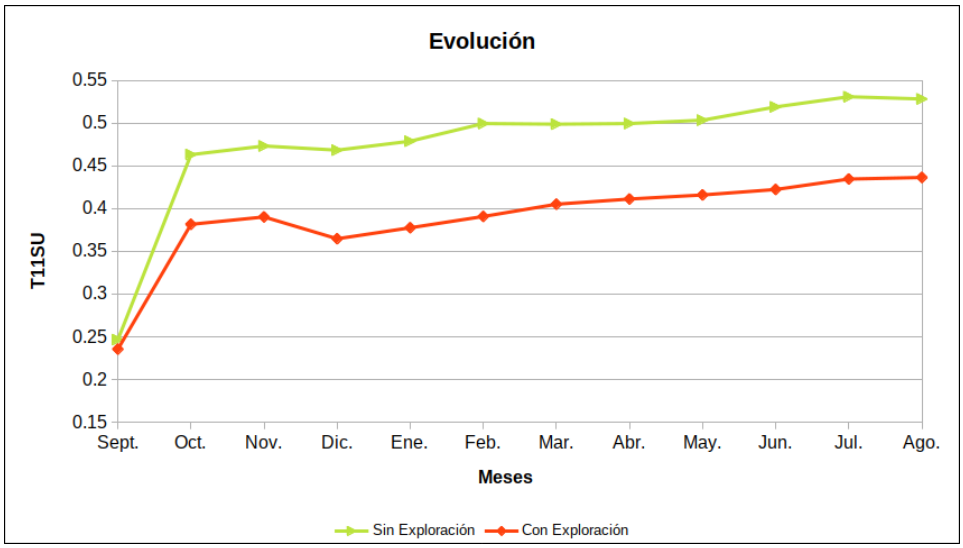


Figura 6.3: Comparación de la evolución en el tiempo al emplear la exploración con MLRI en la colección TREC.

en cuanto a la medida de evaluación empleada para reportar los resultados, en la tabla hay valores que no pudieron ser obtenidos. En tales casos en las celdas correspondientes aparece un guión. Las últimas tres filas de la tabla fueron reservadas para las formas de representación propuestas y validadas en el presente trabajo.

En la tabla, los valores resaltados en negrita se asocian a los valores de calidad más altos, mientras que en cursiva se resaltan los valores alcanzados con nuestras propuestas que son mejores a los reportados en el estado del arte.

Del análisis de los resultados resumidos en la Tabla 6.3 se desprenden algunos de interés. Primeramente, podemos observar que el Indexado Aleatorio obtiene los resultados inferiores en esta colección de documentos. Segundo, aún cuando los valores obtenidos al combinar las relaciones multinivel con el Indexado Aleatorio disminuyeron ligeramente, los resultados alcanzados superan a los reportados en la literatura en la medida *Macro F1*. Además, podemos notar que la calidad obtenida con las relaciones multinivel de términos superan a los reportados hasta el momento en el estado del arte.

Método	Macro F1	T11SU
Zhang et al. [2014]	-	0.5
Li et al. [2012]	0.51	-
Gao et al. [2015]	0.46	-
Li et al. [2015]	0.47	-
Zhang [2004]	-	0.52
Relaciones	***0.559***	***0.531***
MLRI	0.303	0.354
Relaciones + MLRI	*0.523*	0.436

Tabla 6.3: Comparativa de los resultados reportados en trabajos del estado del arte con la colección TREC.

6.4 Conclusiones parciales

El emparejamiento estricto de las relaciones de términos multinivel reduce el recuerdo que es posible ser alcanzado con ellas. Un método para solventar esta limitante fue explorado en el presente capítulo. Como parte de este procedimento de enriquecimiento de las relaciones de términos se puede aumentar el recuerdo que obtiene el clasificador, aunque esto repercute en la calidad final del algoritmo, fundamentalmente en aquellos tópicos compuestos por muy pocos documentos Relevantes en el flujo de Información, o en aquellos donde el interés de información es muy específico.

La comparativa con los métodos actualmente reportados en la literatura permiten mostrar la factibilidad y potencialidad de los métodos propuestos.

