Método Semisupervisado para Clasificación de Documentos Usando Resúmenes Automáticos

por

Emmanuel Anguiano Hernández

Tesis sometida como requisito parcial para obtener el grado de

Maestro en Ciencias en el Área de Ciencias Computacionales en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por:

Dr. Luis Villaseñor Pineda, INAOE

©INAOE 2010
El autor otorga al INAOE el permiso de reproducir y distribuir copias en su totalidad o en partes de esta tesis

Método Semisupervisado para Clasificación de Documentos Usando Resúmenes Automáticos

Tesis de Maestría

Por:

Emmanuel Anguiano Hernández

Asesor:

Dr. Luis Villaseñor Pineda

Instituto Nacional de Astrofísica Óptica y Electrónica
Coordinación de Ciencias Computacionales

Tonantzintla, Puebla. Diciembre 2010

Para Aida y Felipe, mis padres.
Quienes me dieron el primero y más grande conocimiento.

Agradecimientos

Quisiera agradecer a todas las personas que de un modo u otro estuvieron cerca de mí durante los dos años que duró la preparación de este trabajo, estoy seguro de haber aprendido mucho de cada una de ellas. A todos, gracias por apoyarme, aconsejarme, acompañarme y compartir.

Un agradecimiento muy especial a los doctores Luis Villaseñor Pineda y Manuel Montes y Gómez quienes me guiaron y motivaron durante el proceso de realización de esta investigación y que siempre estuvieron dispuestos a apoyar resolviendo dudas, proporcionando material o compartiendo un comentario. Más que unos maestros han sido unos verdaderos amigos.

Al comité que revisó este trabajo: Dr. Francisco Martínez Trinidad, Dr. Eduardo Morales Manzanares y Dr. Saúl Pomares Hernández, sus comentarios me ayudaron a aclarar muchas de las ideas involucradas en este trabajo y a profundizar en el conocimiento sobre el mismo.

A los compañeros de ONCEGEN, con quienes siempre pude compartir una charla, aclarar dudas y pasar un buen rato. Particularmente a los compañeros del Laboratorio de Tecnologías del Lenguaje: Gaby, Omar y Fernando, cuya dedicación y esfuerzo han sido inspiradores.

Al CONACyT por el apoyo brindado mediante la beca 271106 para realizar mis estudios de maestría.

A mis padres Aida y Felipe y mi hermano Ricardo, por su apoyo incondicional sin el cuál nada de esto habría sido posible. A ellos mi más grande agradecimiento.

Finalmente quisiera agradecer a quienes, aunque no intervinieron directamente en la realización de este trabajo, compartieron conmigo esta etapa: a David y Adán por las grandes charlas, a Ivanto y Oliver por la música, a Ligia por los consejos, a la Guerrillera, a los Robotos, a Tell, a Daniel, a Dulce... Gracias.

E.

Resumen

La gran cantidad de información disponible en forma de documentos de texto demanda un ordenamiento para ser accesible. La Clasificación de Texto se encarga de separar conjuntos de documentos en categorías predefinidas según sus características. Los algoritmos que cumplen esta tarea se denominan Clasificadores y existe una amplia variedad de ellos con características distintivas y diferentes niveles de desempeño para condiciones diversas. Una dificultad de este tipo de algoritmos es que requieren de grandes cantidades de información en su entrenamiento para producir un buen clasificador. La información que se les proporciona suelen ser documentos clasificados manualmente. Obtener estos documentos es costoso pues requiere que expertos humanos asignen la categoría correspondiente en el dominio del problema.

Para solucionar este problema se han desarrollado métodos semisupervisados que emplean un conjunto pequeño de documentos etiquetados manualmente más un conjunto numeroso de documentos no etiquetados para entrenarse. Debido a que los documentos no etiquetados pueden contener información ruidosa que interfiera en el entrenamiento del clasificador, es necesario un método que permita filtrar la información adecuada y retirar el ruido.

Utilizando resúmenes automáticos para separar la información relevante de los documentos, se desarrolló un método para clasificación basado en *self-training* que mejora el desempeño del clasificador con respecto al caso base en el que no se emplean resúmenes. Los resultados experimentales demuestran que el método es efectivo, que emplea un número pequeño de iteraciones y pocos documentos etiquetados.

En este trabajo se muestra el respaldo teórico del algoritmo propuesto, se hace una revisión de trabajos relacionados, se exponen los experimentos previos a partir de los cuales se obtuvo información que determinó las condiciones del sistema final, así como los resultados de los experimentos sobre diferentes conjuntos de datos con características diversas.

Abstract

The vast amount of information avaible as text documents demands an ordenation in order to keep it accesible. Text Categorization is the task of separate document sets into predefined categories in function of their characteristics. Algorithms wich performs this task are called Classifiers, and there are a wide variety of them with distinctive features and different levels of performance for almost conditions. A dificulty with this type of algorithms is that they require large amounts of information in their training phase to produce good enought classifiers and training documents use to be manually classified. Obtaining training documents is expensive because it requires human experts to perform the classifications.

In order to solve this problem, semisupervised methods have been developed. It uses just a small set of manually labeled documents in addition with a large pool of unlabeled documents for training. Due to unlabeled documents may contain noisy information that interferes whit the training of the classifier, we need a method to filter the right infomation and remove the noise.

Using automatic summarization to remove noise and keep useful information from documents, we developed a self-training based method for document categorization which performs better when compared with traditional scheme without summaries. Experimental results shows that the method is efective, use few labeled and unlabeled documents and few iterations.

In this work we show the theoretical background of the proposed algorithm, a review of related work and the previous experiments. From these experiments we get the conditions for the final system. Finally we expose the definitive system and its results over many different corpora.

Tabla de Contenido

Capítulo 1
Introducción

Clasificar significa *dividir u ordenar por clases*. *Dividir* es sinónimo de *partir*, también de *distribuir*, que a su vez significa *designar y entregar a cada uno lo que le corresponde según voluntad, conveniencia, regla o derecho*. *Ordenar* por su parte significa *colocar de acuerdo con un plan o de modo conveniente*. Así, clasificar no es un acto que se realiza arbitrariamente sino que involucra y refleja el contexto cultural de quién clasifica, su voluntad, conveniencia y reglas a las que está sometido. Clasificar es una actividad que puede involucrar una complejidad bastante alta.

Sin embargo, todos los días clasificamos cosas de forma consciente o inconsciente: los objetos que colocamos en los distintos compartimentos de una mochila, las monedas con las que pagamos el transporte público, el material de lectura que debe ser revisado para cumplir con cierta tarea; y todas esas decisiones y clasificaciones tienen como objetivo disponer de las cosas de un modo conveniente para mantener el orden.

Cuando se dispone de una gran cantidad de información sobre una gran variedad de temas y se desea acceder a información sobre un tema específico es necesario que la información se encuentre ordenada a fin de que no sea necesaria una búsqueda exhaustiva que implique una gran cantidad de tiempo y esfuerzo. La condición deseable sería que esta basta colección de información se encontrara ordenada según algún criterio de manera que localizar la información deseada no representara un problema mayor que el moverse entre las diferentes categorías o clases que forman la colección. Esta es la importancia de clasificar información.

La investigación para descubrir y desarrollar métodos que permitan clasificar información en forma de documentos de texto se llama Clasificación de Texto (TC, *Text Categorization*) y es una rama del área de Ciencias Computacionales conocida como Procesamiento del Lenguaje Natural (NLP). Los algoritmos de aprendizaje automático empleados para realizar esta tarea se llaman clasificadores y existe una

gran variedad de ellos con características, ventajas, condiciones de uso y capacidades propias.

De forma análoga a la manera como las personas clasifican o toman decisiones de acuerdo con su contexto cultural, voluntad, conveniencia y reglas a las que están sometidas, un clasificador requiere de información que le permita decidir correctamente la categoría a la que corresponde un documento que debe ser clasificado de entre las disponibles en un conjunto predefinido. Esta información suele encontrarse en forma de ejemplos que se proporcionan al clasificador, a partir de los cuales el programa debe reconocer las características que determinan la pertenencia de un documento a una clase determinada. El proceso en el que se proporciona información para que un clasificador aprenda a clasificar se llama entrenamiento y el conjunto de ejemplos de los cuales debe aprender se llama conjunto de entrenamiento. La etapa en la que se evalúa al clasificador se llama prueba y un conjunto de documentos cuyas clases se conocen y sobre los cuales se mide el desempeño del clasificador se llama conjunto de prueba.

En general conforme mayor y de mejor calidad es la información que se proporciona al clasificador en la etapa de entrenamiento mejor será su desempeño. Sin embargo, obtener conjuntos de documentos que puedan usarse para entrenar clasificadores requiere que especialistas humanos inviertan una gran cantidad de tiempo y esfuerzo en asignar a los documentos etiquetas con la categoría correcta. Este procedimiento es muy costoso y para evitarlo se han desarrollado estrategias que permiten utilizar como conjunto de entrenamiento un número reducido de documentos etiquetados junto a una cantidad adicional (usualmente grande) de documentos cuyas etiquetas se desconocen.

Semisupervisados es el nombre de estos enfoques que son un punto intermedio entre la clasificación supervisada en que el clasificador aprende de un conjunto de ejemplos y la clasificación no supervisada, comúnmente llamada *Clustering* o agrupamiento debido a que el resultado de este tipo de procesos es una separación del conjunto de documentos que se desea agrupar en subconjuntos que comparten características. En un proceso de agrupamiento no hay clases o categorías predefinidas.

La idea general de los sistemas de clasificación semisupervisada es utilizar la información disponible en los documentos etiquetados para entrenar un clasificador inicial. Luego utilizar este clasificador para etiquetar los documentos de entrenamiento que no están etiquetados y utilizar la información que de ellos pueda obtenerse para mejorar

al clasificador inicial. Existen diversos algoritmos que permiten obtener información del conjunto de documentos no etiquetado, en la sección del marco teórico se describen los más usuales. Entre ellos, *self-training* supone una plataforma adecuada para los experimentos aquí propuestos. *Self-training* emplea un criterio de selección para elegir entre los documentos no etiquetados que han sido categorizados por el clasificador inicial a aquellos en los que se tiene una mayor *confianza*, retira estos documentos del conjunto no etiquetado y los agrega al conjunto etiquetado con las clases que el clasificador inicial les ha asignado. Una vez que se tiene un nuevo conjunto etiquetado se entrena una nueva versión del clasificador que, por tener una mayor cantidad de información, deberá tener un mejor desempeño. El proceso es iterativo y se repite hasta que se cumple alguna condición de paro.

Otra característica relevante de *self-training* es que asigna un índice de confianza a los documentos clasificados en función de la clasificación realizada, lo que a su vez depende de la calidad del clasificador, de modo que si un documento de cierta clase es categorizado en otra y luego elegido para formar parte del conjunto etiquetado, el error se propagará en las iteraciones siguientes. Del mismo modo, un documento que está correctamente etiquetado y es elegido para integrarse al conjunto etiquetado puede contener suficiente información no relevante para su categoría de forma que desvíe al clasificador en la siguiente iteración y propague el error.

Con el objetivo de filtrar la información relevante de cada documento con ello mejorar el desempeño del clasificador, se propone reemplazar los documentos con sus resúmenes automáticos en las diferentes etapas del sistema. Una serie de experimentos determinará el impacto de los resúmenes en las fases de entrenamiento y prueba con la finalidad de conocer el esquema adecuado para integrar los resúmenes en el sistema semisupervisado.

Los objetivos formales del trabajo se enuncian en la siguiente sección.

1.1. Objetivos

1.1.1. Objetivo General

- Desarrollar, implementar y evaluar un método de clasificación semisupervisada para documentos de texto incorporando resúmenes automáticos.

1.1.2. Objetivos Específicos

- Seleccionar a partir de la investigación bibliográfica, un método confiable para elaborar resúmenes automáticos.
- Determinar experimentalmente la forma de integrar los resúmenes automáticos en la arquitectura del sistema semisupervisado para que su impacto sea positivo.
- Implementar un método para seleccionar los documentos *confiables* que se integrarán al conjunto etiquetado en cada iteración.
- Desarrollar un sistema que implemente el método semisupervisado que incorpora resúmenes automáticos y comparar su desempeño contra el caso base, es decir cuando no se utilizan resúmenes.

1.2. Organización del Documento

Este trabajo está organizado de la siguiente manera: En el capítulo 2 se presenta el soporte teórico necesario para comprender el contenido de esta tesis. Se describe el problema de la clasificación de documentos, los esquemas de representación de documentos, selección de atributos y algunos métodos de clasificación comúnmente empleados. También se presenta la idea general de los sistemas de clasificación semisupervisada, se comentan algunos de los enfoques utilizados y se describen sus algoritmos generales. Finalmente se presentan los conceptos y enfoques comúnmente empleados en Elaboración Automática de Resúmenes (TS, *Text Summarization*).

El capítulo 3 contiene un análisis de los trabajos relacionados en los que se emplean resúmenes automáticos en clasificación de documentos, como técnica de clasificación, como estrategia para reducir el espacio de representación de los documentos y para mejorar la clasificación. También se revisan algunos trabajos donde se presentan enfoques semisupervisados que no incluyen el uso de resúmenes.

Una serie de experimentos previos a la implementación del algoritmo propuesto que tienen como objetivo ampliar el conocimiento referente al impacto de los resúmenes automáticos en clasificación supervisada es desarrollada en el capítulo 4. Los experimentos, resultados y conclusiones son discutidos y usados para determinar las características que el sistema semisupervisado deberá tener.

En el capítulo 5 se presenta el método propuesto, se describe cada uno de sus módulos y su funcionamiento general. Se realizan experimentos sobre 4 conjuntos de datos de naturaleza distinta para evaluar su alcance en diferentes condiciones. Experimentos con distintos parámetros permiten obtener una serie de conclusiones discutidas al finalizar el capítulo.

Finalmente, el capítulo 6 contienen una discusión adicional sobre el comportamiento del sistema, las conclusiones derivadas del trabajo y algunas perspectivas para trabajo futuro.

Capítulo 2

Marco Teórico

2.1. Clasificación Automática de Texto

La Clasificación Automática de Textos consiste en asignar documentos, de acuerdo con sus características, a un conjunto de categorías o clases previamente definidas. Formalmente se emplean algoritmos de aprendizaje automático para aproximarse a función objetivo $\Phi : D \times C \rightarrow \{T, F\}$ que describe cuál debe ser la correspondencia entre el conjunto de documentos $D = \{d_1, d_2, ..., d_{|D|}\}$ y el de categorías $C = \{c_1, c_2, ..., c_{|C|}\}$ a las que pueden asignarse, mediante una función $\hat{\Phi} : D \times C \rightarrow \{T, F\}$ denominada clasificador. Si la evaluación del clasificador para un documento d_1 y una clase c_j $\Phi(d_i, c_j) = T$ se dice que d_i pertenece a la clase c_j. Si $\Phi(d_i, c_j) = F$ entonces d_i no pertenece a c_j. Dependiendo del número de categorías a las que pueda asignarse un documento, la clasificación puede ser monoclase (cada documento se puede asignar solamente a una categoría) o multiclase (un documento puede pertenecer a más de una categoría) aunque esta última suele abordarse como $|C|$ diferentes problemas de clasificación binaria en los que cada documento de D puede o no pertenecer a cada una de las categorías de C [33].

Se dice que un algoritmo de clasificación es supervisado si emplea información de un conjunto de documentos D_{tr} (conocido como conjunto de entrenamiento) cuyas categorías se conocen para identificar las relaciones entre D y C que luego empleará para etiquetar los documentos de un conjunto de prueba D_{te} que además cumple $D_{tr} \cap D_{te} = 0$. Si no se emplea un conjunto de entrenamiento o las categorías del conjunto se desconocen la clasificación es no supervisada. Finalmente si en el conjunto de entrenamiento está formado por documentos cuyas clases se conocen juntos con documentos cuya clase se desconoce se trata de clasificación semisupervisada.

2.1.1. Representación de Documentos

Para que los documentos del conjunto D puedan ser procesados por un algoritmo de aprendizaje automático deben llevarse a una representación adecuada, este procedimiento se conoce como *indexado* de documentos y es usado en otras tareas de procesamiento de lenguaje natural como Recuperación de Información (IR, *Information Retrieval*) [34].

Como paso previo al indexado de los documentos, es necesario preprocesarlos eliminando toda la información no útil que depende de la naturaleza de los documentos y puede consistir en etiquetas de metatexto, comentarios, símbolos y caracteres no alfabéticos. También es necesario sustituir letras mayúsculas y en muchas ocasiones lematizar los términos sustituyéndolos por su lema o raíz eliminando sufijos. Adicionalmente suelen retirarse de los documentos palabras neutrales o vacías que no aportan información sobre su naturaleza como los artículos, preposiciones y conjunciones.

El Indexado consiste en representar cada documento d_i como un vector de características o atributos $d_i = \langle t_{1i}, t_{2i}, ...t_{|T|i} \rangle$ donde cada atributo t representa uno de los $|T|$ términos o palabras que aparecen en la colección de documentos D_{tr} usada para entrenar al clasificador. Al conjunto de términos T se le conoce como vocabulario de la colección. El peso t_{ki} asignado a cada uno de los términos en un documento puede determinarse de varias formas incluyendo en él diferentes tipos de información acerca de los documentos. Los esquemas de pesado más usados en clasificación de textos incluyen los siguientes:

- Binario: el atributo t_k toma el valor de 1 si el término correspondiente aparece en el documento d_i y 0 en caso de no aparecer.

- Frecuencia del Término (TF): el atributo t_k toma el valor correspondiente al conteo de apariciones del término en el documento d_i.

- $TFIDF$: además de información descriptiva sobre el documento, este esquema de pesado incluye información sobre la frecuencia de aparición del término en los otros documentos de la colección con la finalidad de penalizar a aquellos que aparecen en muchos documentos (no aportan mucha información) y premiar aquellos que solo aparecen en pocos y son por tanto más informativos.

Cuando el conjunto de documentos ha sido indexado se tiene una matriz cuyos elementos representan el peso de cada término en cada uno de los documentos. Debido a que el tamaño del vocabulario de una colección de documentos suele ser muy grande (miles, decenas o centenas de miles) se emplean estrategias de selección de atributos para reducir la dimensionalidad de los vectores que los representan, lo que permite remover atributos que no aportan información o seleccionar a aquellos con características relevantes para la clasificación como una alta capacidad discriminativa entre las diferentes clases del conjunto. Diferentes estrategias de selección de atributos como ganancia de información, frecuencia en documentos y otros métodos basados en información estadística se han usado con buenos resultados.

2.1.2. Selección de Atributos

La selección de atributos consiste en reducir la cantidad de atributos usados para representar a los documentos de D desechando aquellos que no proporcionan información útil sobre la correspondencia entre los documentos y las clases. Para conocer cuáles atributos aportan información y cuáles no (o lo hacen en menor medida) se asigna un índice a cada uno de ellos. Este índice proporciona una medida de la capacidad informativa, discriminativa o la importancia del atributo para hacer una buena clasificación. Generalmente, una vez que se conoce el índice de importancia de cada uno de los atributos, se eligen aquellos con un valor por encima de cierto umbral y se descarta a todos los demás [45].

Existen varias estrategias de selección de atributos empleadas a menudo en clasificación automática de textos [44], la mayoría basados en información estadística del conjunto de entrenamiento, los siguientes representan algunos de los más comunes:

- Frecuencia en Documentos (DF): El índice asignado a cada término corresponde al conteo de documentos del conjunto de entrenamiento en el que dicho término aparece al menos una vez. Se elige a los términos que aparecen en más de n documentos para representarlos y el resto se descarta.