Capítulo 7

Conclusiones

Cada día le resulta más difícil a los usuarios mantenerse actualizados con respecto a los temas que les son de interés, debido al enorme volumen de documentos que se genera cada día. Por esta razón, cada día se vuelven más necesarios los sistemas de Filtrado de Información. No obstante, el enfoque estático en esta actividad no se ajusta a las necesidades del día a día, por lo que ganan mayor preponderancia los métodos adaptativos. Sin embargo, estos métodos presentan varios retos que deben ser atendidos con la finalidad de obtener la mayor calidad posible.

Uno de los retos más grandes que deben atenderse cuando se diseña una posible solución para la tarea del filtrado está relacionado con la escasez de información disponible para realizar una correcta modelación del interés de información del usuario.

Otro elemento a tomar en consideración está relacionado con el desbalance que se produce entre los diferentes subtópicos que se presentan entre los documentos que forman el interés del usuario.

Por último, los documentos que son de interés para el usuario no se encuentran homogéneamente distribuidos a lo largo del flujo de documentos. En cada momento pueden surgir nuevos subtópicos. Además, los documentos que abordan éstos se encuentran irregularmente distribuidos a lo largo del flujo de documentos. Por ello, los dos problemas anteriores no se limitan a un período concreto en el desempeño del sistema, sino que se pueden repetir periódicamente a lo largo del flujo de información.

Varios han sido los enfoques dados en la literatura al problema del filtrado de información, y en particular a la construcción y representación del perfil del usuario. Sin embargo, no se ha explorado adecuadamente una representación que tome en conside-

ración la estructura interna de los documentos.

En la presente investigación se propusieron representaciones para el perfil del usuario que toman en consideración esta estructura interna que los humanos damos a los documentos para estructurar de forma lógica el contenido que deseamos plasmar en los mismos. Durante el proceso de construcción de la representación multinivel se propusieron procedimientos para tomar en consideración los principales problemas intrínsecos de la tarea del filtrado adaptativo.

Primeramente, propusimos una extensión de los conjuntos frecuentes, en los cuales se toma además en consideración el nivel de contexto en los cuales los términos suelen coocurrir entre los documentos Relevantes que se emplean para la construcción del perfil del usuario. Se propusieron, además, métodos para tomar en cuenta el problema del desbalance entre los subtópicos. Además se propusieron varias estrategias para combatir el problema del Inicio en Frío en el diseño de los procedimientos.

Seguidamente se propuso una extensión al modelo del Indexado Aleatorio para tomar en consideración, al igual que en las relaciones de términos, las diferentes relaciones que se establecen entre los términos en los diferentes niveles de granularidad que se presentan en los documentos.

Por último, se propuso un método para combinar ambos tipos de representaciones e intentar solventar una de las limitaciones de las relaciones de términos multinivel, y es el caso del emparejamiento estricto entre los términos que forman las relaciones y los que integran los documentos.

Con base en los estudios realizados, y a los resultados obtenidos, podemos concluir que:

- El tomar en consideración diferentes niveles de contextos permite obtener mejores resultados en la tarea del filtrado.

- El empleo de un valor de soporte para seleccionar las relaciones que formarán parte del perfil no es una solución efectiva debido a la escasez de información y la no homogeneidad inherente de la tarea.

- El empleo de recursos externos para enriquecer la información disponible en los documentos es una solución para atacar el problema del Inicio en Frío.

- El uso de un modelo semántico basado en el Indexado Aleatorio es una solución adecuada cuando se tiene que filtrar tópicos de ámbito general, y permite un mejor proceso de modelación cuando se parte de tan poca información como es el caso de una simple consulta.

- El empleo del Indexado Aleatorio Multinivel para enriquecer la información presente en las relaciones de términos multinivel permite aumentar el recuerdo que se puede alcanzar con las mismas.

7.1 Contribuciones

Las constribuciones de la presente investigación son las siguientes:

- Un modelo de representación basado en relaciones de términos multinivel, así como un procedimiento para obtenerlas tomando en consideración el desbalance entre los subtópicos de interés.

- Una propuesta de uso de recursos externos para combatir la escasez de información para la construcción del perfil.

- Un modelo semántico basado en el Indexado Aleatorio en el cual se toma en consideración la estructura interna de los documentos, nombrado como Indexado Aleatorio Multinivel.