- Ganancia de Información (IG): El índice de IG es una medida de la información que se obtiene acerca de la categoría a la que pertenece un documento debido a la ausencia o presencia de un atributo determinado. Se calcula con la ecuación siguiente en la que $P(c_j)$ es la probabilidad de que el documento pertenezca a la clase c_j, $P(t_k)$ es la probabilidad de que el termino t_k aparezca en el documento,

$P(c_j|t_k)$ es la probabilidad de que el documento pertenezca a la clase c_j dado que contiene al término t_k y $P(c_j|\bar{t_k})$ es la probabilidad de que pertenezca a la clase pero no aparezca el término t_k.

$$IG(t) = -\sum_{j=1}^{|C|} P(c_j) log P(c_j) + P(t_k) \sum_{j=1}^{|C|} P(c_j|t_k) log P(c_j|t_k)$$

$$+P(\bar{t_k}) \sum_{j=1}^{|C|} P(c_j|\bar{t_k}) log P(c_j|\bar{t_k})$$

Una vez calculado para cada atributo del conjunto de entrenamiento, se eligen aquellos cuyo índice está por encima de cierto umbral, normalmente 0.

- Información Mutua (MI): Asigna a cada atributo un índice por cada categoría. Este índice está relacionado con la medida de coocurrencia entre el atributo y la clase y se calcula como

$$I(t, c) = log \frac{A \times N}{(A + C) \times (A + B)}$$

dónde A es el número de veces que el término t y la clase c coocurren, B es el número de veces que t ocurre pero no c y C es el número de veces que c ocurre pero no t, N es el número de documentos en D_{tr}. Cuando han sido calculados, se eligen los n atributos con mayor MI para cada clase y se desecha el resto.

- Estadística Chi Cuadrada (χ^2): Es otra medida estadística relacionada con la coocurrencia entre cada atributo y las clases. Es diferente de MI porque proporciona un valor normalizado con lo que la medida de relevancia es comparable entre diferentes clases. Se calcula como

$$\chi^2(t, c) = \frac{(N \times (AD - CB)^2)}{(A + C) \times (B - D) \times (A + B) \times (C + D)}$$

donde A, B y C representan los mismos conteos de antes y D es el conteo de casos en los que no ocurre t ni c. Luego se elige a los atributos con mayor χ^2.

Una vez elegidos los atributos representativos de la colección, cada documento se

representa de una manera más compacta y más informativa por lo que se puede proceder a aplicar el algoritmo de aprendizaje automático para construir el clasificador.

2.1.3. Métodos de Clasificación

Un clasificador es generado mediante un proceso inductivo en el que se observan las características de los documentos del conjunto de entrenamiento D_{tr} que los hacen pertenecer a las clases de C con el propósito de poder identificar la categoría de un documento hasta entonces desconocido mediante un análisis de las características que comparte con los documentos del conjunto de entrenamiento.

Existen numerosos tipos de clasificadores con características, ventajas, desventajas y áreas de aplicación específicas, particularmente en clasificación automática de documentos se han usado clasificadores probabilísticos, basados en similitud geométrica, árboles de decisión y máquinas de vectores de soporte con buenos resultados. Entre ellos, los clasificadores probabilistas basados en la regla de Bayes y las máquinas de vectores de soporte han mostrado los mejores resultados con un costo computacional moderado, a continuación se describen el funcionamiento y características de cada uno de ellos.

Naïve Bayes

El enfoque probabilístico aplicado en clasificación de texto asume que los documentos de cada clase son generados por un modelo paramétrico y utiliza los datos de entrenamiento para calcular los parámetros óptimos de dicho modelo [34]. Una vez que se establece el modelo a partir de la información que se conoce, el clasificador invierte el modelo generativo empleando la regla de Bayes para calcular la probabilidad a posteriori de que el modelo de una clase haya generado al documento que se desea clasificar [6]. Entonces el problema de clasificación se convierte en una simple selección de la clase más probable.

En otros términos, un clasificador bayesiano calcula la probabilidad de que un documento pertenezca a cada una de las clases y lo asigna a aquella cuya probabilidad sea mayor, esto es,

$$c = \arg \max_{c_j \in C} P(c_j|d_i)$$

luego, aplicando el teorema de Bayes tenemos que,

$$c = \arg\max_{c_j \in C} \frac{P(d_i|c_j)P(c_j)}{P(d_i)}$$

como el denominador es constante para todas las clases puede evaluarse únicamente el numerador de la razón, es decir.

$$c = \arg\max_{c_j \in C} P(d_i|c_j)P(c_j)$$

Se dice que un clasificador bayesiano es naïve (ingenuo, simple) ya que supone la independencia condicional de todos los términos de un documento dada la clase, suposición que generalmente es falsa, por lo que la probabilidad del documento dada la clase puede verse como el producto de las probabilidades de los términos que en él aparecen, esto es, bajo la suposición naïve [27], la ecuación anterior puede escribirse como

$$c = \arg\max_{c_j \in C} P(c_j) \prod_{k=1}^{n} P(t_{ki}|c_j)$$

dónde la probabilidad de la clase está determinada como la fracción de documentos en el conjunto de entrenamiento que pertenecen a dicha clase y la probabilidad de un término dada la clase se calcula como la razón entre el número de documentos que contienen al término t_k, D_{t_k} y la sumatoria de documentos que contienen a cada uno de los términos del vocabulario D_{tl} y pertenecen a la clase c_i. Agregando un suavizado Laplaciano para evitar probabilidades nulas, esta probabilidad se calcula como:

$$P(t_{ki}|c_j) = \frac{1 + |D_{t_k}|}{|T| + \sum_{l=1}^{|T|} |D_{t_l}|}$$

donde $|T|$ es el tamaño del vocabulario de la colección.

2.1.4. Máquinas de Vectores de Soporte

Los clasificadores basados en máquinas de vectores de soporte realizan operaciones en el espacio de representación de los objetos que clasificarán para encontrar hiperplanos que separen los ejemplos positivos de una categoría de aquellos que no pertenecen a la clase con el mayor margen de separación posible. Cuando la separación lineal no es posible, recurren a separaciones no lineales por medio de una función denomina-

da kernel que les permite trasladar los objetos a espacios de mayor dimensionalidad donde la separación sea posible.

Las máquinas de vectores de soporte están basadas en el principio de minimización de riesgo estructural de Vapnik [39] y la idea principal consiste en encontrar una hipótesis h para la cuál se pueda garantizar el mínimo error cuando una muestra aleatoria y desconocida es sometida a dicha hipótesis, se trata entonces de un problema de optimización.

En clasificación automática de textos las hipótesis formuladas son los hiperplanos de separación entre las muestras que pertenecen a una clase y los que no, por lo tanto es necesario hallar una hipótesis que permita separar correctamente todas las muestras del conjunto de entrenamiento con el mayor margen posible de forma que al probar un nuevo objeto (lo cuál se traduce en clasificar un nuevo documento) el error esperado sea el mínimo posible. La herramienta matemática empleada para resolver el problema de optimización es compleja y puede hallarse una explicación detallada de ella en [16]. Las maquinas de vectores de soporte han mostrado un excelente desempeño cuando son empleadas para clasificar documentos debido a que las soluciones que encuentran no son función de los atributos de cada objeto sino de el grado de separabilidad existente entre los conjuntos formados por los objetos de cada clase, este dato es fundamental para entender algunos de los resultados experimentales obtenidos.

2.1.5. Evaluación de los Clasificadores

Los clasificadores suelen ser evaluados por su efectividad más que por su eficiencia. Debido a que la clasificación de textos es una tarea subjetiva, en la evaluación de la efectividad de un clasificador se considera su capacidad para producir resultados que correspondan con la clasificación que haría un experto en las áreas temáticas sobre las que el clasificador opera. De ahí la importancia de contar con un conjunto de prueba cuyas categorías de hecho son conocidas (aunque no por el clasificador) que sirva para evaluar su desempeño con respecto a la clasificación realizada por los usuarios humanos que las han etiquetado.

Debe señalarse que para evaluar a un clasificador es necesario considerar los resultados de su aplicación sobre el conjunto de prueba. Dichos resultados pueden agruparse, para cada categoría, en cuatro conjuntos resultantes de las combinaciones entre las categorías asignadas por el experto y las que fueron designadas por el clasificador.

- El conjunto TP_{c_j} (True Positive) contiene a los documentos que tanto el clasificador como el experto asignaron a la categoría c_j.

- El conjunto FP_{c_j} (False Positive) que contiene a los documentos que el clasificador etiquetó como pertenecientes a la clase c_j pero que el experto marcó con una etiqueta de categoría distinta.

- El conjunto TN_{c_j} (True Negative) que incluye a los documentos que el clasificador catalogó como no pertenecientes a la categoría c_j al igual que el experto.

- El conjunto FN_{c_j} (False Negative) que contiene a los documentos clasificados como no pertenecientes a la clase c_j pero que de hecho, según el experto, pertenecen.

Existen tres medidas ampliamente utilizadas que permiten medir distintos aspectos de la efectividad de un clasificador, cada una aporta un tipo diferente de información aunque en nuestro contexto (clasificación monoclase) las tres pueden ser empleadas. Las medidas usadas son la exactitud (A), la precisión (π) y el recuerdo (ρ), cada una de ellas más una combinación que reúne a π y ρ son definidas a continuación.

Exactitud

Se define como la proporción de documentos del conjunto de prueba D_{t_e} que fueron clasificados correctamente independientemente de la categoría a la que pertenecen. Es una medida global que no distingue entre categorías. En términos de los conjuntos antes mencionados, la exactitud se define como,

$$A = \frac{\sum_{c_j} TP_{c_j} + \sum_{c_j} TN_{cj}}{|D_{te}|}$$

Precisión

Corresponde a la probabilidad de que un documento clasificado como perteneciente a una categoría c_j pertenezca efectivamente a dicha categoría, se calcula para cada una de las clases como la razón entre el número de documentos correctamente clasificados

y los que el clasificador asigno a dicha categoría, es decir,

$$\pi = \frac{TP_{c_j}}{TP_{c_j} + FP_{c_j}}$$

Recuerdo

Esta medida refleja la capacidad de un clasificador para clasificar en una categoría a todos los documentos que de hecho pertenecen a ella. Corresponde con la probabilidad de que un documento que pertenece a la clase c_j sea correctamente clasificado y se calcula como la razón entre los documentos correctamente clasificados en c_j y los que, según el experto, pertenecen a dicha categoría, dicho de otro modo:

$$\rho = \frac{TPc_j}{TPc_j + FN_{c_j}}$$

Medida F_β

Se emplea para combinar la precisión y el recuerdo en una sola medida que además establece una proporción de importancia relativa entre una y otra (definida por el parámetro β). Regularmente se emplea un $\beta = 1$ que representa una importancia semejante entre π y ρ. La medida F_β se define como:

$$F_\beta = \frac{(1 + \beta^2)\pi\rho}{\beta^2\pi + \rho}$$

Dado que las tres últimas medidas establecen una evaluación de la efectividad para cada categoría de la colección, es necesario combinarlas para tener una medida global que describa el desempeño del clasificador en toda la colección. Existen dos maneras de combinar las medidas individuales de cada categoría, en una de ellas cada categoría es ponderada de forma idéntica a las demás (macropromedio) mientras que en la otra cada categoría es ponderada de acuerdo a la proporción de muestras del conjunto de entrenamiento que corresponden a dicha categoría (micropromedio). Las definiciones en términos de TP, FP, TN y FN de cada categoría para la precisión y el recuerdo se muestran en la tabla siguiente.

$Medida$	$Micropromedio$	$Macropromedio$
$Precision(\pi)$	$\pi = \frac{\sum_{j=1}^{\|C\|} TP_j}{\sum_{j=1}^{\|C\|} TP_j + FP_j}$	$\pi = \frac{\sum_{j=1}^{\|C\|} \pi_j}{\|C\|}$
$Recuerdo(\rho)$	$\rho = \frac{\sum_{j=1}^{\|C\|} TP_j}{\sum_{j=1}^{\|C\|} TP_j + FN_j}$	$\rho = \frac{\sum_{j=1}^{\|C\|} \rho_j}{\|C\|}$

2.2. Clasificación Semisupervisada

En situaciones reales es común que no se cuente con suficientes documentos etiquetados para entrenar un clasificador confiable. Etiquetar los datos manualmente es costoso por lo que se recurre a métodos semisupervisados que combinan pocos documentos etiquetados con documentos no etiquetados para entrenar al clasificador. Los métodos semisupervisados construyen un modelo deficiente con los pocos documentos etiquetados y luego lo modifican con los documentos no etiquetados tratando de que el modelo reforzado sea mejor que el inicial. Existen múltiples métodos de clasificación semisupervisada, los más comunes incluyen modelos generativos combinados con *expectation maximization*, *co-training* y *self-training*.

2.2.1. Modelos Generativos + EM

Entre los primeros trabajos de clasificación semi-supervisada se encuentra el de Nigam et al. [30] en el que se describe el paradigma de un modelo generativo combinado con un algoritmo de la familia de *expectation maximization*.

El modelo generativo describe la forma como están distribuidos los atributos en los documentos de cada clase y se construye observando la distribución de probabilidad condicional de los atributos que describen a los documentos dada la clase. Los algoritmos de la familia de EM se utilizan para estimar iterativamente datos faltantes en una distribución, en este caso las etiquetas de clase de los documentos no etiquetados son los datos faltantes que se desea estimar a partir del modelo generativo [25] [11]. El algoritmo propuesto es el siguiente:

1.- Construir un clasificador inicial estimando un modelo generativo a partir de los documentos etiquetados.

2.- Repetir hasta que los parámetros del clasificador no cambien:

- Usar el clasificador actual para etiquetar probabilísticamente los documentos no etiquetados.
- Volver a calcular los parámetros del modelo generativo a partir de las etiquetas asignadas probabilísticamente.

Los resultados demuestran que el método es efectivo pues mejoran el desempeño del clasificador cuando integran documentos no etiquetados al conjunto de entrenamiento [35] [41].

2.2.2. Self-training

Self-training es otra técnica de aprendizaje semisupervisado que consiste en entrenar un clasificador con los pocos documentos etiquetados y usarlo para etiquetar a los que no lo están [10] [43]. Debido a que el clasificador es deficiente, muchos de los documentos clasificados tendrán etiquetas equivocadas sin embargo un subconjunto de los documentos clasificados estarán correctamente clasificados. Como no es posible comprobar si un documento está o no correctamente clasificado, se utiliza algún criterio para asignar un grado de confianza en las etiquetas. Los documentos con etiquetas más confiables se retiran del conjunto no etiquetado y son agregados al conjunto de entrenamiento con el que se entrena un nuevo clasificador. Este procedimiento se repite hasta una condición de paro que puede ser agotar los documentos no etiquetados o cierto número de iteraciones. Puede resumirse en el siguiente algoritmo:

- Repetir hasta cumplir una condición de paro.
  - **1**.- Entrenar un clasificador con los pocos documentos etiquetados del conjunto de entrenamiento.
  - **2**.- Etiquetar los documentos no etiquetados
  - **3**.- Seleccionar los documentos confiables mediante algún criterio e integrarlos al conjunto de entrenamiento

2.2.3. Co-training

Otro método de aprendizaje semisupervisado comúnmente empleado en clasificación de texto es *co-training*. Este asume que los atributos que describen a los documentos pueden ser divididos en dos conjuntos, que con cada conjunto puede ser usado

para entrenar un clasificador y que los conjuntos son condicionalmente independientes dada la clase. Inicialmente se entrenan dos clasificadores, cada uno con un conjunto de atributos distinto, cada uno clasifica los documentos no etiquetados y selecciona aquellos en los que tiene mayor confianza (como en *self-training*), estos documentos se agregan al conjunto de entrenamiento del otro clasificador y se vuelve a entrenar, el proceso se repite hasta una condición de paro, al final los documentos en cuya clasificación ambos clasificadores estuvieron de acuerdo pueden usarse para entrenar uno nuevo [4] [23].

2.3. Elaboración Automática de Resúmenes

Un resumen automático es un texto producido a partir de uno o varios documentos que contiene la información importante contenida en los documentos de origen con una extensión considerablemente menor y usualmente inferior al 50 % [8] [31].

La investigación para elaborar resúmenes automáticos comenzó en la segunda mitad de los años 50, cuando Lunh propuso que la frecuencia de aparición de alguna palabra en un documento proporcionaba una medida de la significancia en el documento, mientras que para una oración completa, el factor de significancia podía determinarse a partir del número de palabras significativas que aparecen en ella y la distancia que las separa debido a la aparición de palabras no significativas [26]. De este modo las oraciones del documento pueden ordenarse y seleccionar las más significativas para formar un resumen automático.

En el mismo año, Baxendale [3] examinó 200 párrafos para encontrar que en el 85 % de los casos la primera y en 7 % la última de las oraciones del documento contenían el tema del que trata el documento. Este atributo basado en posición se sigue usando en sistemas de aprendizaje automático más complejos. Once años después se desarrollaba el primer experimento de elaboración automática de resúmenes (TS, *text summarization*) donde, para evaluar al sistema, se comparaban los resúmenes automáticos con resúmenes elaborados manualmente siguiendo una estructura determinada. En este experimento, además de los atributos de frecuencia y posición, se agregaban otros como la presencia de *cue-words* (palabras como *significativamente* o *difícilmente*) y el título o encabezado de los documentos. Más tarde en los 90 se desarrollan métodos que incorporan herramientas de aprendizaje automático como los clasificadores bayesianos o árboles de decisión. En estos se extraían atributos de

los documentos que debían ser resumidos y se aprendía cuando una oración debe pertenecer o no al resumen del documento dados sus atributos. Para reconocer dicha información, el sistema era entrenado con resúmenes de documentos elaborados manualmente [8].

Trabajos más recientes como el de Conroy y O'leary [7] modelan el problema de extraer oraciones de un documento usando modelos ocultos de Markov para tomar en cuenta dependencias locales entre oraciones. El modelo es construido usados estados alternantes de pertenencia o no al resumen, los estados de no pertenencia están anidados y los de pertenencia no por lo que se puede pasar de un estado de pertenencia a uno de no pertenencia pero no se puede hacer la transición inversa. La matriz de transición entre dos estados i y j es determinada empíricamente a partir de un conjunto de entrenamiento y asociado a cada estado existe una función de salida dependiente del estado que actúa sobre un vector de atributos observados. Los atributos con los que describen las oraciones del documento son solo tres, la posición de la oración en el documento, el número de términos en la oración, y la semejanza de los términos de la oración dados los términos del documento[7] [24]. Otros trabajos como el de Svore utilizan Redes Neuronales y atributos provenientes de otros algoritmos. Este logró sobrepasar un *baseline* impuesto en 2002 que consistía en utilizar las primeras n oraciones de un cable de noticias y atribuido a la convención existente entre los periodistas de colocar la información más importante en los primeros párrafos [38]. De un modo distinto, los trabajos de Mihalcea [28] [29] abordan el problema de determinar la relevancia de cada oración en el documento construyendo un grafo cuyas conexiones son determinadas por la similitud o traslape en las palabras de cada oración para calcular iterativamente la influencia de ésta sobre el resto del documento.

Sin embargo, a pesar de haber un gran avance en técnicas de procesamiento de lenguaje natural, muchas de las ideas de los trabajos seminales de elaboración automática de resúmenes siguen siendo utilizadas. Los resúmenes automáticos pueden elaborarse a partir de un documento (monodocumento) o a partir de varios (multidocumento). Existen dos enfoques fundamentales para elaborar un resumen automático: el extractivo y el abstractivo. Ambos enfoques son empleados en la generación de resúmenes monodocumento y multidocumento [40].

- Resúmenes Extractivos: La idea fundamental es identificar las partes importantes del documento o conjunto de documentos a partir del cual se elaborará el resumen; una vez que se han identificado, un nuevo documento es construi-

do concatenando los extractos así identificados. Los resúmenes producidos por este tipo de algoritmos no mantienen una continuidad lingüística debido a su naturaleza, usualmente son usados para alimentar a otros sistemas de tratamiento de documentos o como resúmenes informativos sobre la naturaleza de los documentos a partir del cuál son generados.

- Resúmenes Abstractivos: Este tipo de algoritmos sí tiene como objetivo producir un documento que sea legible por un usuario humano por lo que emplean herramientas avanzadas de generación de lenguaje. La calidad de los resúmenes abstractivos depende de la base de información con que cuente el sistema y en general no pueden obtenerse resúmenes con la calidad de uno elaborado por humanos. Tradicionalmente, los sistemas generadores de resúmenes abstractivos tienen tres etapas:

  1.- Identificación del Tópico: consiste en identificar de alguna manera el tema principal de el o los documentos de origen, una vez identificados los extractos con la información necesaria se produce un resumen esquemático o jerárquico que indica la evolución del discurso.

  2.- Interpretación: requiere el uso de herramientas o recursos lingüísticos externos para generar información que no existe de manera implícita en el documento original. Normalmente en esta etapa la información extraída de la identificación de tópicos es fusionada o comprimida.

  3.- Generación: mediante herramientas generativas de lenguaje se dota de coherencia semántica a la información producida en la etapa anterior. El resultado es un documento legible y coherente.

2.3.1. Técnicas para Elaborar Resúmenes Extractivos

La tarea principal cuando se desea elaborar resúmenes extractivos consiste en identificar la información relevante para el documento. Muchos enfoques han sido utilizados para resolver este problema, desde algunos muy simples como seleccionar el título del documento hasta métodos supervisados y basados en grafos, a continuación se describen algunos de los más relevantes.

2.3.2. Posición

Se trata de una de las técnicas más antiguas para elaborar resúmenes automáticos [13]. Está basada en la idea de regularidad en la estructura de documentos lo que establece que ciertas secciones de documentos como los encabezados, títulos, párrafos iniciales o finales tienden a contener información relevante. A pesar de la simplicidad de la técnica, elegir las primeras oraciones de un documento para formar su resumen es una técnica con efectividad comparable a la de otras técnicas más complejas.

2.3.3. Frecuencia

También es una de las técnicas que fueron empleadas en los primeros trabajos sobre elaboración de resúmenes extractivos. Esta sustentada por la ley de Zipf que establece que la frecuencia de aparición de ciertas palabras es, en general, inversamente proporcional al tamaño de dicho conjunto de palabras (pocas palabras aparecen muy frecuentemente, muchas palabras aparecen poco frecuentemente) [32]. A grandes rasgos las técnicas basadas en frecuencia identifican las palabras o frases más frecuentes de un documento y asumen que dichos extractos contienen la información relevante del documento. Esta técnica es muy efectiva con ciertos tipos de documentos (sobre todo de carácter informativo) aunque es deficiente con otros.

2.3.4. Palabras Clave

Para aplicar estrategias basadas en palabras clave es necesario conocer cierta información descriptiva sobre el documento, ésta puede ser el título, una petición, una serie de palabras empleadas para describir los tópicos del documento o para introducir conceptos clave. Cuando se conocen estas palabras clave, identificar los extractos que formarán al resumen es una tarea limitada a identificar los párrafos, oraciones o secciones con mayor similitud.

2.3.5. Enfoques Supervisados

Es posible identificar parámetros de un documento que determinan la relevancia de una oración o frase. Algunos pueden ser el tamaño de la oración, la posición, frecuencia, traslape con el título o con otras oraciones [40] [24] [8]. Caracterizando una serie de estos parámetros en un conjunto de documentos donde se conoce su

información relevante es posible entrenar un clasificador que luego sea utilizado para clasificar las oraciones de otros documentos como relevantes o no. Con la información relevante de cada documento se construye su resumen.

2.3.6. Basados en Grafos

Las oraciones, secciones o párrafos de un documento pueden representarse como nodos de un grafo. Por medio de algoritmos que determinan la influencia de cada nodo en el grafo, mediante medidas de similitud o distancia se puede asignar a cada oración del documento un ranking *local*. El resumen del documento puede construirse usando las n oraciones más relevantes y descartando el resto.

En [28] y [29] Mihalcea propone y evalúa un método para elaborar resúmenes automáticos usando grafos. El método está basado en una medida de *ranking* para páginas web denominado *Hyperlinked Induced Topic Search* [19], el objetivo es determinar un par de valores asociados a cada oración del documento, representada como un nodo en el grafo, a partir de la fortaleza de los arcos que le conectan a otras oraciones. El peso de los arcos está determinado por una medida de similitud entre dos oraciones. Los valores que caracterizan a cada oración se calculan iterativamente con las siguientes ecuaciones:

$$HITS_A(V_i) = \sum_{V_j \in In(V_i)} w_{ji} HITS_H(V_j)$$

$$HITS_H(V_i) = \sum_{V_j \in Out(V_i)} w_{ij} HITS_A(V_j)$$

En ellas, V_i es un nodo del grafo $G = (V, E)$, $In(V_i)$ es el conjunto de arcos entrantes a V_i, es decir, las oraciones que tienen términos comunes con aquella representada por el nodo V_i, mientras $Out(V_i)$ son los arcos salientes u oraciones con las cuales V_i tiene términos comunes. $HITS_A$ es el *valor de autoridad* para el nodo V_i que está en función de sus arcos entrantes, mientras que $HITS_H$ se llama *valor de hub* y se calcula considerando los arcos salientes de V_i.

La dirección de los arcos captura el orden de aparición de las oraciones en el documento. Este orden es relevante para calcular los pesos de cada nodo. Sin embargo, se ha observado que el valor de $HITS_H$ cuando el grafo es construido siguiendo el orden de lectura del documento es equivalente al de $HITS_A$ cuando se sigue el

orden inverso. De ahí que tomar cualquiera de los valores siguiendo el orden adecuado proporciona el mismo ordenamiento final de las oraciones.

Una vez calculado el índice $HITS$ de cada nodo del grafo, puede interpretarse como un *valor de importancia o relevancia* asociado a cada oración del documento, lo que permite ordenarlas y elegir las k oraciones más importantes para integrar el resumen del documento. En los experimento de [29] y [8] k es un valor fijo entre 1 y 10, sin embargo, como se verá en la sección de experimentos, conviene utilizar un k dinámico en función de la extensión del documento original.

Los resultados de [29] y [8] demuestran que este método para elaborar resúmenes automáticos es superior en desempeño cuando se utiliza en clasificación de documentos (reemplazando los documentos por sus resúmenes) a otros métodos basados en posición, frecuencia, palabras clave, enfoques supervisado y análisis semántico latente (LSA). Debido a que este método es independiente del tamaño del conjunto de entrenamiento, emplea únicamente información presente en el documento y tiene un desempeño superior a otros fue usado como estrategia para elaborar los resúmenes de este trabajo.

Capítulo 3

Trabajo Relacionado

El objetivo de este capítulo es presentar y comentar los trabajos que han funcionado como antecedentes en la elaboración de la investigación reportada en esta tesis. Aunque no existe un trabajo en que se halla explorado directamente la interacción entre resúmenes automáticos y clasificación semisupervisada de documentos, hay evidencia para sostener la hipótesis planteada en este trabajo reportada en artículos dónde se emplean estrategias y técnicas propias de elaboración de resúmenes (TS) con clasificación de documentos. Adicionalmente, con la revisión de trabajos sobre clasificación semisupervisada se ha verificado la factibilidad de los experimentos propuestos para probar la hipótesis.

En la subsección 3.1 de este capítulo se analizan las diferentes formas en las que trabajos previos han relacionado las áreas de elaboración automática de resúmenes y clasificación de documentos: adaptando estrategias de sumarización automática para clasificar, usando resúmenes como representaciones reducidas de los documentos y empleando sumarización para mejorar el desempeño de un clasificador. La sección 3.2 presenta algunas características de los sistemas semisupervisados, particularmente los basados en *self-training* que han sido empleados como antecedentes.

3.1. Clasificación de Documentos + Resúmenes Automáticos

Bajo la suposición de que un resumen contiene la información más relevante de un documento dado con una extensión significativamente menor, se ha estudiado la posibilidad de que remplazar documentos por sus resúmenes en un sistema de clasificación tenga un impacto positivo en el desempeñó del sistema. Esto debido a

que los resúmenes, en general, contienen una menor cantidad de palabras y a que desechan información que puede considerarse no útil o redundante para describir al documento.

La efectividad de combinar técnicas de elaboración automática de resúmenes con sistema de clasificación de documentos está en función de la calidad de los resúmenes generados. Pero debido a que producir resúmenes comparables con los elaborados por expertos (caso ideal) es muy costoso y requiere de herramientas avanzadas de procesamiento de lenguaje, existe un compromiso entre los recursos necesarios para generar los resúmenes y el beneficio obtenido en la clasificación.

Sin embargo, técnicas muy simples de elaboración de resúmenes como seleccionar el título o las primeras oraciones del documento han mostrado ser efectivas cuando se usan para clasificar ciertos tipos de documento -como noticias- debido a su naturaleza estructurada.

Pocos trabajos han reportado el uso de herramientas de elaboración automática de resúmenes (TS) y clasificación de documentos (TC), sin embargo, en ellos es posible identificar 3 distintos enfoques: emplear técnicas de TS para clasificar, técnicas de TS para reducir el espacio de representación de los documentos y TS para mejorar la clasificación. En las subsecciones siguientes se comentan los trabajos donde cada enfoque es explorado.

3.1.1. Técnicas de Elaboración Automática de Resúmenes para Clasificar Documentos

En trabajos de este tipo, técnicas tradicionalmente empleadas para elaborar resúmenes automáticos son adaptadas para clasificar documentos.

El primer trabajo donde se incorporan herramientas de elaboración automática de reśumenes y clasificación de texto es el de Ker y Chen [18]. En él, se emplea una técnica basada en posición y frecuencia para determinar la importancia de cada término en cada categoría del conjunto de entrenamiento. Luego, los documentos del conjunto de prueba se reemplazan por sus resúmenes (títulos) y se calcula un valor de pertenencia de cada documento para cada una de las categorías combinando el valor de importancia de cada término con su frecuencia de aparición en el resumen, lo cual es una técnica usada para medir la importancia de cada oración en el documento cuando se hacen resúmenes automáticos. El documento es clasificado en la categoría para la

que tiene el mayor valor de pertenencia. Los resultados obtenidos son comparados con la clasificación tradicional usando kNN y no consiguen superar al caso base pero tiene un desempeño aceptable.

Este enfoque basado en la medida de pertenencia o relevancia de cada documento a cada clase también ha sido explorado con algunas variantes en los trabajos de Jiang et al. [15] [14]. En éstos tanto los documentos del conjunto de entrenamiento como los de prueba son reemplazados por resúmenes automáticos producidos con las oraciones más relevantes de cada documento. Dicha relevancia es medida a partir de la cantidad de sustantivos, verbos y la distancia de separación entre ellos para cada oración. Sus resultados sobre un corpus balanceado, son superiores a los de un clasificador kNN usado como caso base.

3.1.2. Resúmenes Automáticos para Reducir el Espacio de Representación de los Documentos

Estos trabajos emplean herramientas de elaboración automática de resúmenes para obtener representaciones reducidas de los documentos a clasificar, particularmente se usan como una estrategia de selección de atributos.

Kolcz et al. en [22] hacen un estudio comparativo para medir el impacto de diferentes técnicas de elaboración automática de resúmenes basadas en posición, frecuencia y palabras clave, cuando son usadas como estrategia de selección de atributos. Las técnicas analizadas incluyen elaborar los resúmenes únicamente con el título del documento, con el primer o primeros dos párrafos, con el párrafo que incluye la mayor cantidad de palabras que aparezcan en el título, con el primer y último párrafos, con el párrafo que incluye la mayor cantidad de palabras clave y con las *mejores oraciones* definidas como aquellas que contienen al menos cuatro palabras clave y 3 palabras del título. Utilizan información mutua (MI) [44] como método base para sus comparaciones y demuestran que aunque las técnicas de selección de atributos emplean menor tiempo de procesamiento que MI (con excepción de *mejor oraciones* y el párrafo con más palabras clave) la efectividad de los métodos es muy semejante. Excepciones de esto son los casos en los que se elabora el resumen únicamente con el título y únicamente con el primer párrafo del documento, ya que emplean tiempo considerablemente menor en ser procesado y obtienen resultados ligeramente superiores a los de MI.

3.1.3. Resúmenes Automáticos para Mejorar la Clasificación

Después de experimentar con técnicas de elaboración automática de resúmenes aplicándolas para hacer clasificación y usar resúmenes automáticos como una estrategia de selección de atributos, algunos trabajos demostraron que es posible mejorar el desempeño de un sistema de clasificación ya sea seleccionando -mediante resúmenes- información más adecuada, reduciendo el ruido que se proporciona al clasificador o bien enriqueciendo la representación de los documentos.

Los primeros trabajos en los que se consigue mejorar el desempeño de un sistema de clasificación mediante el uso de herramientas de elaboración automática de resúmenes son los de Ko et al. [21] [20]. En ellos, se propone enriquecer los pesos de los términos con los que se representa a los documentos tanto en el conjunto de entrenamiento como en el de prueba, con información sobre la relevancia de la oración de la cual proceden en el documento. La información sobre la relevancia de cada oración se obtiene aplicando estrategias de elaboración automática de resúmenes. Particularmente, calculan un índice de relevancia de cada oración, de acuerdo con la similitud que mantiene con el título del documento y los valores de TF, DF y χ^2 de cada uno de los términos que en ella aparecen. Luego calculan un nuevo peso para cada término de acuerdo a su valor TF y el índice de relevancia de la oración de la cual proceden. Con este nuevo esquema de pesado, emplean clasificadores NB, Rocchio, kNN y SVM, los resultados sobrepasan ligeramente a los obtenidos por los métodos originales.

Poco después, Mihalcea propuso un método basado en grafos y medidas para rankeo de páginas web que permite elaborar resúmenes automáticos de gran calidad [28]. El algoritmo propuesto representa cada oración del documento como nodo de un grafo, los arcos y sus pesos representan relaciones de similitud entre las diferentes oraciones del documento. Una vez que se conocen los pesos de los arcos del nodo, se pueden calcular iterativamente, usando medidas de rankeo de páginas web como *HITS* [19] o *PageRank* [5], los pesos de cada nodo, que representan la relevancia de la oración correspondiente en el documento.

En un trabajo siguiente [29], Mihalcea y Hassan aplicaron y evaluaron el desempeño de la técnica propuesta en clasificación de documentos. Reemplazaron los documento por resúmenes extraídos automáticamente usando su método y lo compararon con otros métodos reportados en la literatura [22] mostrando tanto que los resúmenes así producidos son más efectivos (evaluándolos con la herramienta ROUGE) y

permiten mejorar la clasificación significativamente cuando se compara con el uso de documentos completos y estrategias de extracción más simples como tomar el primer párrafo o el título del documento.

Posteriormente, en un estudio más amplio, Shen et al. [36] compararon el desempeño de diversas técnicas para elaborar resúmenes automáticos en la tarea de reducir el ruido para hacer clasificación de páginas web. Las técnicas analizadas incluyen,

- Título: Los resúmenes de las páginas web estaban formados únicamente por el título del documento.
- Meta-datos: La información incluida en el campo de meta-datos de la página es usada como resumen de la misma.
- Descripción: Se trata de una descripción breve elaborada por usuarios humanos que también está incluida en la página web.
- Content Body: Las oraciones que constituyen el resumen son aquellas cuya representación vectorial es más semejante al vector que representa el documento completo.
- Estrategia de Luhn: Se trata de una adaptación de la técnica propuesta por Luhn para seleccionar información relevante en un documento a partir de las oraciones que contienen palabras de una lista elegida por su frecuencia de aparición [26].
- Análisis Semántico Latente: Un modelo de las relaciones entre términos del documento es construido para determinar un patrón, luego se seleccionan las oraciones que mejor representan a dicho patrón para constituir el resumen [9].
- HITS: Se trata de el método basado en grafos y *rankeo* de páginas web empleado en [29] (ver sección 2.3.6).
- Enfoque Supervisado: Las oraciones que formaran el resumen son seleccionadas por un clasificador. Cada oración es descrita con atributos como la posición en el párrafo, el tamaño, la frecuencia de sus términos, similitud con el título del documento entre otros. Es semejante al empleado en [24].
- Ensamble: Las oraciones son rankeadas usando una combinación lineal de los índices obtenidos usando las estrategias anteriores (título, metadatos, descrip-

ción, *content body*, estrategia de Luhn, *LSA*, *HITS* y enfoque supervisado), los pesos de los factores se obtienen empíricamente.

Los experimentos demuestran que la técnica HITS tiene un mejor desempeño que las demás, con excepción del ensamble que es ligeramente superior.

3.2. Clasificación Semi-Supervisada + Resúmenes Automáticos

El uso de técnicas de elaboración automática de resúmenes o resúmenes automáticos en clasificación semisupervisada no está documentada por lo que el trabajo relacionado más próximo corresponde a aplicaciones del algoritmo *self-training* en diferentes contextos a partir de los cuales es posible elegir las condiciones adecuadas para emplearlo en éste trabajo.

Uno de los trabajos más próximos es el de Guzmán-Cabrera et al. [10] que entrena un clasificador a partir de unos cuantos documentos etiquetados y genera una petición para realizar una búsqueda web con la información relevante de los documentos de cada clase, luego clasifica los *snippets* (fragmentos del contenido de una página web que el buscador muestra en los resultados de una búsqueda) devueltos por la máquina de recuperación de información, selecciona e integra los más confiables al conjunto de entrenamiento y reentrena el clasificador, repitiendo la operación hasta que una condición de paro es alcanzada. Dos características hacen especialmente relevante a este trabajo: La generación de una petición por cada clase a partir de la cual se buscara nueva información que alimente al clasificador y el hecho de que la información añadida al clasificador en cada iteración son documentos breves y cercanos por su naturaleza a un resumen más que a un documento completo.

En [37] Solorio propone una mejora para *self-training* empleando un ensamble de clasificadores en un algoritmo denominado *Orderer Classification* que permite seleccionar documentos más confiables para integrar al conjunto de entrenamiento basándose en una medida de entropía entre los resultados de la clasificación de los diferentes clasificadores que forman el ensamble.

Otros trabajos donde se hacen descripciones sobre el algoritmo *self-training* incluyen a [43] que lo emplea para entrenar un clasificador y compara su desempeño contra *Transfer Learning* y [42] que emplea el algoritmo para entrenar un clasificador

de oraciones subjetivas. Así como revisiones de trabajos relacionados con aprendizaje semisupervsado aplicado en clasificación de textos como [25], [35], [11] y [46].

A partir de las características observadas en los trabajos revisados se eligió *self-training* como algoritmo semisupervisado para el sistema. El método basado en grafos para elaborar resúmenes propuesto en [29] ha mostrado ser más efectivo que otros cuando los documentos de un sistema de clasificación son reemplazados por sus resúmenes por lo que será incorporado al sistema, finalmente, el uso de resúmenes automáticos para reducir el espacio de representación combinado con el enfoque en que se desea mejorar un sistema de clasificación mediante resúmenes dan evidencia de que es posible incorporar resúmenes en un sistema semisupervisado. El capítulo siguiente presenta una serie de experimentos donde los parámetros adecuados para incorporar resúmenes en el sistema semisupervisado han sido determinados.

Capítulo 4

Resúmenes Automáticos en Clasificación Supervisada

Como una etapa previa al desarrollo del sistema que implementa el método de clasificación semisupervisada con resúmenes automáticos, se realizó una serie de experimentos con el objetivo de analizar el impacto de los resúmenes automáticos en clasificación supervisada. Aunque este enfoque ha sido explorado en diferentes trabajos [22] [20] [29] [36] se ha considerado que una exploración minuciosa de algunos de los efectos de dicha combinación proporcionan información útil para el desarrollo del sistema.

Los artículos arriba citados emplean un enfoque bajo el cual los documentos son reemplazados por sus resúmenes automáticos tanto en el conjunto de entrenamiento como en el conjunto de prueba del sistema de clasificación; además de este esquema, en nuestros experimentos se analiza el efecto de usar resúmenes únicamente en alguno de los dos conjuntos, lo que condujo a algunos resultados importantes. También se profundiza el estudio de los resúmenes automáticos como estrategia de selección de atributos, que a diferencia de otros métodos tradicionalmente usados, no depende de información estadística del conjunto de entrenamiento por lo que es aplicable cuando se tienen pocos documentos de entrenamiento.