- Una propuesta de combinación de las representaciones anteriores para eliminar el problema de la concordancia exacta entre las relaciones y los documentos de interés.

7.2 Trabajo Futuro

Los resultados obtenidos como resultado de la presente investigación abre las puertas a nuevas líneas de investigación. Entre ellas encontramos:

- En las relaciones de términos todos sus componentes son positivos, y por ende, son elementos que deben aparecer en los documentos de interés. Sin embargo, existen situaciones en las cuales no solo deben estar algunos elementos presentes, sino que deben combinarse con la ausencia de otros para que se pueda modelar de forma correcta el interés del usuario.

- El procedimiento de extracción de las relaciones es sensible a la presencia de ruido en los documentos disponibles para la construcción del perfil. En el futuro se deberá trabajar en el desarrollo de una estrategia que posibilite que el algoritmo sea menos sensible al ruido.

- Como parte de los resultados mostrados en el capítulo 5, se pudo apreciar que el emplear un modelo semántico ayuda a solventar el problema del emparejamiento estricto entre las relaciones y los documentos. Sin embargo, los modelos semánticos no permiten encausar el sentido en el que se desea que sean expandidas las relaciones. Una línea de investigación de interés está relacionada con el diseño de otras formas de combinar uno o varios modelos para aumentar el recuerdo que es posible obtener a través de las relaciones, sin que esto implique una disminución excesiva de la precisión del clasificador.

7.3 Publicaciones

Las siguientes publicaciones se derivaron de la presente investigación:

- Fonseca Bruzón, Adrian and López-López, Aurelio and Medina Pagola, José. *Exploring Random Indexing for Profile Learning.* Future and Emergent Trends in Language Technology: First International Workshop, FETLT 2015. LNAI 9577. Quesada, F. José and Martín Mateos, Francisco-Jesús and Lopez-Soto, Teresa(Eds). pp. 77 - 85. 2016. ISBN: 978-3-319-33500-1.

- Adrian Fonseca Bruzón, Aurelio López López, José E. Medina Pagola. *Evaluación de diversas variantes de Indexado Aleatorio aplicadas a la categorización de documentos en el contexto del Aprendizaje en Línea.* Revista Cubana de Ciencias Informática. pp. 162 - 171. Enero, 2016. ISSN: 2227-1899.

- Adrian Fonseca Bruzón, Aurelio López López, José E. Medina Pagola. *Multilevel term analysis for profile learning in adaptive document filtering.* Journal of Intelligent & Fuzzy Systems 34 (5). pp. 3015 - 3026. 2018. IOS Press. ISSN 1875-8967.

Bibliografía

Albakour, M., Macdonald, C., Ounis, I., et al. (2013). On sparsity and drift for effective real-time filtering in microblogs. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pages 419–428. ACM.

Algarni, A., Li, Y., and Xu, Y. (2008). Adaptive information filtering based on ptm model (aptm). In Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on, volume 3, pages 37–40. IEEE.

Allan, J. (1996). Incremental relevance feedback for information filtering. In Proceeding of the Nineteenth Annual International ACM SIGIR. Conference on Research and Development in Information Retrieval, pages 270–278. ACM.

Amine, A., Elberrichi, Z., Simonet, M., Bellatreche, L., and Malki, M. (2009). Som-based clustering of textual documents using wordnet. In Handbook of Research on Text and Web Mining Technologies, pages 189–200. IGI Global.

Ault, T. and Yang, Y. (2000). knn at trec. In Proceeding of the 9th Text REtrieval Conference (TREC-9), pages 127–134.

Ault, T. and Yang, Y. (2001). Knn, rocchio and metrics for information filtering at trec-10. In Proceeding of the Tenth Text REtrieval Conference (TREC-10), pages 84–93. National Institute of Standards and Technology.

Baayen, R. H. (1996). The randomness assumption in word frequency statistics. Research in Humanities Computing, 5:17–31.

Becker, J. and Kuropka, D. (2003). Topic-based vector space model. In Proceedings of the 6th international conference on business information systems, pages 7–12.

Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: Two sides of the same coin? Commun. ACM, 35(12):29–38.