4.1. Conjuntos de Datos

Dos colecciones de documentos fueron empleadas para los experimentos de esta etapa, una colección de noticias (R8) y una de páginas web (WebKB), que por su naturaleza distinta permiten evaluar el alcance de los resultados obtenidos.

R8 es una colección de noticias, subconjunto de la partición ModApte de Reuters-21578 [1] que contiene a las 8 clases con mayor número de documento monoclase de la colección original, está desbalanceada. La distribución de documentos en los conjuntos de prueba y entrenamiento puede verse en la tabla 1.

	Conjunto	
Categoría	Entrenamiento	Prueba
earn	2701	1040
acq	1515	661
trade	241	72
crude	231	112
money-fx	191	76
interest	171	73
ship	98	32
grain	41	9
Total	5189	2075

Tabla 4.1: Distribución en las categorías de R8

WebKB es una colección de páginas web de 4 universidades, la colección original contiene 7 categorías pero en los experimentos realizados se retiró la clase *other* debido a que los documentos con dicha etiqueta no forman una categoría temática. La colección no está particionada en conjuntos prueba y entrenamiento por lo que cada clase fue dividida entre los dos conjuntos asignando los documentos con números de identificación pares al conjunto de prueba y los impares al conjunto de entrenamiento. La distribución final se puede ver en la tabla 2.

	Conjunto	
Categoría	Entrenamiento	Prueba
student	821	820
faculty	562	562
course	465	464
project	252	252
department	91	91
staff	68	69
Total	2259	2258

Tabla 4.2: Distribución en las categorías de WebKB

4.1.1. Preprocesamiento

Para poder aplicar el algoritmo de extracción automática de resúmenes es necesario que los documentos puedan separarse por oraciones. Los documentos originales de Reuters-21578 contienen encabezados de metatexto y texto plano. En el metatexto está incluida la información sobre las categorías de cada documento y pertenencia al conjunto de prueba o entrenamiento, esta información fue empleada para particionar el corpus, otra información referente a fechas, origen y localización de la noticia fue descartada. En el texto plano que forma el cuerpo de la noticia se consideró que las oraciones están separadas por puntos -que no indiquen abreviaturas- y cambios de línea o párrafo. Los documentos resultantes fueron llevados a minúsculas, se retiraron signos no alfabéticos y palabras vacías.

Los documentos de WebKB están en formato HTML por lo que toda la información está mezclada con etiquetas. Se consideró que las oraciones estaban separadas por cambios de línea o párrafo, puntos y etiquetas que indiquen cambios de línea, párrafo y sección como $< a >$, $< p >$ y $< br >$. Todas las demás etiquetas fueron descartadas. El texto resultante fue convertido a minúsculas, se retiraron signos no alfabéticos y palabras vacías. Dada la organización del texto en una página web, la estrategia para dividir el documento en oraciones produjo una separación que incluía muchas oraciones que contenían solo una palabra procedentes de secciones tales como títulos o menús, condición que pudo afectar el desempeño del sistema. La situación se comenta detalladamente en la discusión de los resultados.

En el experimento 1 no se utilizó ninguna estrategia de selección de atributos.

4.2. Experimento 1: Entrenamiento con Resúmenes Automáticos y Clasificación de Resúmenes Automáticos

En los trabajos reportados, los documentos de los conjuntos de entrenamiento y prueba del sistema de clasificación son reemplazados por sus resúmenes automáticos. Esto implica que el espacio de representación de los documentos está reducido al conjunto de atributos que describen a los resúmenes usados en el entrenamiento del clasificador y que los vectores que representan a los documentos, tanto de entre-

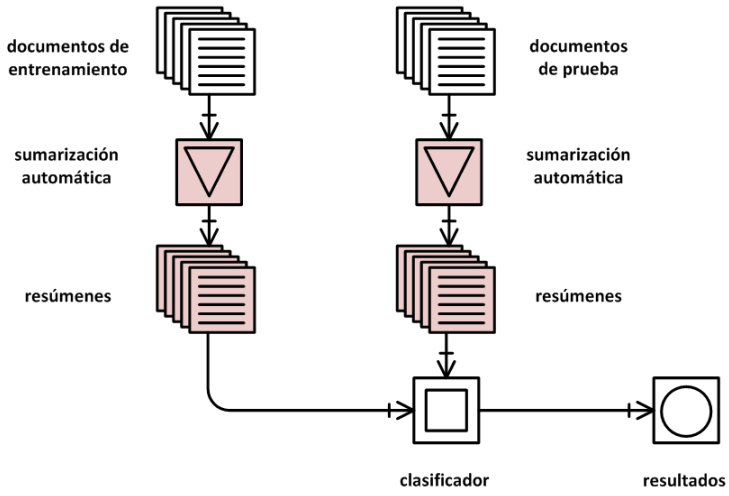

Figura 4.1: El esquema Res-Res en el que los documentos son sustituidos por sus resúmenes automáticos tanto en el conjunto de prueba como en el de entrenamiento. Este es el esquema utilizado en todos los trabajos reportados.

namiento como de prueba, contienen únicamente la información de dichos atributos por lo que una parte de la información que se encontraba en los documentos completos, ya sea en forma de atributos o como contribución al peso de los mismos, se pierde al representar los resúmenes de los documentos sobre el espacio de atributos del conjunto usado para entrenar al clasificador.

Con el objetivo de evitar esa perdida de información se propone el uso de resúmenes en uno de los conjuntos a la vez. Es decir, en el primer caso, usar resúmenes sólo para entrenar el clasificador mientras que los documentos a clasificar estarán completos aunque representados en el espacio de dimensiones de los resúmenes con que el clasificador fue entrenado. En el segundo caso el clasificador se entrenará con documentos completos y los documentos a clasificar serán reemplazados por sus resúmenes automáticos representados con los atributos de los documentos completos. Se emplea el conteo de apariciones del término en el documento TF o resumen como peso del atributo correspondiente para acentuar el efecto de los esquemas propuestos.

Otra característica de los trabajos donde se combinan resúmenes automáticos con clasificación es que los resúmenes suelen tener un tamaño fijo en número de oraciones. Esto implica la exclusión de documentos cuya extensión está por debajo de cierto umbral y limita el alcance de los resultados obtenidos debido a que la extensión del resumen no está en función del tamaño del documento original. Esto puede verse más claramente con un ejemplo hipotético a partir de la implementación de [29]. Supon-

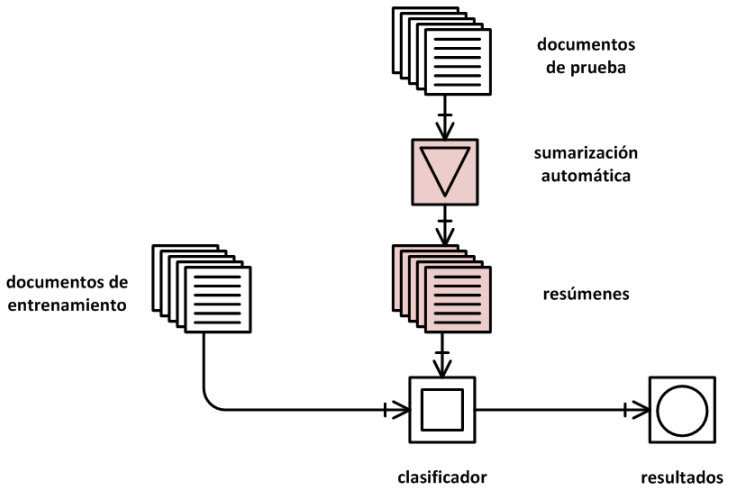

Figura 4.2: En el esquema Doc-Res el clasificador es entrenado con documentos completos. En la etapa de prueba los documentos son reemplazados por sus resúmenes.

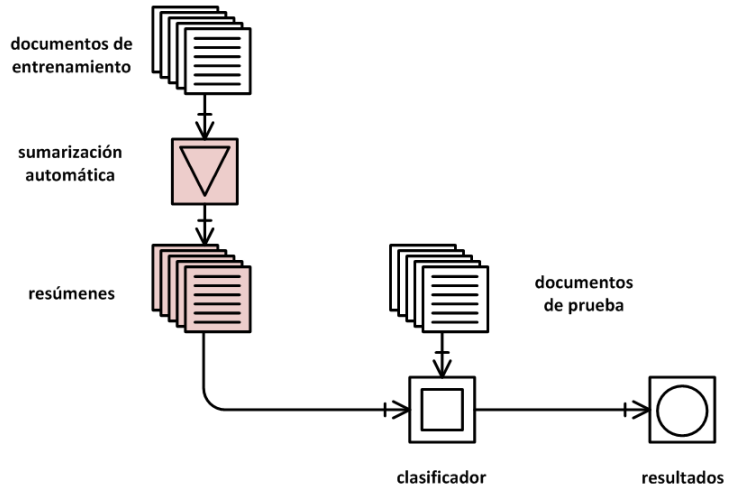

Figura 4.3: El esquema Res-Doc emplea resúmenes automáticos en lugar de documentos completos para entrenar al clasificador. Los documentos a clasificar así como los de prueba no son reemplazados por sus resúmenes..

gamos que en uno de los casos estudiados los resúmenes tendrán una extensión de 5 oraciones. Dado que los documentos no tienen un tamaño fijo, es fácil encontrar un documento que contiene 10 oraciones por lo que el resumen estará formado por el 50 % más representativo de dicho documento. Por el otro lado también es fácil encontrar un documento con 50 oraciones, lo que implica que en su resumen contendrá únicamente el 10 % de la información original. Para resolver este problema, los resúmenes utilizados en los experimentos aquí reportados no tienen una extensión fija a cierto número de oraciones sino proporcional al número de oraciones del documento original. Se analizan resúmenes con extensiones entre 10 % y 90 % del tamaño de los documentos originales. El algoritmo empleado para elaborar los resúmenes automáticos en todos los casos es la adaptación de HITS basada en grafos propuesta en [29].

Así, en este experimento hay cuatro esquemas para incorporar los resúmenes automáticos en el sistema de clasificación:

1.- Caso Base (Doc-Doc): Esquema de clasificación tradicional, sin uso de resúmenes, funciona como referencia para evaluar los otros casos.

2.- Resúmenes de Ambos Conjuntos (Res-Res): Los documentos en el conjunto de entrenamiento y en el conjunto de prueba son reemplazados por sus resúmenes automáticos.

3.- Clasificación de Resúmenes (Doc-Res): En este esquema los documentos del conjunto de prueba son reemplazados por sus resúmenes automáticos. El clasificador es entrenado usando documentos completos como es usual.

4.- Entrenamiento con Resúmenes (Res-Doc): Los documentos del conjunto de entrenamiento son reemplazados por sus resúmenes automáticos. Documentos completos se usan para probar el clasificador.

Los clasificadores usados son las implementaciones WEKA de Naive Bayes y SVM con el kernel por defecto. Los resultados obtenidos se muestran y comentan en la siguiente sección.

4.2.1. Resultados

En las siguientes tablas, la primera columna muestra el tamaño de los resúmenes empleados, en la segunda aparece la evaluación del método base que es independiente

	Esquema del Uso de Resúmenes			
Tamaño de Resumen	Doc-Doc	Res-Res	Doc-Res	Res-Doc
10 %	.810	.824	.522	**.836**
20 %	.810	.834	.736	**.835**
30 %	.810	**.847**	.795	.841
40 %	.810	**.849**	.810	.846
50 %	.810	.841	.808	**.843**
60 %	.810	.837	.807	**.838**
70 %	.810	**.843**	.811	.840
80 %	.810	.839	.812	**.840**
90 %	.810	.833	.811	**.836**

Tabla 4.3: Micropromedio de Medida F1 de los clasificadores del experimento 1 en R8 con Naive Bayes

	Esquema del Uso de Resúmenes			
Tamaño de Resumen	Doc-Doc	Res-Res	Doc-Res	Res-Doc
10 %	**.607**	.488	.554	.496
20 %	**.607**	.52	.481	.575
30 %	**.607**	.522	.459	**.607**
40 %	**.607**	.542	.459	.606
50 %	**.607**	.556	.456	**.607**
60 %	**.607**	.567	.461	.606
70 %	**.607**	.573	.466	.599
80 %	**.607**	.587	.471	.603
90 %	**.607**	.593	.464	.601

Tabla 4.4: Micropromedio de Medida F1 de los clasificadores del experimento 1 en WebKB con Naive Bayes

del tamaño de resumen debido a que no hay resúmenes involucrados, la columna 3 muestra los resultados de la evaluación cuando se emplean resúmenes tanto en el conjunto de prueba como en el de entrenamiento, la cuarta columna muestra los resultados de la clasificación de resúmenes y la última columna contiene los resultados cuando se entrena con resúmenes automáticos y se prueba documentos completos. La tabla 4.3 muestra la medida F1 de los experimentos sobre R8, la tabla 4.4 contiene los resultados de WebKB.

Una inspección de los resultados mostrados permite observar que,

- Reemplazar documentos por resúmenes automáticos (Res-Res) permite mejorar la clasificación en R8 pero no en WebKB. Este hecho puede deberse a la natura-

leza de los documentos: una noticia es un documento estructurado y consistente, formados por oraciones claras y que no requiere elementos externos a ella para completar su discurso por lo que el resumen de una noticia cumple con la condición de contener la información relevante con una extensión menor; por otro lado, el texto de una página web puede no contener un significado por sí mismo, siendo este complementado por el formato, hipervínculos, posición en el documento y otra información que no puede ser capturada en una versión reducida. Entonces, un resumen basado únicamente en el texto difícilmente captura la información relevante de una página web.

- Entrenar al clasificador con resúmenes automáticos y utilizar documentos completos para probarlos (Res-Doc) produce una mejor clasificación en la mayor parte de los casos en comparación a cuando se usan resúmenes en ambos conjuntos (Res-Res) y estos dos enfoques tienen un mejor desempeño que en el caso cuando se entrena al clasificador con documentos completos y se usan resúmenes para probarlo (Doc-Res). Este fenómeno puede observarse en ambos conjuntos de datos. Analizando cada caso puede hallarse una justificación.

  - Res-Res: Los conjuntos de entrenamiento y prueba están representados sobre el conjunto de atributos que describen localmente a los resúmenes del conjunto de entrenamiento y sus pesos están relacionados con la distribución de dichos atributos en los mismos resúmenes. En el conjunto de prueba los documentos tiene una distribución muy cercana aunque no idéntica; de forma que al resumirlos, atributos que no mantienen esa distribución localmente (es decir en cada documento) pero que son determinantes para su clasificación se pierden.

  - Res-Doc: En este caso, nuevamente los conjuntos de entrenamiento y prueba están representados sobre el conjunto de atributos que describe localmente a los resúmenes del conjunto de entrenamiento, la diferencia radica en que no habrá pérdida de atributos en el conjunto de prueba debido a que no se aplica sumarización automática sobre sus documentos.

  - Doc-Res: Aquí los conjuntos de entrenamiento y prueba se representan sobre el conjunto completo de atributos de los documentos de entrenamiento. Debido a que los documentos de prueba están resumidos, el problema de

la pérdida de atributos que aparece en Res-Res se hace más notable, tal como los resultados demuestran.

Así, entrenar con resúmenes y probar con documentos completos minimiza la pérdida de información al mismo tiempo que mantiene los beneficios de usar resúmenes automáticos.

- En R8, cuando se usa el esquema Res-Res, resúmenes con una extensión entre el 30 % y 70 % tienen un mejor desempeño. Este comportamiento también se observa con Res-Doc y da una referencia acerca de la cantidad de información relevante contenida en un documento con respecto a su tamaño. Esto no significa que resúmenes muy pequeños no sean útiles (en ambos casos, con resúmenes de 10 % también se sobrepasa al caso base) sino que el tamaño óptimo de un resumen para documentos de este tipo (noticias) podría encontrarse entre los límites señalados.

Ahora, estos resultados muestran el comportamiento de un clasificador basado en distribuciones probabilísticas al incorporar resúmenes automáticos en el esquema de clasificación. Como se ha visto, la distribución de los atributos cuando se usan resúmenes automáticos tiende a cambiar. Un clasificador que determina la categoría de los documentos mediante un análisis distinto, como SVM, permite comprobar algunas hipótesis y comparar el impacto de los resúmenes cuando diferentes tipos de características de los documentos son tomadas en cuenta para realizar la clasificación. Las tablas 4.5 y 4.6 muestran los resultados de los mismos experimentos usando un clasificador SVM.

Los resultados mostrados en las tablas 4.5 y 4.6 son consistentes con las observaciones hechas a partir de 4.3 y 4.4. Además, puede observarse que la ventaja del esquema Res-Doc contra Res-Res se hace más notable al utilizar SVM debido a que este clasificador no depende de la distribución de frecuencias de los términos, sino que emplea información sobre la ubicación en espacial de los vectores que representan a los documentos para distinguir entre las diferentes clases. Esta condición se hace más notable en el experimento siguiente.

	Esquema del Uso de Resúmenes			
Tamaño de Resumen	Doc-Doc	Res-Res	Doc-Res	Res-Doc
10 %	.842	.825	.645	**.876**
20 %	.842	.852	.737	**.886**
30 %	.842	.872	.775	**.891**
40 %	.842	.862	.799	**.877**
50 %	.842	.864	.817	**.870**
60 %	.842	.863	.823	**.864**
70 %	.842	.861	.830	**.862**
80 %	.842	.860	.836	**.861**
90 %	.842	.855	.837	**.856**

Tabla 4.5: Micropromedio de Medida F1 de los clasificadores del experimento 1 en R8 con Support Vector Machines

	Esquema del Uso de Resúmenes			
Tamaño de Resumen	Doc-Doc	Res-Res	Doc-Res	Res-Doc
10 %	**.794**	.570	.606	.650
20 %	**.794**	.620	.561	.705
30 %	**.794**	.659	.572	.749
40 %	**.794**	.684	.572	.766
50 %	**.794**	.714	.597	.772
60 %	**.794**	.740	.606	.783
70 %	.794	.770	.637	**.797**
80 %	.794	.787	.637	**.797**
90 %	**.794**	.786	.613	.793

Tabla 4.6: Micropromedio de Medida F1 de los clasificadores del experimento 1 en WebKB con Support Vector Machines

4.3. Experimento 2: Resúmenes Automáticos como Herramienta para Selección de Atributos

Como se ha visto en el experimento anterior, el esquema de clasificación en el que se incorporan resúmenes automáticos únicamente en el conjunto de entrenamiento tiene un desempeño superior al caso en el que se usan en ambos conjuntos. Debido a que la cantidad de atributos hallados en los resúmenes automáticos es inferior a la que se encuentra en un conjunto de documentos completos, puede desarrollarse una estrategia de selección de atributos a partir del uso de resúmenes automáticos. El uso de resúmenes automáticos como selección de atributos ha sido parcialmente explorado en el trabajo de Kolcz [22] donde se emplearon técnicas simples basadas en posición y frecuencia de palabras para elaborar resúmenes de los conjuntos de entrenamiento y prueba, llegando a la conclusión de que el desempeño es comparable con el de otra técnica como información mutua. Sin embargo, empleando una técnica de elaboración de resúmenes más efectiva y aplicando el esquema de clasificación que mejora los resultados (Res-Doc) se pretende desarrollar y evaluar una nueva estrategia para selección de atributos.

Los resúmenes automáticos seleccionan la información más relevante de cada documento, debido a esto, su uso como herramienta para seleccionar atributos tiene la ventaja de ser independiente del tamaño del conjunto de entrenamiento. Esto permite que pueda aplicarse en dominios donde otras técnicas (IG, MI, CHI) no funcionarían por requerir grandes cantidades de información para construir modelos estadísticos suficientemente generales, como en el caso de tener conjuntos de entrenamiento con pocos elementos.

El objetivo de este experimento es evaluar el uso de resúmenes automáticos como herramienta para seleccionar atributos, comparar su desempeño contra el de una técnica de uso común (IG) y explorar los alcances para situaciones en las que se cuenta con pocos datos de entrenamiento.

Para explorar estas posibilidades se realizarán pruebas sobre R8 y dos subconjuntos con menor cantidad de documentos de entrenamiento. R8-41 contiene 41 documentos de entrenamiento en cada categoría (41 es el número de documentos en la categoría más pequeña de R8, razón por la que todas fueron reducidas a ese tamaño) y R8-10 contiene únicamente 10 documentos de cada clase lo que representa un caso

en el que se cuenta con muy pocos documentos por categoría. Los documentos que forman parte de cada categoría fueron elegidos aleatoriamente. En todos los casos, el conjunto de prueba está formado por los 2075 documentos del conjunto de prueba original de R8 (tabla 4.1).

Dos son los casos base utilizados como referencias en este experimento:

1.- Utilizar todos los atributos de los documentos de entrenamiento.

2.- Emplear únicamente los atributos con ganancia de información mayor que 0 [44].

En la primera parte del experimento se analizan los resultados de los casos base y del uso de resúmenes automáticos con radios de compresión entre 10 % y 90 %. Debido a que el número de atributos seleccionados por ganancia de información es siempre menor que el de atributos hallados en los resúmenes automáticos, en una segunda etapa del experimento se compara el uso de resúmenes como selección de atributos contra el uso de la misma cantidad de atributos con mayor IG. Esto es, si los resúmenes de 10 % contenían 706 atributos, su desempeño se comparaba contra el uso de los 706 atributos con mayor ganancia de información. Estos resultados aparecen etiquetados como TOP-IG.

4.3.1. Resultados

En la tabla 4.7 se muestran los resultados de los casos base contra los que se compara la técnica propuesta en cada conjunto de datos. En la fila superior se encuentra el número de atributos, exactitud y medida F1 obtenidos cuando se emplean todos los atributos de los documentos (no se aplican resúmenes automáticos) de cada conjunto de entrenamiento. En la fila inferior aparecen las mismas medidas cuando se realiza una selección de atributos basada en el criterio de $IG > 0$. Todos los resultados de este experimento corresponden al uso del clasificador SVM. Los resultados de Naive Bayes son consistentes y pueden consultarse en el anexo A.

Puede observarse en la tabla anterior que cuando se tiene un conjunto de entrenamiento suficientemente numeroso como el caso de R8 (5189 documentos), un método de selección de atributos como IG puede mejorar el desempeño del clasificador reduciendo el número de atributos hasta en un 90 %. Por el otro lado, también se observa que un método basado en información estadística como IG tiene un mal desempeño

	Atributos	Exactitud	Medida F1
Conjunto de Entrenamiento: R8			
Todos los Atributos	17,336	85.25	.842
$IG > 0$	1,691	86.51	.857
Conjunto de Entrenamiento: R8-41			
Todos los Atributos	5,404	78.75	.782
$IG > 0$	54	42.89	.539
Conjunto de Entrenamiento: R8-10			
Todos los Atributos	2,305	71.71	.702
$IG > 0$	20	35.57	.424

Tabla 4.7: Casos Base con el Clasificador SVM: Usando todos los atributos y con el criterio de Ganancia de Información > 0.

cuando se aplica en conjuntos de entrenamiento de tamaño reducido pues los resultados de la clasificación sobre R8-41 y R8-10 muestran una importante caída. En estos conjuntos, IG reduce radicalmente el número de atributos debido a que no cuenta con suficientes datos que le permitan hacer una buena evaluación de la calidad de los atributos. Así, aunque reduce los atributos hasta en un 99 %, la calidad de los clasificadores con ellos entrenados decae significativamente.

En las tablas 4.8, 4.9 y 4.10 se muestran los resultados de aplicar resúmenes automáticos como estrategia de selección de atributos en los diferentes conjuntos, el número de atributos seleccionados por los resúmenes y la comparación contra el uso del mismo número de atributos con mayor IG. Los resultados marcados en negritas en la columna correspondiente al método propuesto representan aquellos en los que la mejora obtenida es estadísticamente significativa (usando una prueba-z [17] con confianza del 95 %) con respecto al uso de todos los atributos y a los atributos seleccionados por el criterio de ganancia de información > 0.

Como se observa en las tablas, la selección de atributos basada en resúmenes automáticos puede mejorar el desempeño del sistema de clasificación tanto cuando se compara con el uso de todos los atributos como en comparación con una estrategia tradicional como ganancia de información. En el conjunto R8 se observa que el micropromedio de medida F1 alcanza hasta 0.891 cuando se utilizan resúmenes automáticos en comparación con el 0.857 obtenido con el criterio de $IG > 0$ aunque este permite una reducción del 90 % de los atributos mientras que los resúmenes solo reducen el conjunto en poco más del 50 %. Por el otro lado, cuando se emplea el

	R8		
Tamaño de Resumen	Número de Atributos	Método Propuesto	Top-IG
10 %	8,289	**.876**	.846
20 %	9,701	**.886**	.846
30 %	11,268	**.891**	.848
40 %	12,486	**.877**	.848
50 %	13,320	**.870**	.848
60 %	14,560	.864	.846
70 %	15,626	.862	.845
80 %	16,339	.861	.847
90 %	17,063	.856	.843

Tabla 4.8: Micropromedios de Medida F1 de los clasificadores del experimento 2 donde se compara el uso de resúmenes contra IG en R8 con SVM.

	R8-41		
Tamaño de Resumen	Número de Atributos	Método Propuesto	Top-IG
10 %	1,943	**.842**	.817
20 %	2,445	**.834**	.790
30 %	3,089	**.836**	.789
40 %	3,569	**.842**	.789
50 %	3,919	**.819**	.791
60 %	4,348	.798	.787
70 %	4,671	.800	.786
80 %	5,004	.803	.780
90 %	5,263	.784	.781

Tabla 4.9: Medida F1 (Micropromedio) del experimento 2 donde se compara el uso de resúmenes contra IG en R8-41 con SVM.

	R8-10		
Tamaño de Resumen	Número de Atributos	Método Propuesto	Top-IG
10 %	706	**.776**	.572
20 %	902	**.709**	.659
30 %	1,178	.654	.618
40 %	1,392	**.766**	.631
50 %	1,523	**.763**	.700
60 %	1,722	.683	.717
70 %	1,890	.685	.716
80 %	2,082	.693	.698
90 %	2,230	**.712**	.703

Tabla 4.10: Micropromedios de Medida F1 del experimento 2 donde se compara el uso de resúmenes contra IG en R8-10 con SVM.

mismo número de atributos con mayor IG que los obtenidos con los resúmenes, la clasificación no mejora con respecto a $IG > 0$ y los resúmenes continúan teniendo un mejor desempeño.

Los resultados de R8-41 y R8-10 además muestran que IG no es aplicable cuando se tienen pocos documentos en el conjunto de entrenamiento ya que tras aplicar este método los resultados de clasificación decayeron hasta 50 % en exactitud. En R8-41, el utilizar el mismo número de atributos con mayor IG que los obtenidos con los resúmenes mostró resultados superiores a los obtenidos con $IG > 0$ aunque inferiores a los que se obtienen usando resúmenes automáticos. Esto puede justificarse en el hecho de que 41 documentos de cada clase todavía pueden contener información suficiente para asignar un índice suficientemente significativo a cada atributo aunque se requiere un criterio distinto a $IG > 0$ para seleccionarlos. Se observa también que los resúmenes siguen mostrando un desempeño superior.

Finalmente, en R8-10 los documentos del conjunto de entrenamiento son tan pocos que tanto el criterio de $IG > 0$ como el uso del mismo número de atributos que los encontrados en los resúmenes resulta insuficiente para obtener una buena clasificación. En cambio, los resúmenes demuestran ser una buena herramienta para seleccionar atributos pues mejoran significativamente los resultados ya que obtienen hasta 0.776 como medida F1 mientras el caso base con todos los atributos está evaluado en 0.702 mientras que $IG > 0$ obtiene 0.424 y el uso del mismo número de atributos con mayor IG tiene solo 0.572.

4.4. Conclusiones

En este capítulo se han mostrado los resultados de una serie de experimentos con clasificación supervisada que tuvieron como objetivo determinar algunas de las condiciones con las que será posible desarrollar el sistema de clasificación semi-supervisada que incorpora resúmenes automáticos. Una segunda motivación vino del hecho de que muchos aspectos del fenómeno no habían sido explorados en la literatura y podrían arrojar resultados determinantes para el trabajo que se realiza.

Como se ha visto, existe más de una forma de integrar resúmenes automáticos en un sistema de clasificación de documentos, cada uno con características propias que modifican el comportamiento del sistema. Con los experimento aquí realizados se ha tratado de explorar la mayor parte de dichas variantes a fin de conocer en cuál de ellas es posible aprovechar al máximo la capacidad de los resúmenes automáticos para concentrar la información de los documentos de los que proceden.

También se han explorado formas distintas a las reportadas para determinar la extensión de los resúmenes. Concretamente, extensiones proporcionales al tamaño de los documentos originales, de forma que la información contenida en el resumen no esté acotada a una extensión fija sino que mantenga una proporción con el original.

Del mismo modo, y una vez determinado el esquema para incorporar los resúmenes automáticos en el sistema de clasificación, se exploró la posibilidad y el alcance de usar a los resúmenes como una herramienta para seleccionar atributos.

De los experimentos realizados ente capítulo se pueden obtener las siguientes conclusiones:

- **Se ha realizado una exploración de los esquemas mediante los cuales se pueden integrar resúmenes automáticos en un sistema de clasificación.** Ampliando las posibilidades reportadas en la literatura donde se emplea únicamente un esquema en el que los documentos son reemplazados por sus resúmenes tanto en el conjunto de entrenamiento como en el de prueba, se ha determinado que el esquema en el que se entrena al clasificador con resúmenes y se prueba con documentos completos representados sobre el conjunto de atributos de los resúmenes de entrenamiento produce mejores resultados debido a que se minimiza la perdida de información inherente al proceso de sumarización en los documentos de prueba.

- **Se ha demostrado que los resúmenes automáticos pueden mejorar la clasificación.** Esto comprueba los resultados reportados en trabajos previos aunque se amplia el conocimiento sobre el fenómeno al explorar esquemas distintos a los reportados. Así mismo se muestra que esta conclusión es válida en documentos para los cuales elaborar un resumen tiene sentido como noticias, no siendo así el caso de documento cuyo contenido radica también en partes ajenas a ellos como las páginas web.

- **Se ha demostrado la utilidad de los resúmenes automáticos como herramienta de selección de atributos.** Especialmente en casos donde otros métodos basados en información estadística no permiten hacer una selección adecuada como en el caso de tener conjuntos de entrenamiento con pocos documentos. El uso de resúmenes automáticos demostró ser consistente y mejorar la clasificación significativamente con una reducción en el número de atributos de hasta el 70 %.

- **Empíricamente se ha mostrado que el clasificador SVM mejora el efecto de incorporar resúmenes automáticos en un sistema de clasificación.** Debido a las modificaciones en la distribución de los atributos en los documentos al sumarizarlos, los resultados de Naive Bayes son variables mientras los de SVM muestran una mayor consistencia y mejor desempeño al incorporar resúmenes.

Como se verá en el capítulo siguiente, los resultados aquí mostrados determinaron algunas de las características del sistema que implementa el método semisupervisado.

Capítulo 5

Clasificación Semisupervisada con Resúmenes Automáticos

El objetivo principal de este trabajo es el desarrollo de un método de clasificación semisupervisada que incorpora resúmenes automáticos. Como se ha mostrado en el capítulo anterior, incorporar resúmenes automáticos confiables en un sistema de clasificación puede contribuir a mejorar el desempeño. Además, si los resúmenes son usados únicamente para entrenar al clasificador, su impacto puede aumentarse debido a que el clasificador recibirá una menor cantidad de datos innecesarios mientras mantiene la información relevante de los documentos con los que se entrena. Aplicar este enfoque en un problema de clasificación semisupervisada es el objetivo de este capítulo.

En la primera sección se describe la arquitectura del sistema propuesto así como cada una de las partes que lo integran. Luego se describen los conjuntos de datos utilizados para evaluarlo. En la tercera sección se describen los experimentos realizados y se muestran los resultados obtenidos. Finalmente, se discuten los resultados y se presentan algunas conclusiones.

5.1. Sistema Propuesto

El sistema que se propone para implementar el método de clasificación semisupervisada está basado en *self-training*. Como todas las estrategias semisupervisadas, *self-training* parte del supuesto de que se cuenta con un conjunto reducido de documentos etiquetados D_L más un conjunto suficientemente grande de documentos que no están etiquetados D_U. El objetivo es obtener y utilizar la mayor cantidad de in-

formación procedente de los documentos no etiquetados para mejorar al clasificador inicial $\hat{\Phi}_0$ que es entrenado utilizando únicamente el reducido conjunto etiquetado. Esto se consigue utilizando el clasificador inicial para clasificar al conjunto no etiquetado, luego, mediante algún criterio, seleccionar entre estos documentos a aquellos en los que se tiene un mayor grado de confianza con respecto a la etiqueta asignada D_C, retirarlos del conjunto no etiquetado y agregarlos, junto con los originalmente etiquetados, al conjunto con el que se entrenará un nuevo clasificador $\hat{\Phi}_1$. El proceso se repite hasta cumplir una condición de paro tal como que el clasificador no pueda mejorarse, que se agoten los documentos no etiquetados o bien cumplir un número determinado de iteraciones.

Como se sabe a partir de los experimentos del capítulo anterior, reemplazar a los documentos del conjunto de entrenamiento por sus resúmenes automáticos y clasificar documentos completos es una buena estrategia para mejorar el desempeño del clasificador. Entonces, incorporando esta idea a un esquema semisupervisado, el algoritmo de *self-training* que usa resúmenes automáticos es el siguiente:

1.- Elaborar los resúmenes automáticos del conjunto inicial de documentos etiquetados.

2.- Entrenar un clasificador $\hat{\Phi}_0$ con los resúmenes del conjunto inicial de documentos etiquetados.

3.- Clasificar los documentos del conjunto no etiquetado D_U.

4.- Evaluar, mediante algún criterio, la confianza de las etiquetas asignadas a los documentos no etiquetados.

5.- Seleccionar los documentos más confiables de cada categoría D_C.

6.- Retirar los documentos confiables del conjunto no etiquetado $D_{U_i} = D_{U_{i-1}} - D_C$.

7.- Elaborar los resúmenes automáticos de los documentos confiables.

8.- Agregar los resúmenes de los documentos confiables al conjunto etiquetado $D_{L_i} = D_{L_{i-1}} + D_C$.

9.- Entrenar un nuevo clasificador $\hat{\Phi}_i$ utilizando el nuevo conjunto etiquetado.

10.- Si la condición de paro se ha cumplido terminar. De otro modo, regresar a 3.

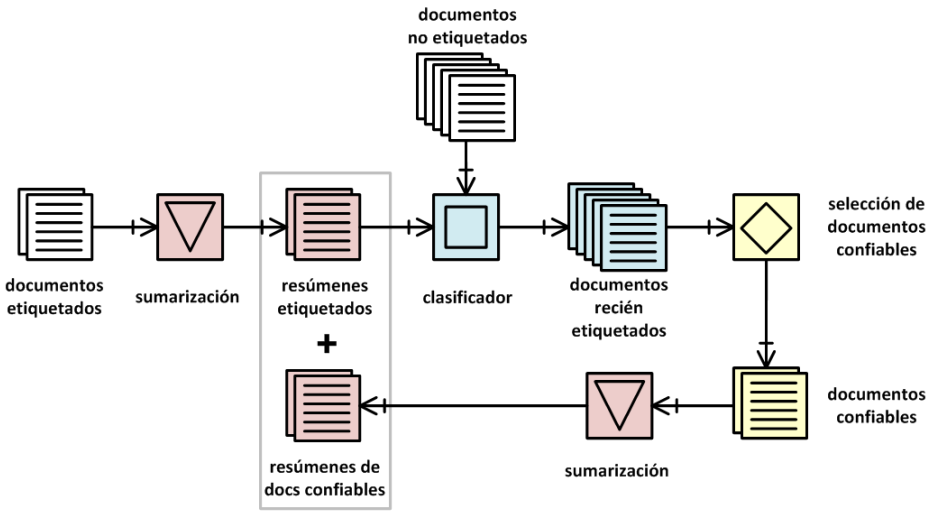

Figura 5.1: Arquitectura del sistema semisupervisado de clasificación de documentos que incorpora resúmenes automáticos en la etapa de entrenamiento.

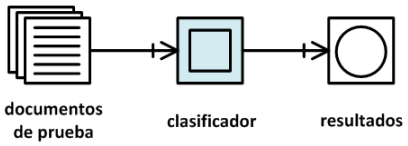

Figura 5.2: Etapa de prueba del sistema propuesto

Las figuras 5.1 y 5.2 muestran la arquitectura de la implementación propuesta. En la sección siguiente se describe cada uno de sus componentes.

5.1.1. Resúmenes Automáticos

Los resúmenes automáticos son elaborados mediante una implementación del algoritmo basado en grafos propuesto en [28], que han demostrado producir resúmenes más confiables que otros métodos [29] [8]. En los experimentos realizados se prueban extensiones entre 10 % y 90 % del tamaño de los documentos originales.

5.1.2. Clasificador

Los resultados de los experimentos en clasificación supervisada demostraron que *support vector machines* tiene un mejor desempeño cuando se emplean resúmenes

automáticos. Particularmente cuando el conjunto de entrenamiento es reducido. Esta ventaja se debe a que SVM busca los planos que maximizan la separación entre los documentos de las distintas categorías del conjunto de entrenamiento y dichos planos no dependen directamente de la cantidad de documentos que se tiene. En contraste, clasificadores como Naive Bayes, dependen directamente de las distribuciones de probabilidad de los atributos en los documentos por lo que un conjunto de entrenamiento reducido tiene un impacto mayor sobre clasificadores de este tipo.

5.1.3. Método de Selección de Documentos Confiables

En algunos sistemas de clasificación semi-supervisada que utilizan clasificadores probabilísticos, la confianza en la etiqueta que el clasificador asigna a un documento puede conocerse a partir de la probabilidad con la que el clasificador marcó al documento como perteneciente a su clase. Este criterio es muy ambiguo porque depende directamente de la calidad del clasificador y a su vez, de la cantidad y calidad de los documentos empleados en el entrenamiento. Debido a que en esta implementación no se emplean clasificadores probabilistas, se desarrolló un nuevo criterio de selección adecuado para la naturaleza del clasificador empleado.

El criterio de selección está dividido en dos etapas. En la primera de ellas se calcula la distancia de cada uno de los resúmenes de los documentos que originalmente no estaban etiquetados $d \in D_U$ hacia cada uno de los clústers formados por los resúmenes de los documentos del conjunto de entrenamiento de cada categoría C_j. La distancia entre el resumen del documento y el clúster es la utilizada en el *coeficiente Silhouette* [2] [12] y se calcula como,

$$dist(d, C_j) = \frac{\sum_{d_L \in C_j} dist(d, d_L)}{|C_j|}$$

Los documentos cuya etiqueta coincide con la del clúster hacia el cual guardan una distancia menor pasan a la siguiente etapa. En la etapa siguiente, los documentos recién etiquetados de cada categoría son ordenados ascendentemente según la distancia que guardan al clúster al que pertenecen. Dependiente del número de documentos que se integrará al conjunto etiquetado en cada iteración, los k documentos con menor distancia al clúster son elegidos como confiables. La medida de distancia $dist(d, d_L)$ es la distancia euclidiana entre el resumen y cada uno de los documentos del clúster.

Este criterio permite que los clústers formados por los documentos etiquetados en cada iteración sean lo más homogéneos posible con respecto a los documentos empleados en el conjunto etiquetado inicial.