Berardi, G., Ceccarelli, D., Esuli, A., and Marcheggiani, D. (2015). On the impact of entity linking in microblog real-time filtering. In Proceedings of the 30th Annual ACM Symposium on Applied Computing, pages 1066–1071. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Cancedda, N., Cesa-Bianchi, N., Conconi, A., Gentile, C., Goutte, C., Graepel, T., Li, Y., Renders, J. M., Taylor, J. S., and Vinokourov, A. (2003). Kernel methods for document filtering. In Proceeding of the Eleventh Text REtrieval Conference (TREC-11), pages 373–382. National Institute of Standards and Technology.

Changala, R. and Rao, D. R. (2018). Pattern deploying methods for text mining. International Journal of Soft Computing, 13(2):61–68.

Cohen, T., Schvaneveldt, R., and Widdows, D. (2010). Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of biomedical informatics, 43(2):240–256.

Collins-Thompson, K., Ogilvie, P., Zhang, Y., and Callan, J. (2002). Information filtering, novelty detection, and named-page finding. In Proceeding of the Eleventh Text REtrieval Conference (TREC-11), pages 107–118.

Cossu, J.-V., Bonnefoy, L., Bost, X., and Bèze, M. E. (2015). How to merge three different methods for information filtering? arXiv preprint arXiv:1510.07385.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391.

Fan, F., Feng, Y., Yao, L., and Zhao, D. (2016). Adaptive evolutionary filtering in real-time twitter stream. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pages 1079–1088. ACM.

Figuerola, C. G., Berrocal, J. L. A., Rodríguez, A. F. Z., Rodríguez, E., and Reina, G. (2004). Algunas técnicas de clasificación automática de documentos. Cuadernos de documentación multimedia, ISSN-e, pages 1575–9733.

Gao, Y., Xu, Y., and Li, Y. (2015). Pattern-based topics for document modelling in information filtering. IEEE Transactions on Knowledge and Data Engineering, 27(6):1629–1642.

Han, Z., Yang, M., Kong, L., Qi, H., and Li, S. (2016). A hybrid model for microblog real-time filtering. Chinese Journal of Electronics, 25(3):432–440.

Hanani, U., Shapira, B., and Shoval, P. (2001a). Information filtering: Overview of issues, research and systems. User modeling and user-adapted interaction, 11(3):203–259.

Hanani, U., Shapira, B., and Shoval, P. (2001b). Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction, 11(3):203–259.

Harish, B. S., Guru, D. S., and Manjunath, S. (2010). Representation and classification of text documents: A brief review. IJCA, Special Issue on RTIPPR (2), pages 110–119.

Hassel, M. and Sjöbergh, J. (2006). Towards holistic summarization: Selecting summaries, not sentences. Proceedings of Language Resources and Evaluation.

Higgins, D. and Burstein, J. (2007). Sentence similarity measures for essay coherence. In Proceedings of the 7th International Workshop on Computational Semantics, pages 1–12.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pages 289–296. Morgan Kaufmann Publishers Inc.

Hogan, G. (2012). Building Better Essays. Cengage Learning.

Horváth, T. and de Carvalho, A. C. (2017). Evolutionary computing in recommender systems: a review of recent research. Natural Computing, 16(3):441–462.

Jones, G. J. and Brown, P. J. (2003). Context-aware retrieval for ubiquitous computing environments. In Workshop on Mobile and Ubiquitous Information Access, pages 227–243. Springer.

Jonnalagadda, S., Leaman, R., Cohen, T., and Gonzalez, G. (2010). A distributional semantics approach to simultaneous recognition of multiple classes of named entities. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 224–235. Springer.

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. Natural Language Engineering, 2(1):15–59.

Kejriwal, M. and Szekely, P. (2017). Information extraction in illicit web domains. In Proceedings of the 26th International Conference on World Wide Web, pages 997–1006. International World Wide Web Conferences Steering Committee.

Lanquillon, C. and Renz, I. (1999). Adaptive information filtering: Detecting changes in text streams. In Proceedings of the eighth international conference on Information and knowledge management, pages 538–544. ACM.

Lee, Y., Nam, K. W., and Ryu, K. H. (2017). Fast mining of spatial frequent wordset from social database. Spatial Information Research, 25(2):271–280.