5.2. Conjuntos de Datos

Los documentos empleados en los experimentos realizados con el método de clasificación semisupervisado que emplea resúmenes automáticos se describen en esta sección. Entre ellos, se pueden hallar colecciones de documentos de naturaleza distinta. Cabe indicar que, como los experimentos previos demostraron, el método es aplicable siempre que la naturaleza de los documentos permita realizar un resumen de ellos. Esto es, documentos como páginas web para los cuales realizar un resumen automático carece o tiene poco sentido dejan esta condición de manifiesto en los resultados obtenidos de sus experimentos. Otras colecciones sobre las que se aplicó el método incluyen noticias en inglés (R8), cablegramas y noticias (Meter), noticias en español (Desastres Naturales) y artículos con el formato de Wikipedia (Wikis). La descripción de cada colección se muestra a continuación.

5.2.1. R8

Se trata de la misma colección empleada en los experimentos supervisados (ver sección 4,1). Noticias en 8 categorías desbalanceadas. La partición original [1] incluye un subconjunto de entrenamiento y uno de prueba. Para los experimentos aquí realizados, el conjunto de prueba se mantuvo en tanto que el conjunto de entrenamiento fue dividido, una pequeña porción se usó como conjunto etiquetado mientras que al resto se le retiraron las etiquetas de categoría y se usó como conjunto no etiquetado. El número de documentos etiquetados varía según los experimentos y se comenta en la sección correspondiente.

5.2.2. Desastres Naturales

Se trata de una colección de noticias en español referentes a desastres naturales con cuatro categorías balanceadas: *huracanes, inundaciones, sismos y forestal*. Tiene 10 documentos de entrenamiento y 50 de prueba en cada categoría. Al igual que con las colecciones anteriores, el conjunto de entrenamiento fue dividido para adaptar

los conjuntos al sistema semisupervisado. La dificultad de este corpus radica en que las categorías *huracán* e *inundaciones* poseen un vocabulario con muchos traslapes por lo que los resúmenes deberán capturar información que permita diferenciar una categoría de la otra. La colección fue preprocesada convirtiendo todas las letras a minúsculas, se retiraron signos no alfabéticos y se hizo una separación por oraciones semejante a R8. No se retiraron palabras vacías.

5.2.3. Meter

Esta colección es utilizada para evaluar sistemas de detección de plagio. Está formada por un conjunto de cablegramas y uno de noticias derivadas de ellos divididos en 2 categorías muy amplias, *showbiz* (noticias del espectáculo que incluyen música, tv, cine entre otras) y *court* (noticias sobre la corte que incluyen robos, asesinatos, raptos, etc.). Debido a que no es un corpus propio de la tarea, se consideró que los cablegramas podrían emplearse como documentos de entrenamiento y las noticias derivadas como documentos de prueba. La distribución final puede verse en la tabla siguiente. Del mismo modo que en los otros conjuntos de datos, el subconjunto de entrenamiento fue separado en dos, uno etiquetado usado para entrenar al clasificador inicial y un conjunto no etiquetado al que se le retiraron las etiquetas. La complejidad de este corpus radica en la diversidad de las temáticas de los documentos de cada categoría ya que, como se mencionó, cada uno abarca un conjunto de subcategorías con vocabularios y distribuciones propios. La colección fue preprocesada de la misma forma que R8.

	Conjunto	
Categoría	Entrenamiento	Prueba
court	660	770
showbiz	111	175
Total	771	945

Tabla 5.1: Distribución de los documentos en las categorías de Meter

5.2.4. Wiki

Se trata del corpus más pequeño de todos, está constituido por 5 categorías relacionadas con ciencias computacionales: *object-oriented programmation (oop), page-*

rank algorithm (pagerank), dynamic programation (dynamic), bayes theorem (bayes) y *vectorial model (vector)*. El corpus fue creado para evaluar sistemas de detección automática de plagio por lo que algunos documentos han sido derivados y reformulados de otros. Esto es, una parte de su contenido ha sido extraído directamente de otro documento y modificado para evitar el plagio. La reformulación de los documentos que han sido derivados de otros presenta diferentes medidas que van desde *nula* (se han insertado fragmentos del documento original sin modificar) hasta *alta* (un gran número de modificaciones han sido realizadas sobre los fragmentos de contenido extraído del documento original). Cada categoría está integrada por un documento procedente de la Wikipedia y 19 artículos referentes al mismo tema incluidos documentos derivados (con diferentes grados de reformulación) y no derivados. Estos últimos comparten el contenido temático pero proceden de fuentes distintas. En los experimentos realizados, los documentos de Wikipedia siempre fueron usados como conjunto etiquetado en tanto que para formar los conjuntos no etiquetado y de prueba se probaron dos esquemas. En el primero de ellos se utilizaron los documentos derivados como conjunto no etiquetado y los documentos no derivados como conjunto de prueba; en el segundo esquema se invirtieron los conjuntos . La tabla 5.2 muestra la distribución de los documentos en las diferentes categorías.

	Tipos de Documentos		
Categoría	Originales	Derivados	No Derivados
bayes	1	13	6
dynamic	1	12	7
oop	1	10	9
pagerank	1	10	9
vector	1	12	7

Tabla 5.2: Distribución de los documentos en la colección Wiki

5.3. Experimentos

Una vez que el método ha sido implementado y procesados los conjuntos de datos, fue realizada una serie de experimentos para evaluar el desempeño del método en diversas condiciones. Los experimentos realizados tienen como objetivo analizar el comportamiento del sistema bajo diferentes condiciones para probar que se comporta

de acuerdo con las hipótesis planteadas. Las condiciones particulares en las que fueron llevados a cabo se especifican en las secciones siguientes. En todos los experimentos, el conjunto de datos fue dividido en un conjunto etiquetado, un conjunto no etiquetado y uno de prueba. El conjunto etiquetado es sumarizado, usado para entrenar al clasificador y crece al integrar resúmenes de documentos confiables en cada iteración. El conjunto no etiquetado es clasificado y decrece en cada iteración proporcionando documentos confiables que son agregados al conjunto etiquetado. El conjunto de prueba se mantiene en todo momento y es usado para evaluar el desempeño del clasificador entrenado en cada iteración.

5.3.1. Casos Base

La evaluación de un sistema semisupervisado no está estandarizada debido a que cada método presenta características particulares como la cantidad de documentos etiquetados y no etiquetados que deben proporcionarse al sistema, el tiempo de convergencia o las condiciones de paro.

Debido a que el objetivo de los experimentos es evaluar el desempeño del sistema cuando se incorporan resúmenes automáticos, el caso base más cercano con el que se pueden realizar comparaciones objetivas consiste en seguir el mismo algoritmo propuesto con la condición de no utilizar resúmenes automáticos. Este caso corresponde a un *self-training* tradicional implementado con los módulos antes descritos y las mismas condiciones (ver secciones siguientes) que cada uno de los experimentos en los que se utilizan resúmenes automáticos. En todos los experimentos reportados se analiza el caso base consistente en no usar resúmenes automáticos.

Adicionalmente, se consideró como otro punto de referencia para evaluar el desempeño del sistema al clasificador entrenado con el conjunto etiquetado inicialmente, que es a partir del cuál se mide el efecto de integrar información procedente del conjunto no etiquetado, lo que también determina la efectividad del esquema de selección de documentos confiables y del algoritmo semisupervisado en general.

5.3.2. Documentos Iniciales

Una característica que determina la necesidad de utilizar un sistema semisupervisado es el reducido número de documentos etiquetados con los que se cuenta inicialmente. Aunque tampoco existe una convención acerca de lo que *número reducido*

significa, puede observarse que se trata de un término muy dependiente del contexto, así, en trabajos como [42] se usan 100 muestras iniciales cuando se trata de clasificar oraciones, en [25] se usan 5 para noticias, 20 en [30] también para noticias o entre 10 y 100 en [10] para noticias en inglés y entre 1 y 10 para noticias en español.

Debido a esta ambigüedad en los trabajos reportados y teniendo como antecedente que los resúmenes automáticos son efectivos para mejorar la clasificación aún en conjuntos de entrenamiento que cuentan con no más de 10 documentos por categoría (capítulo anterior) se recurrió a utilizar la menor cantidad posible de muestras en función del tamaño del corpus empleado. Así en un corpus numeroso como R8 se usaron 1 y 5 documentos de cada categoría, en Meter 1, 3 y 5; y solamente 1 en los corpora de Desastres Naturales y Wiki.

Finalmente, debido a que el crecimiento del conjunto etiquetado es función de los documentos que están etiquetados inicialmente, en todas las situaciones tales documentos fueron elegidos manualmente tratando de que fuesen representativos de la categoría a la que pertenecen.

5.3.3. Condición de Paro del Sistema

La condición que debe cumplirse para que el sistema detenga su funcionamiento tampoco es un estándar y se determina en función tanto de la cantidad de documentos no etiquetados disponibles, el tiempo del que se dispone para hacer crecer el conjunto etiquetado y la calidad del clasificador con respecto a la versión anterior (pues no tiene sentido mantener al sistema funcionando mientras su desempeño empeora).

En los experimentos reportados y debido a las limitaciones en el tamaño de algunos corpora, tres condiciones de paro fueron implementadas:

- No existen más documentos no etiquetados o no son suficientemente confiables.
- La evaluación del clasificador sobre el conjunto de prueba empeora con respecto a la versión anterior.
- Se alcanzan 20 iteraciones (10 en caso de algunos corpora con pocos documentos).

Al cumplirse cualquiera de ellas, el sistema se detiene.

5.3.4. Documentos integrados en cada Iteración

Otra de las variantes en las implementaciones de *self-training* es el número de documentos que serán integrados al conjunto etiquetado en cada iteración. Este parámetro está determinado normalmente por el criterio de selección de dichos documentos en la clasificación realizada que es a su vez función del clasificador.

En la implementación propuesta, el criterio de selección de documentos confiables ofrece dos alternativas: seleccionar todos los documentos que cumplan la primera etapa del criterio, es decir, todos aquellos cuya distancia al clúster de la categoría con la que fueron etiquetados sea mínima con respecto a los otros clústers o bien, seleccionar entre ellos los k documentos más cercanos al clúster de su clase. Debido a que en el primer caso no se tiene control sobre el número de documentos integrados, en los experimentos realizados el criterio de selección permite integrar hasta k documentos por categoría. Cabe señalarse que k es una cota superior en el número de documentos que se integran para cada categoría, si no es posible seleccionar tal cantidad, cualquier número de documentos confiables menor a k será seleccionado.

Un valor de $k = 1, 5$ y 10 es usado en los distintos experimentos.

5.3.5. Resultados

Los siguientes resultados reportan el micropromedio de la *Medida F1* de la clasificación sobre el conjunto de prueba en cada iteración del sistema usando resúmenes con extensiones entre 10 % y 90 % del tamaño del documento original. También se muestran los resultados del caso base en cada iteración. El resultado mostrado en la iteración 0 corresponde al otro caso base: cuando se evalúa el clasificador entrenado únicamente con los documentos etiquetados disponibles inicialmente.

R8

Los experimentos realizados sobre R8 incluyen dos variantes en el número de documentos etiquetados inicialmente, en el primer caso se emplea únicamente un documento de cada categoría, en el segundo experimento se utilizan 5 de cada clase. El número de documentos integrados al conjunto de entrenamiento en cada iteración se fijó en 1, 5 y 10 documentos con el propósito de observar las variaciones en el comportamiento del sistema.

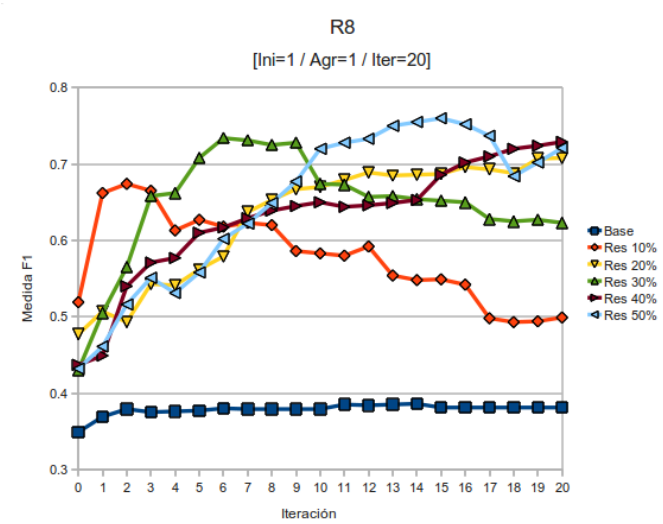

Figura 5.3: Micropromedio de la Medida F1 de los resultados de la clasificación del conjunto de prueba en cada una de las 20 iteraciones (Iter=20) del sistema semisupervisado en R8 usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contenía únicamente un documento de cada categoría (Ini=1) y se agregó hasta uno más por categoría en cada iteración (Agr=1).

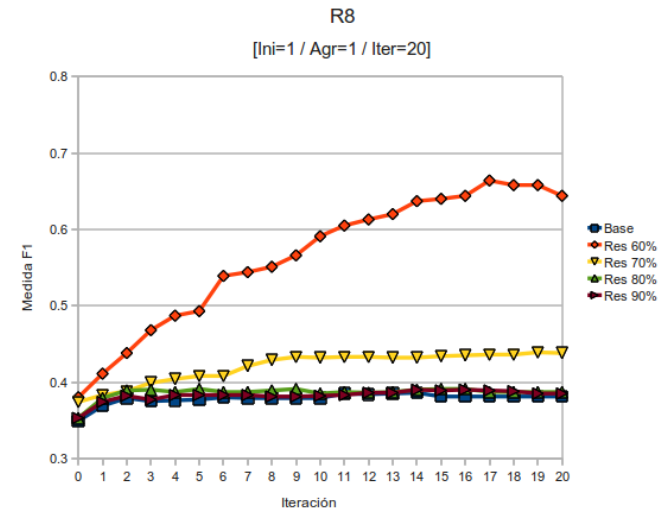

Figura 5.4: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

Las figuras 5.3 y 5.4 muestra los resultados del método en R8 cuando se emplea únicamente 1 documento de cada categoría en el conjunto de entrenamiento inicial y se agrega hasta 1 documento de cada categoría en cada iteración.

Puede observarse que el clasificador del caso base mejora su desempeño únicamente en 0.037 con respecto al caso inicial en 20 iteraciones.

Por otro lado, las ventajas de incorporar resúmenes automáticos en el sistema de clasificación supervisada pueden observarse desde el clasificador inicial de los resúmenes más pequeños: En la columna correspondiente a los resúmenes de 10 %, el clasificador inicial, entrenado con los resúmenes de 1 documento de cada categoría (8 en total) tiene una medida F1 superior a la obtenida cuando no se emplean resúmenes automáticos. En lo sucesivo, el uso de resúmenes demuestra mejorar el desempeño del sistema notablemente.

Mientras en el caso base, el desempeño del sistema mejora muy lentamente al integrar nuevos documentos al conjunto de entrenamiento, cuando se utilizan resúmenes automáticos para entrenar al clasificador el sistema mejora su desempeño notablemente en pocas iteraciones: para los resúmenes de 10 % el clasificador mejora su medida F1 en 0.14 en la primera iteración, los resúmenes de 20 % lo hacen en 0.22 para la iteración 20, los de 30 % mejoran en 0.3 para la sexta iteración mientras los de 40 % pasan de 0.44 a 0.73 en 20 iteraciones; los resúmenes de 50 % alcanzan una mejora semejante mientras los de 60 % mejoran hasta en 0.28 su evaluación. Finalmente los resúmenes entre 70 % y 90 % obtienen mejoras menos significativas con respecto al caso base debido a que su contenido se aproxima cada vez más al de los documentos completos.

Se puede observar que en algunos casos después de mejorar, la evaluación del clasificador tiende a decaer, aunque debido al número de iteraciones no puede apreciarse lo que ocurre si el sistema continúa funcionando de este modo. Sin embargo, un cambio en el número de documentos agregados al conjunto etiquetado en cada iteración muestra el desempeño del sistema al agregar más información procedente de los documentos que inicialmente no estaban etiquetados lo que puede verse como una continuación de los resultados de esta primera serie de experimentos. Las figuras 5.5 y 5.6 muestran el comportamiento del sistema cuando se agregan 5 documentos confiables en cada iteración mientras las figuras 5.7 y 5.8 presentan los resultados agregando 10 documentos confiables en cada iteración.

Las gráficas de las figuras 5.5 y 5.6 muestran un comportamiento semejante a

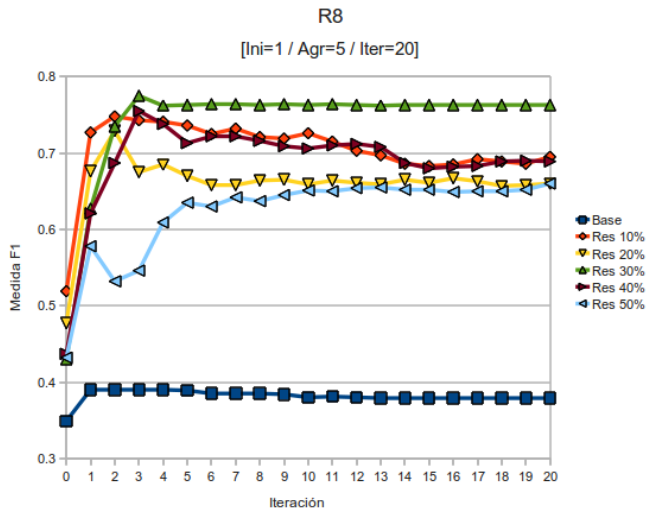

Figura 5.5: Micropromedio de la Medida F1 de los resultados de la clasificación del conjunto de prueba en cada una de las 20 iteraciones (Iter=20) del sistema semisupervisado en R8 usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contenía únicamente un documento de cada categoría (Ini=1) y se agregaron hasta cinco más por categoría en cada iteración (Agr=5).

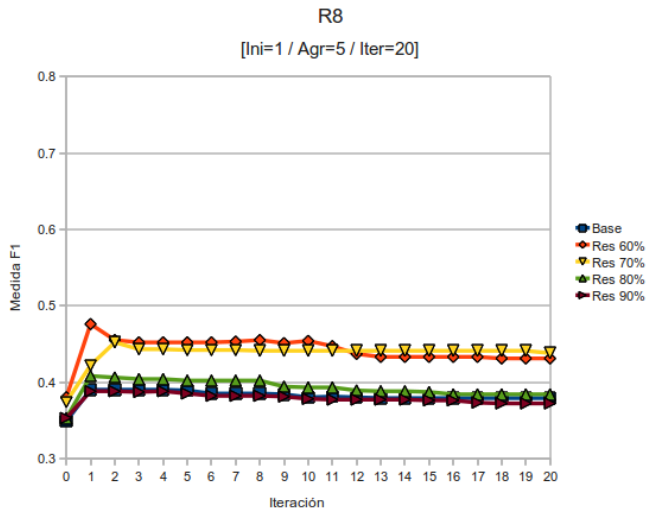

Figura 5.6: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

las de 5.3 y 5.4. Aunque puede apreciarse más claramente el momento en el que el sistema se estabiliza. En el apéndice A se muestran las tablas correspondientes a los resultados de los mismos experimentos.

Como antes se comentó, los diferentes pares de gráficas muestran acercamientos distintos del mismo fenómeno. En las figuras 5.3 y 5.4 puede observarse la etapa en que la evaluación del sistema va en crecimiento gracias a la información de los documentos que inicialmente no estaban etiquetados y que es agregada en cada iteración, mientras que en las figuras 5.5 y 5.6 puede observarse al sistema cuando su comportamiento se estabiliza y comienza a decaer. Cabe señalarse que en ambos casos, los resúmenes con extensión menor al 60 % tienen un comportamiento creciente y mejoran el desempeño del clasificador notablemente, mientras los resúmenes entre el 60 % y el 90 % tienden a comportarse como el caso base debido a que integran al resumen la información menos relevante de cada documentos por lo que su contenido es cada vez más cercano a los documentos de los cuales han sido extraídos. Las figuras 5.7 y 5.8 muestran el comportamiento del sistema al permitir agregar hasta 10 documentos al conjunto etiquetado en cada iteración.