Lewis, D. D., Yang, Y., Rose, T. G., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research, 5(Apr):361–397.

Li, Y., Algarni, A., Albathan, M., Shen, Y., and Bijaksana, M. A. (2015). Relevance feature discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, 27(6):1656–1669.

Li, Y., Algarni, A., Wu, S.-T., and Xue, Y. (2009). Mining negative relevance feedback for information filtering. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, pages 606–613. IEEE Computer Society.

Li, Y., Algarni, A., and Xu, Y. (2011). A pattern mining approach for information filtering systems. Information Retrieval, 14(3):237–256.

Li, Y., Zhou, X., Bruza, P., Xu, Y., and Lau, R. Y. (2012). A two-stage decision model for information filtering. Decision Support Systems, 52(3):706–716.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine learning, 2(4):285–318.

McNamee, P., Piatko, C., and Mayfield, J. (2002). Jhu/apl at trec 2002: Experiments in filtering and arabic retrieval. In Proceeding of the Eleventh Text REtrieval Conference (TREC-11), pages 358–363. National Institute of Standards and Technology.

Mikolov, T., Chen, K., Corrado, G., , and Dean, J. (2013). Efficient estimation of word representations in vector space. In ICLR Workshop.

Moen, H., Marsi, E., and Gambäck, B. (2013). Towards dynamic word sense discrimination with random indexing. In Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, pages 83–90.

Mohd Azmi, N. F., Sam, S. M., Adli, S., and Sjarif, N. N. A. (2017). Adaptive user profiling: Experimenting on multiple interests and changing interests. International Journal of Advances in Soft Computing & Its Applications, 9(1).

Montejo-Ráez, A., Perea-Ortega, J. M., Díaz-Galiano, M. C., and Ureña-López, L. A. (2010). Experiments with google news for filtering newswire articles. In Peters, C., Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., and Roda, G., editors, Multilingual Information Access Evaluation I. Text Retrieval Experiments, pages 381–384, Berlin, Heidelberg. Springer Berlin Heidelberg.

Musto, C. (2010). Enhanced vector space models for content-based recommender systems. In Proceedings of the fourth ACM conference on Recommender systems, pages 361–364. ACM.

Nanas, N. and de Roeck, A. (2010). A review of evolutionary and immune-inspired information filtering. Natural Computing, 9(3):545–573.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In Proceedings of the Language Resources and Evaluation Conference (LREC2012).

Qamar, A. M., Gaussier, E., and Denos, N. (2010). Batch document filtering using nearest neighbor algorithm. In Peters, C., Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., and Roda, G., editors, Multilingual Information Access Evaluation I. Text Retrieval Experiments, pages 354–361, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rahmatizadeh Zagheli, H., Zamani, H., and Shakery, A. (2017). A semantic-aware profile updating model for text recommendation. In Proceedings of the Eleventh ACM Conference on Recommender Systems, pages 316–320. ACM.

Rekabsaz, N., Bierig, R., Lupu, M., and Hanbury, A. (2017). Toward optimized multimodal concept indexing. In Transactions on Computational Collective Intelligence XXVI, pages 144–161. Springer.

Ricci, F., Rokach, L., and Shapira, B. (2015). Recommender Systems: Introduction and Challenges, pages 1–34. Springer US, Boston, MA.

Robertson, S. E., Walker, S., Beaulieu, M., Gatford, M., and Payne, A. (1996). Okapi at trec-4. Nist Special Publication Sp, pages 73–96.

Sahlgren, M. (2005). An introduction to random indexing. In Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, (TKE).

Sahlgren, M., Hansen, P., and Karlgren, J. (2002). English-japanese cross-lingual query expansion using random indexing of aligned bilingual text data. In The Philosophical Writings of Gottlob Frege. Citeseer.

Sahlgren, M. and Karlgren, J. (2001). Vector-based semantic analysis using random indexing for cross-lingual query expansion. In Workshop of the Cross-Language Evaluation Forum for European Languages, pages 169–176. Springer.

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley.

Sharma, V. (2018). To study and analyze information filtering approaches for emergency responders during crisis management. PM World Journal, VII(6):1–10.

Soboroff, I., Ounis, I., Macdonald, C., and Lin, J. J. (2012). Overview of the trec-2012 microblog track. In TREC, volume 2012, page 20.

Srikanth, M., Wu, X., and Srihari, R. (2002). Ub at trec-11: Batch and adaptive filtering. In Proceeding of the Eleventh Text REtrieval Conference (TREC-11), pages 557–563.

Steinberger, R., Pouliquen, B., Kabadjov, M., and Van der Goot, E. (2011). Jrc-names: A freely available, highly multilingual named entity resource. In Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (RANLP).

Sun, X. and Zhou, H. (2011). An empirical comparison of two boosting algorithms on real data sets based on analysis of scientific materials. In Advances in Computer Science, Intelligent System and Environment, pages 327–331. Springer.

Tan, J., Wan, X., Liu, H., and Xiao, J. (2018). Quoterec: Toward quote recommendation for writing. ACM Transactions on Information Systems (TOIS), 36(3):34.

Ventura, P. Ú. and Marañón, M. A. O. (2013). Español escrito: idea y redacción: Propuestas para el taller de escritura, volume 5. Univ Pontifica Comillas.

Wai, T. T. and Aung, S. S. (2017). Enhanced frequent itemsets based on topic modeling in information filtering. In Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on, pages 155–160. IEEE.

Wu, L., Huang, X., and Niu, J. (2002). Fdu at trec 2002: Filtering, q&a, web and video tasks. In Proceeding of the Eleventh Text REtrieval Conference (TREC-11), pages 232–247.

Wu, S.-T., Li, Y., Xu, Y., Pham, B., and Chen, P. (2004). Automatic pattern-taxonomy extraction for web mining. In Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, pages 242–248. IEEE Computer Society.

Xu, H., Yang, Z., Wang, B., Liu, B., Cheng, J., Liu, Y., Yang, Z., Cheng, X., and Bai, S. (2002). Trec-11 experiments at cas-ict: Filtering and web. In Proceeding of the Eleventh Text REtrieval Conference (TREC-11), pages 141–151.

Zagheli, H. R., Ariannezhad, M., and Shakery, A. (2017a). Negative feedback in the language modeling framework for text recommendation. In European Conference on Information Retrieval, pages 662–668. Springer.

Zagheli, H. R., Ariannezhad, M., and Shakery, A. (2017b). Negative feedback in the language modeling framework for text recommendation. In Jose, J. M., Hauff, C., Altıngovde, I. S., Song, D., Albakour, D., Watt, S., and Tait, J., editors, Advances in Information Retrieval, pages 662–668, Cham. Springer International Publishing.

Zamani, H. and Shakery, A. (2018). A language model-based framework for multi-publisher content-based recommender systems. Information Retrieval Journal, pages 1–41.

Zhai, C., Jansen, P., Stoica, E., Grot, N., and Evans, D. A. (1998). Threshold calibration in clarit adaptive filtering. In Proceeding of the Seventh Text REtrieval Conference (TREC-7), pages 149–157.

Zhang, L. and Zhang, Y. (2010). Interactive retrieval based on faceted feedback. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, pages 363–370, New York, NY, USA. ACM.

Zhang, L. and Zhang, Y. (2014). Hierarchical bayesian models with factorization for content-based recommendation. arXiv preprint arXiv:1412.8118.

Zhang, L., Zhang, Y., and Xing, Q. (2014). Learning from labeled features for document filtering. CoRR, abs/1412.8125.

Zhang, Y. (2004). Using bayesian priors to combine classifiers for adaptive filtering. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04, pages 345–352, New York, NY, USA. ACM.

Zhang, Y. (2009). Adaptive information filtering. In Text Mining: Classification, Clustering, and Applications, pages 215–242. Chapman and Hall/CRC.

Zhang, Y. and Callan, J. (2000). Yfilter at trec-9. In Proceeding of the Ninth Text REtrieval Conference (TREC-9), pages 135–140.

Zhang, Y. and Callan, J. (2001). The bias problem and language models in adaptive filtering. In Proceeding of the Tenth Text REtrieval Conference (TREC-10), pages 78–83.

Zhong, N., Li, Y., and Wu, S. (2012). Effective pattern discovery for text mining. IEEE Transactions on Knowledge and Data Engineering, 24(1):30–44.