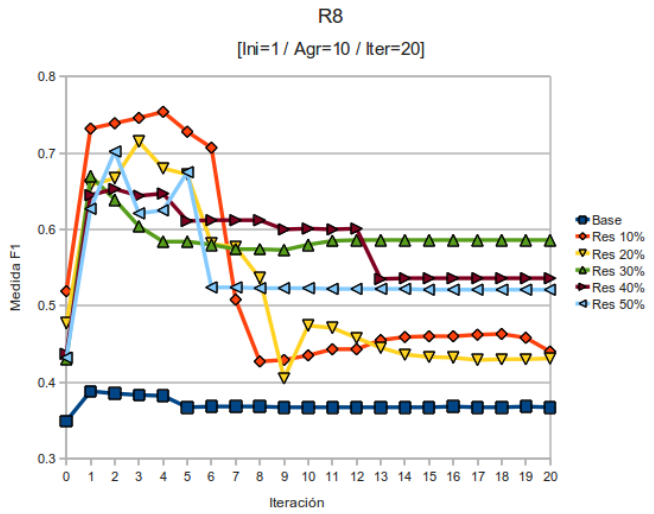

Figura 5.7: Micropromedio de la Medida F1 de los resultados de la clasificación del conjunto de prueba en cada una de las 20 iteraciones (Iter=20) del sistema semisupervisado en R8 usando resúmenes de entre 10 % **y** 50 % **. El conjunto etiquetado inicial contenía únicamente un documento de cada categoría (Ini=1) y se agregaron hasta 10 más por categoría en cada iteración (Agr=10).**

En las figuras 5.7 y 5.8 puede verse como, al agregar un mayor número de documentos al conjunto etiquetado, el comportamiento del clasificador es más rápido: en

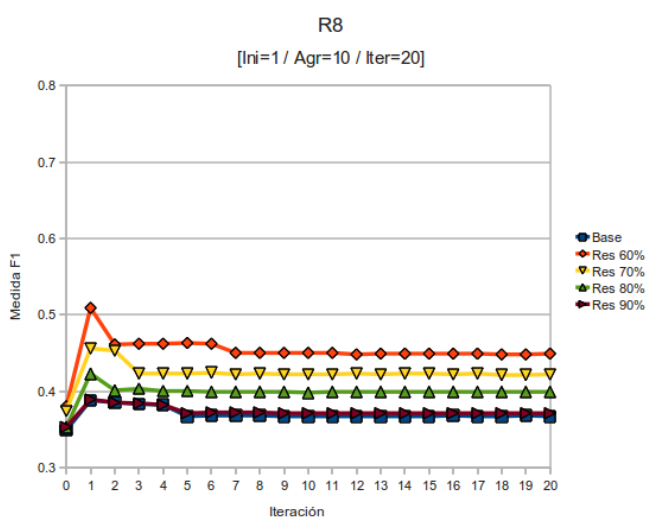

Figura 5.8: Continuación de los resultados de la figura anterior para resúmenes entre 60 % **y** 90 %**.**

pocas iteraciones crece, alcanza un máximo en su desempeño y decae. Esto también puede verse como un efecto de la relajación del criterio de selección de documentos confiables. Debido a que estos son elegidos a partir de la distancia que guardan con el clúster que representa a los documentos etiquetados de la clase a la que fueron asignados, mientras mayor es el número de documentos integrados en cada iteración mayor es la distancia que guardaran hacia el clúster, favoreciendo la formación de clústers menos homogéneos y proporcionando al clasificador información más dispersa respecto a la clase. Esta condición se propaga en cada iteración, provocando que el comportamiento del sistema decaiga después de cierto número de iteraciones.

Otros resultados obtenidos de R8 son los mostrados en las figuras 5.9 a 5.12 que muestran el comportamiento del sistema cuando se utilizan 5 documentos de cada categoría como conjunto etiquetado inicial. Puede observarse en estos casos, que aunque los resúmenes tienen un impacto positivo en el desempeño del clasificador inicial, éste no mejora cuando se agregan nuevos documentos etiquetados al conjunto de entrenamiento. Este fenómeno se comenta en la sección de discusión.

Desastres Naturales

El conjunto de entrenamiento del corpus de Desastres Naturales cuenta únicamente con 10 documento por categoría. Si un documento de cada clase es usado para

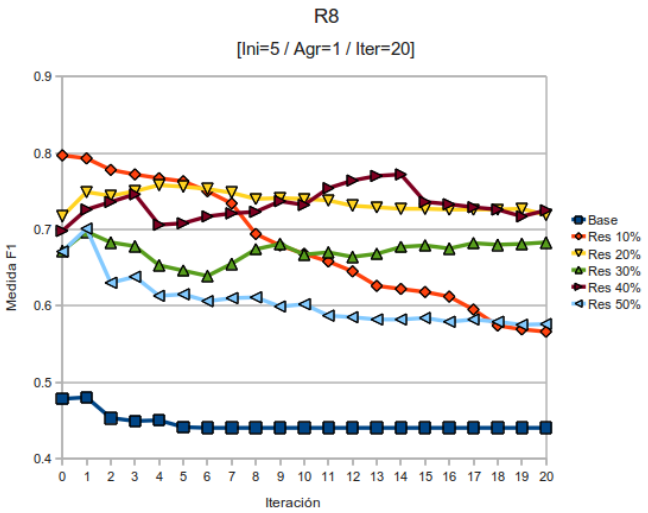

Figura 5.9: Micropromedio de la Medida F1 de los resultados de la clasificación del conjunto de prueba en cada una de las 20 iteraciones (Iter=20) del sistema semisupervisado en R8 usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contenía 5 documentos de cada categoría (Ini=5) y se agregó hasta 1 documento por categoría en cada iteración (Agr=1).

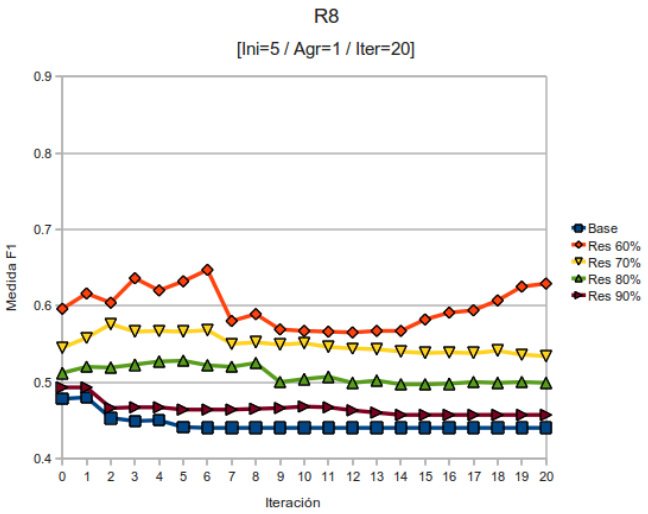

Figura 5.10: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

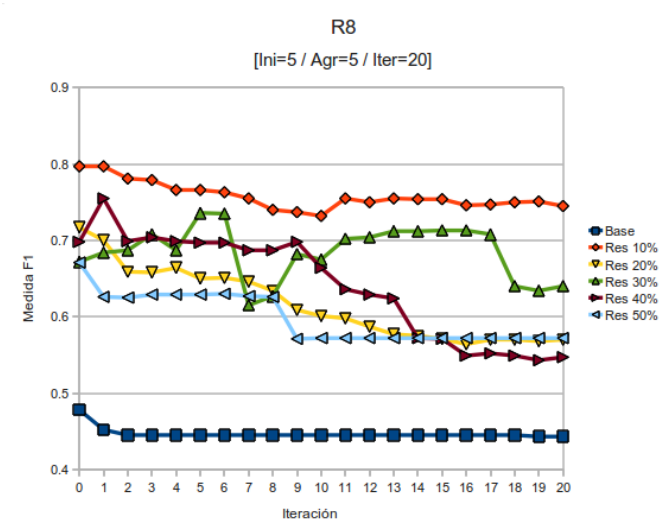

Figura 5.11: Micropromedio de la Medida F1 de los resultados de la clasificación del conjunto de prueba en cada una de las 20 iteraciones (Iter=20) del sistema semisupervisado en R8 usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contenía 5 documentos de cada categoría (Ini=5) y se permitió agregar hasta 5 documentos por categoría en cada iteración (Agr=5).

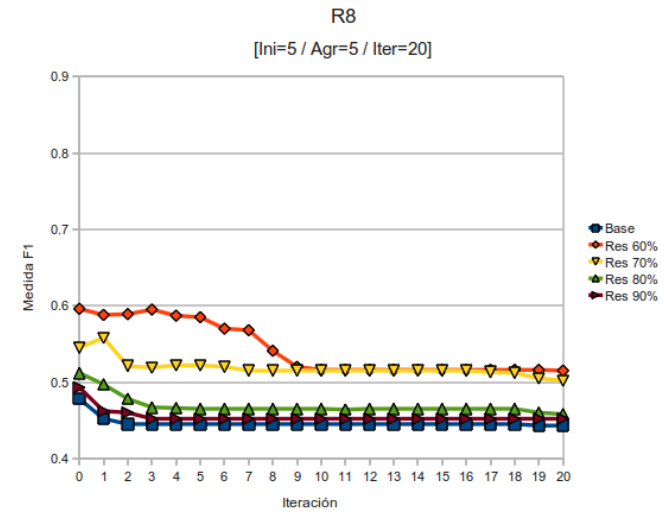

Figura 5.12: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

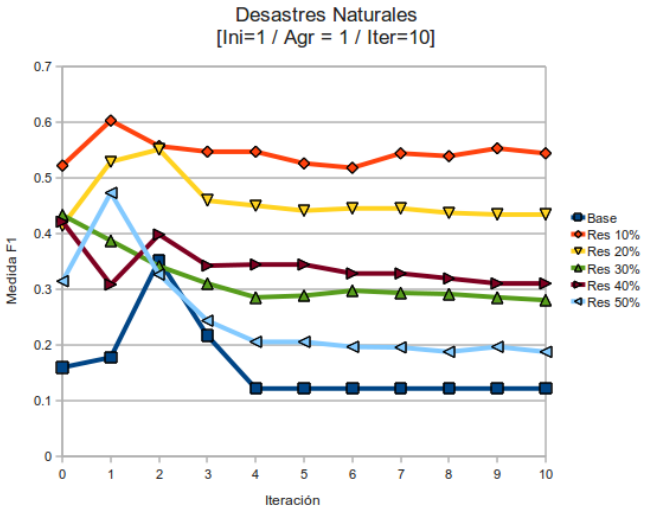

Figura 5.13: Medida F1 de la evaluación del clasificador en cada una de las 10 iteraciones (Iter=10) del sistema semisupervisado en el corpus Desastres Naturales usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 1 documento de cada categoría (Ini=1) y se permitió agregar 1 documento por categoría en cada iteración (Agr=1).

formar el conjunto etiquetado inicial, el conjunto no etiquetado contendrá 9 documentos por categoría como máximo. De modo que para este conjunto de documentos sólo podrá variarse el número de documentos que son agregados en cada iteración. Los experimentos realizados incluyen los casos en que se agrega hasta 1 y hasta 5 documentos por categoría en cada iteración. Las figuras 5.13 a 5.16 muestran los resultados obtenidos. Los casos en que la gráfica se detiene antes de completar el número de iteraciones representan los casos en los que el conjunto no etiquetado se agotó.

Las figuras 5.13 y 5.14 muestran que el efecto de los resúmenes automáticos también es notable desde que se sumarizan los documentos iniciales del conjunto etiquetado, pues para todos los casos en la iteración 0 los resultados de la clasificación cuando se utilizan resúmenes son siempre superiores a los del caso base. También puede observarse que el desempeño del clasificado no es muy eficiente pues se mantiene por debajo de 0.7 aunque en todos los casos en que se usan resúmenes se alcanza una evaluación mejor al máximo obtenido en el caso base.

En este conjunto de resultados puede observarse que resúmenes de menor tamaño tienen un impacto mayor en el clasificador y mientras el tamaño del resumen crece el comportamiento se aproxima al del caso base. Cuando se emplea un documento por

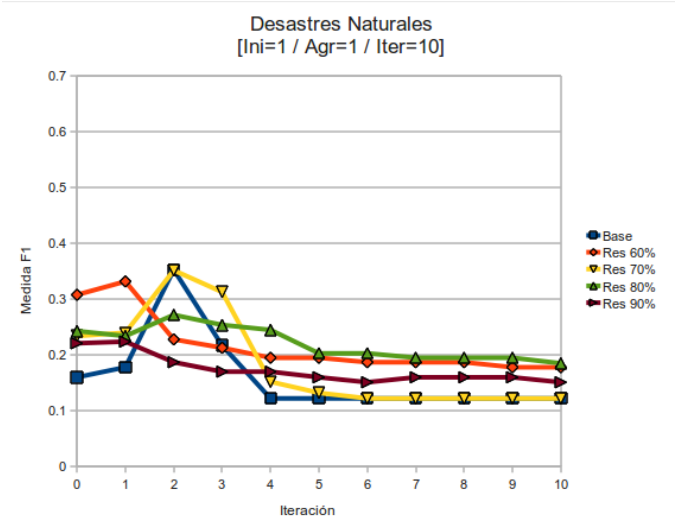

Figura 5.14: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

categoría en el conjunto inicial y se permite agregar hasta 1 más en cada categoría por iteración, la medida F1 del clasificador pasa de 0.52 a 0.60 en la primera iteración con resúmenes de 10 % de la extensión del documento y se mantiene con una mejora de 0.022 para la décima iteración lo que indica que en realidad se está obteniendo información útil de los documentos que originalmente no estaban etiquetados. En los resúmenes de 20 % se pasa de 0.41 a 0.55 en 3 iteraciones mientras que para los resúmenes de tamaño mayor el clasificador tiende a decaer aunque los resúmenes menores al 60 % tienen siempre una pico mayor al del caso base.

Este hecho proporciona una medida de la cantidad de información relevante (para la clasificación) contenida en los documentos de esta colección de noticias. Hablamos de que únicamente un máximo de 20 % del documento es información que permite discriminar a los documentos de las diferentes categorías mientras el resto del documento contiene información que no es relevante o que introduce ruido al clasificador.

Cuando se permite que el criterio de selección agregue hasta 5 documentos por categoría en cada iteración se observan mejoras más significativas aunque en general el comportamiento del sistema es semejante. Las figuras 5.15 y 5.16 muestran las gráficas de estos experimentos.

Las figuras 5.15 y 5.16 muestran que con resúmenes de 10 % se mejora el desempeño del clasificador inicial de 0.52 a 0.66 en solo tres iteraciones lo que representa una

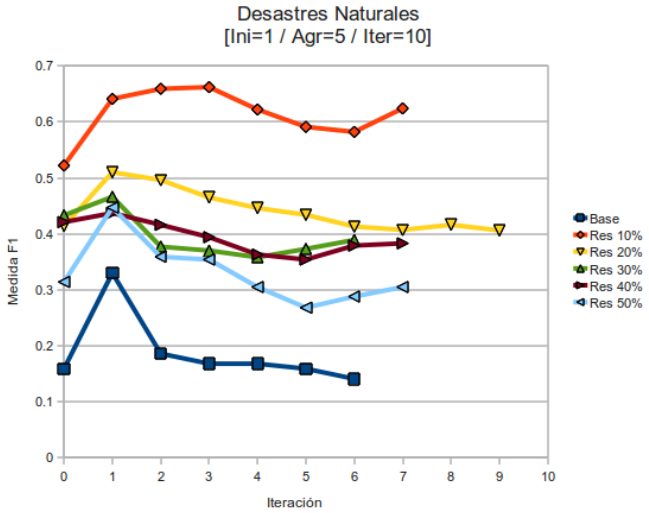

Figura 5.15: Medida F1 de la evaluación del clasificador en cada una de las 10 iteraciones (Iter=10) del sistema semisupervisado en el corpus Desastres Naturales usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 1 documento de cada categoría (Ini=1) y se permitió agregar hasta 5 documentos por categoría en cada iteración (Agr=5).

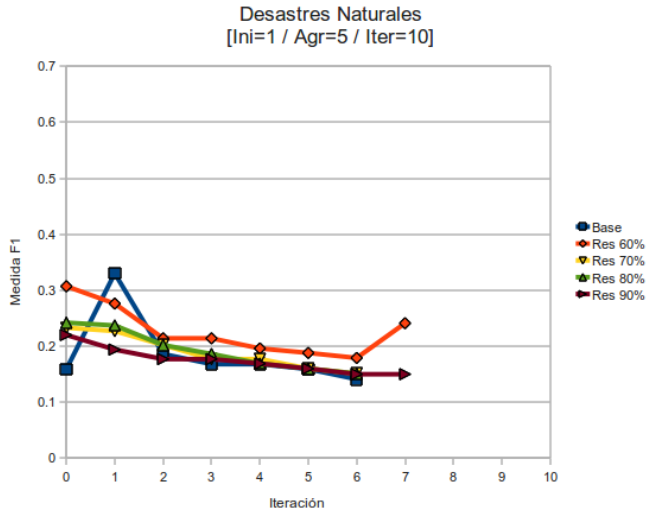

Figura 5.16: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

mejora de 0.33 con respecto al máximo alcanzado por el sistema cuando no se emplean resúmenes automáticos. Con resúmenes de 20 % el sistema pasa de una evaluación de 0.41 a 0.51 en la primera iteración mientras que cuando se emplean resúmenes más grandes el clasificador mejora levemente en las primeras iteraciones y decae para las siguientes. Cabe señalarse que en todos los casos de este experimento se agotaron los documentos del conjunto no etiquetado, razón por la cual las gráficas están truncadas después de cierto número -diferentes para los diferentes tamaños de resúmenes- de iteraciones.

Un detalle importante de estos experimentos es el comportamiento del caso base que a pesar de incorporar documentos al conjunto etiquetado en cada iteración, tiene una pequeña mejora en la primera iteración y luego se mantiene en un valor determinado. Este fenómeno se justifica si se considera que el clasificador inicial es muy malo ya que tiene una exactitud del 25 % lo cual, en un conjunto balanceado con 4 categorías significa que está clasificando a todos los documentos en una sola clase; luego entre ellos selecciona a los documentos confiables y los integra al conjunto etiquetado a pesar de que sus etiquetas sean erróneas por lo que el clasificador de la siguiente iteración estará sesgado y nuevamente clasificará a todos los documentos en la misma categoría propagando la condición en las iteraciones siguientes.

Como observación final, debe señalarse una vez más que el comportamiento del sistema está en función directa de la calidad de los documentos que son etiquetados manualmente y de la cantidad de documentos no etiquetados de los que se dispone para integrar al conjunto etiquetado y presentan las características necesarias para aprobar el criterio de selección. La discusión sobre esta observación se expone en la sección 5,4. Así, si con los documentos iniciales es posible realizar una buena clasificación, el desempeño del clasificador en cada iteración tiende a mejorar; en cambio, si no es posible realizar una buena clasificación desde el principio el mal comportamiento se mantiene ya que los documentos agregados tendrán una elevada probabilidad de estar mal etiquetados.

Meter

Meter es un corpus que contiene únicamente dos grandes categorías. Cada una de ellas está subdividida en categorías más específicas pero en los experimentos no se considera esta división con el propósito de analizar la capacidad del sistema para incorporar información de estas subcategorías como parte de la categoría mayor.

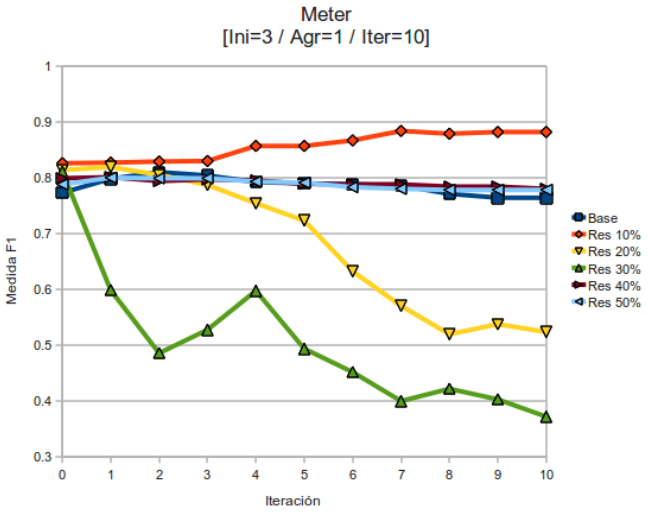

Figura 5.17: Gráficas de la Medida F1 del clasificador en cada una de las 10 iteraciones (Iter=10) del sistema semisupervisado en el corpus Meter usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 3 documentos de cada categoría (Ini=3) y se permitió agregar 1 documento por categoría en cada iteración (Agr=1).

Los experimentos realizados incluyen tres números distintos de documentos iniciales, 1, 3 y 5 por categoría (lo que representa un conjunto inicial de solamente 2, 6 y 10 documentos). En cada caso se analizó el efecto de integrar 1, 5 y 10 documentos confiables en cada iteración. En esta sección se muestran únicamente los resultados para el caso en que el conjunto inicial contuvo solo 3 documentos por categoría ya que en general, el comportamiento del sistema a través de las iteraciones es semejante cuando se emplean diferentes cantidades de documentos inicialmente etiquetados y es más variable para los casos en los que se agrega un número distinto de documentos en cada iteración. Los gráficas de las figuras 5.17 a 5.22 muestran la evaluación del clasificador en cada una de las 10 iteraciones cuando se usan 3 documentos por categoría en el conjunto etiquetado y los casos en que se agregan 1, 5 y 10 documentos adicionales en cada iteración.

En este corpus el efecto de los resúmenes no es tan notable como en los otros, sin embargo, puede observarse que en todos los casos, la evaluación del clasificador inicial (iteración 0) es superior al caso en que no se emplean resúmenes (Apéndice C). Sin embargo, con el paso de las iteraciones el comportamiento del clasificador tiende a mantenerse constante salvo algunas excepciones en que el clasificador decae drásticamente y solo una en el caso de los resúmenes de 10 % cuando se agregan hasta

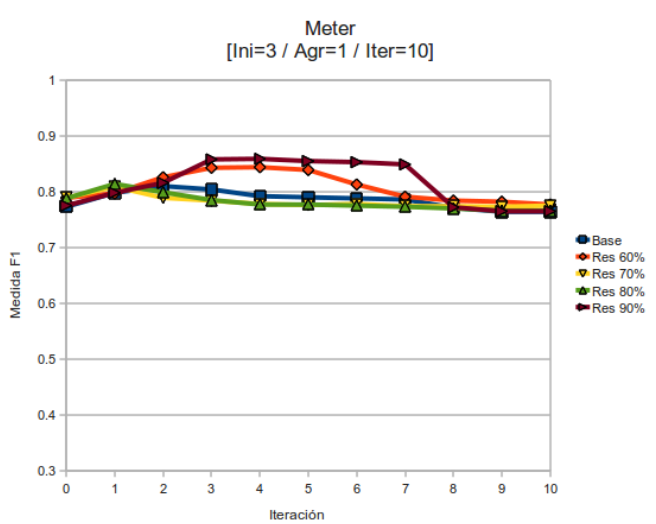

Figura 5.18: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

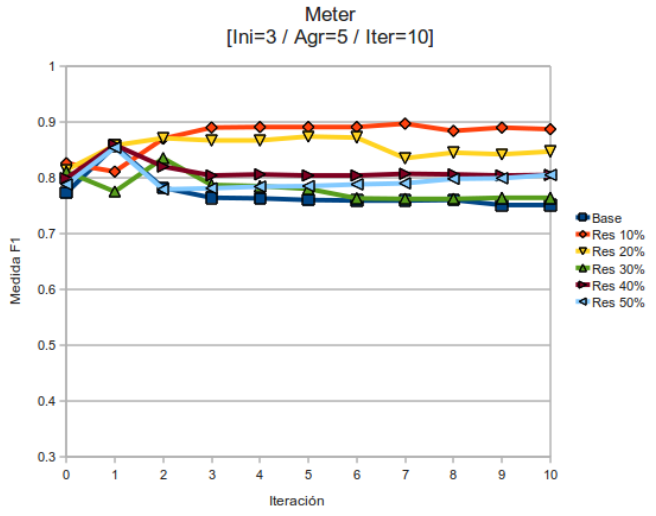

Figura 5.19: Gráficas de la Medida F1 del clasificador en cada una de las 10 iteraciones (Iter=10) del sistema semisupervisado en el corpus Meter usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 3 documentos de cada categoría (Ini=3) y puede agregar 5 documentos por categoría en cada iteración (Agr=5).

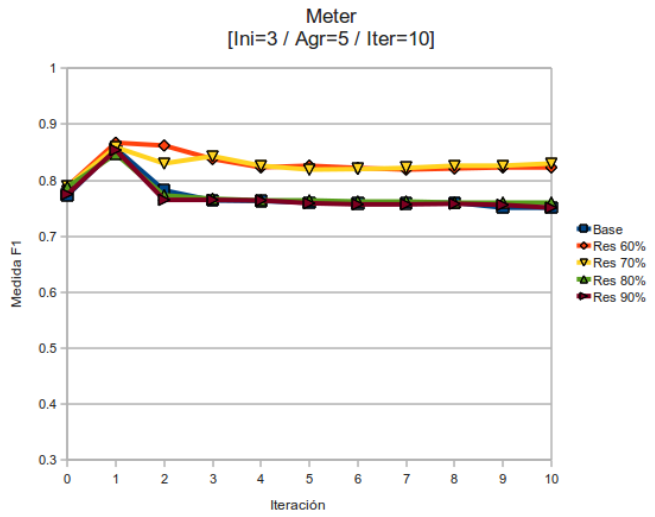

Figura 5.20: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

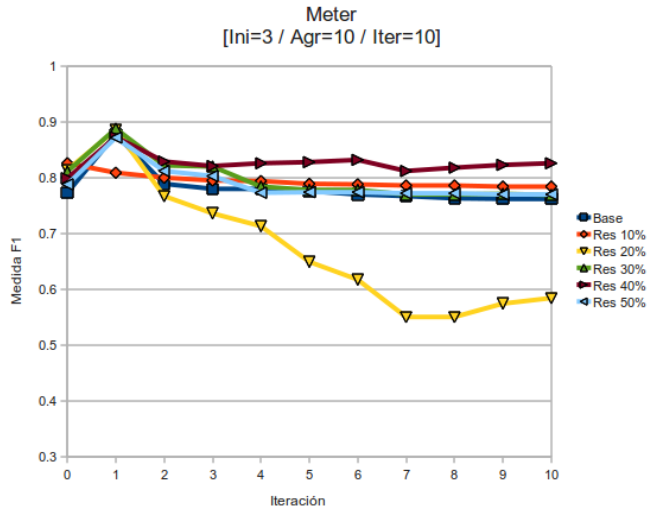

Figura 5.21: Gráficas de la Medida F1 del clasificador en cada una de las 10 iteraciones (Iter=10) del sistema semisupervisado en el corpus Meter usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 3 documentos de cada categoría (Ini=3) y puede agregar 10 documentos por categoría en cada iteración (Agr=10).

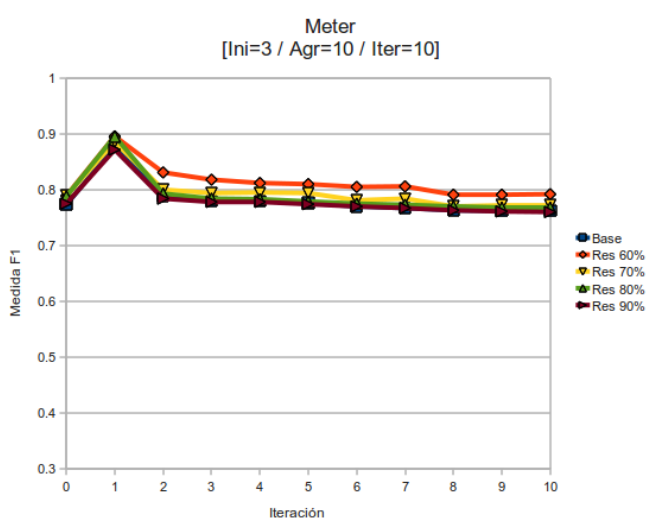

Figura 5.22: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

1 y hasta 5 documentos por iteración en que el clasificador mejora considerablemente.

El caso más notable se presenta cuando se permite agregar hasta 5 documentos por iteración (figuras 5.19 y 5.20) ya que el uso de resúmenes permite al sistema estabilizarse con una evaluación superior a la del caso base, particularmente los resúmenes de 10 % y 20 % pasan de 0.82 a 0.9 en 7 iteraciones y de 0.81 a 0.874 en 5 respectivamente.

Cuando se permite agregar únicamente un documento por categoría la evaluación de los resúmenes de 10 % mejora en cada iteración mientras la de resúmenes de 20 % y 30 % decaen severamente en tanto que las demás se mantienen constantes. De igual modo, cuando se permite agregar hasta 10 documentos por iteración los resúmenes de 20 % tienen una caída mientras los demás se mantienen. La justificación de este fenómeno se encuentra al considerar que la colección contiene únicamente 2 grandes categorías subdivididas jerárquicamente, es decir, los casos en que el clasificador decae pueden deberse a que se agregó al conjunto etiquetado algún o algunos documentos que sesgaron al clasificador (debe recordarse que el conjunto etiquetado inicial contiene solo 6 documentos) propagando el error en las iteraciones siguientes.

Wiki

El conjunto de documentos Wiki es el mas reducido de todos. En consecuencia muestra la limitante del número de documentos no etiquetados disponibles por lo que

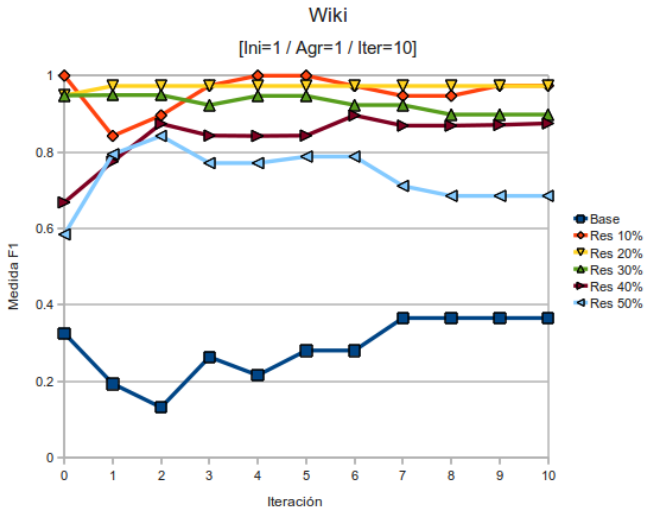

Figura 5.23: Medida F1 de la evaluación del sistema semisupervisado en cada una de las 10 iteraciones (Iter=10) con el corpus Wiki (el conjunto no etiquetado está formado por los documentos *derivados*) usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 1 documento de cada categoría (Ini=1) y se puede agregar 1 documento más en cada iteración (Agr=1).

en los experimentos realizados se cumple la condición de paro por agotamiento de los documentos disponibles. En la sección 5.2.4 puede verse la división este conjunto.

Dos variaciones sobre los experimentos fueron realizadas. En ambas se mantuvo el conjunto originalmente etiquetado que contenía únicamente el documento de Wikipedia de cada categoría. En el primer caso se empleo como conjunto no etiquetado a los documentos derivados del artículo de Wikipedia con algún grado de reformulación (nula, media o alta) mientras los documentos que no son derivados fueron empleados como conjunto de prueba. En el segundo caso los papeles de estos conjuntos se invirtieron. En los dos experimentos se probaron los casos en que 1 o 5 documentos confiables eran agregados en cada iteración. Los resultados del primer caso se muestran en las gráficas de las figuras 5.23 a 5.26 mientras las figuras 5.27 a 5.30 muestran los resultados del segundo de ellos.

Se discutirá primero el caso en que los documentos derivados del artículo de wikipedia fueron empleados como conjunto no etiquetado mientras los documentos no derivados se usaron como conjunto de evaluación.

El efecto positivo de los resúmenes automáticos puede observarse en todas las gráficas desde el clasificador inicial. En el caso que se permite agregar hasta 1 documento en cada iteración (figuras 5.23 y 5.24) se observa que para todos los casos

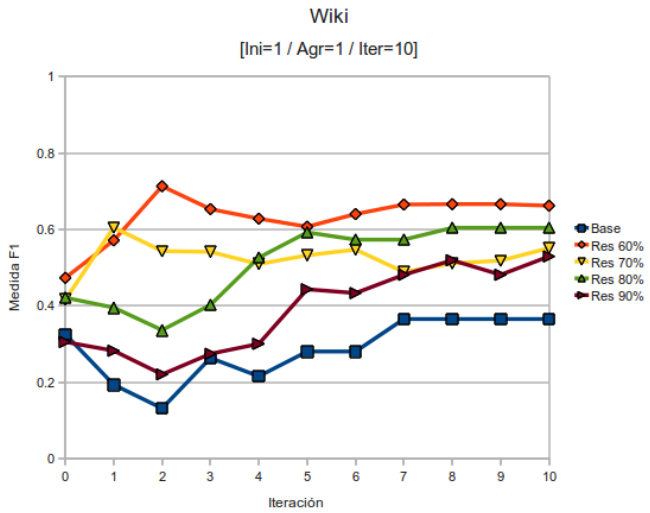

Figura 5.24: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

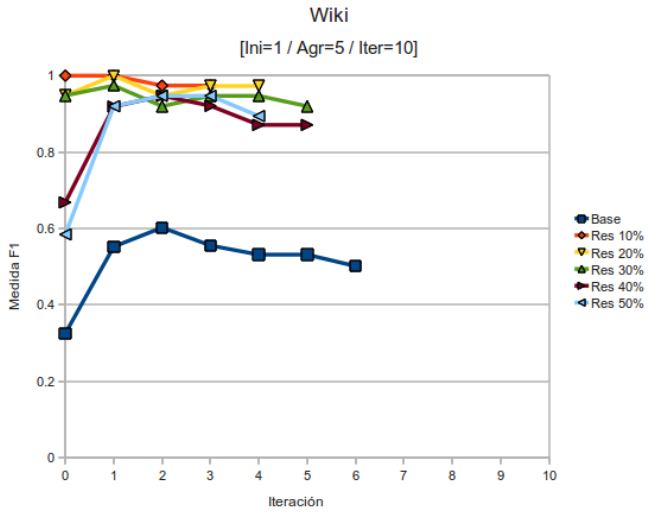

Figura 5.25: Medida F1 de la evaluación del sistema semisupervisado en cada una de las 10 iteraciones (Iter=10) con el corpus Wiki (el conjunto no etiquetado está formado por los documentos *derivados*) usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 1 documento de cada categoría (Ini=1) y se pueden agregar 5 documentos por clase en cada iteración (Agr=1).

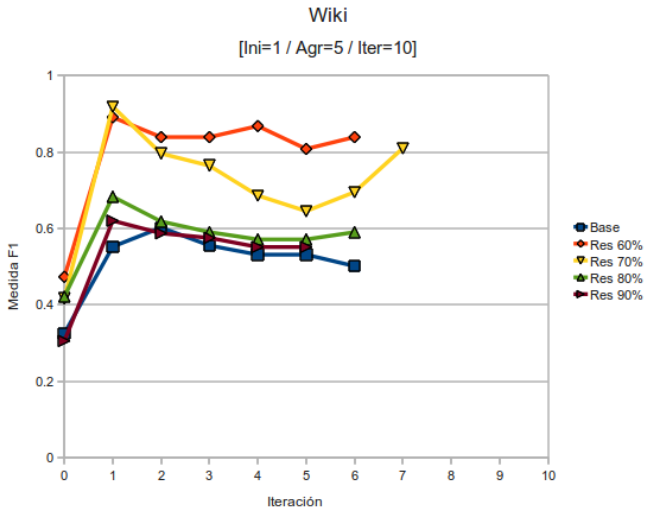

Figura 5.26: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

el desempeño del sistema cuando se emplean resúmenes es superior al del caso base. Particularmente los resúmenes de 10 %, 20 % y 30 % muestran una excelente evaluación en cada iteración por lo que se estabilizan en valores de la medida F1 cercanos a 1. Los resúmenes de 40 %, 50 % y 60 % tienen un buen desempeño y en general su comportamiento es primero creciente, luego estable y en el caso de los resúmenes de 50 % tiene una ligera decaída. Finalmente los resúmenes de tamaño superior muestran un comportamiento que tiende a aproximarse al del caso base, con una ligera decaída y una recuperación gradual aunque siempre con una medida F1 por debajo de 0.6.

Una situación que merece especial atención se presentan cuando alguno de los clasificadores alcanza una medida F1 de 1 y luego decae ligeramente. Esta situación parece extraña pues puede suponerse que si el clasificador tienen una evaluación perfecta en un instante dado no tiene porqué decaer, sin embargo debe señalarse que esta evaluación corresponde al conjunto de prueba que es distinto del conjunto no etiquetado sobre el cuál la evaluación del clasificador es generalmente distinta. Así, es posible que el clasificador tenga una evaluación de 1 en el conjunto de prueba en un instante dado pero no clasifique perfectamente el conjunto no etiquetado del cual se extraerán documentos para enriquecer al clasificador en la iteración siguiente. Como el conjunto de prueba cambia, el clasificador también lo hace y su evaluación puede mostrar variaciones, aunque como se ha visto, se mantiene en un rango de valores

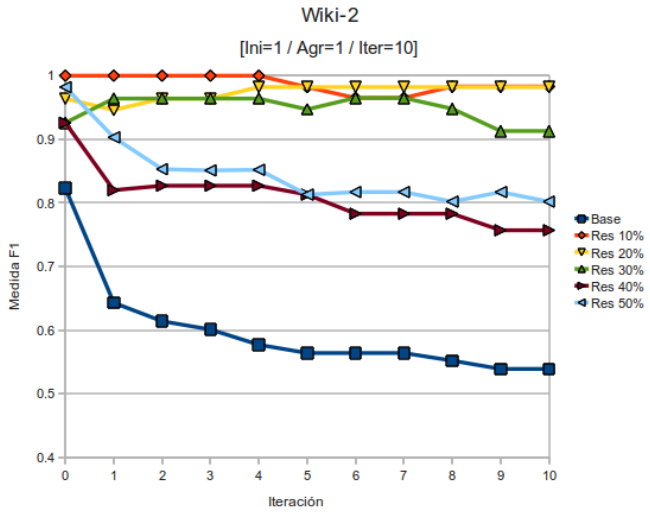

Figura 5.27: Medida F1 de la evaluación del sistema semisupervisado en cada una de las 10 iteraciones (Iter=10) con el corpus Wiki (el conjunto no etiquetado está formado por los documentos *no derivados*) usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 1 documento de cada categoría (Ini=1) y se puede agregar 1 documento más en cada iteración (Agr=1).

cercano.

Cuando se permite agregar hasta 5 documentos de cada categoría por iteración (figuras 5.25 y 5.26) el sistema muestra un comportamiento semejante aunque la subida y decaída ocurren en un menor número de iteraciones debido a que la cantidad de información que se proporciona al clasificador es mayor en cada iteración. En esta situación el sistema se detiene debido al agotamiento del conjunto no etiquetado.

Las gráficas de las figuras 5.27 a 5.30 muestran los resultados del caso cuando los documentos que no están derivados del artículo de Wikipedia fueron empleados como conjunto no etiquetado. Puede observarse que en general el comportamiento es semejante aunque muestran una decaída más pronunciada al pasar las iteraciones que se observa desde el caso base. Sin embargo, una vez más el efecto de los resúmenes queda manifiesto ya que en todos los casos el desempeño del clasificador cuando incluye resúmenes automáticos es superior al baseline.

El hecho de que la caída sea más pronunciada cuando se usa esta partición del corpus se debe a la naturaleza de los documentos ya que ahora los documentos más parecidos al conjunto etiquetado son empleados para evaluar el sistema mientras que el conjunto no etiquetado está formado por documentos que aunque comparten contenido temático presentan mayores diferencias con respecto a los documentos etiquetados.

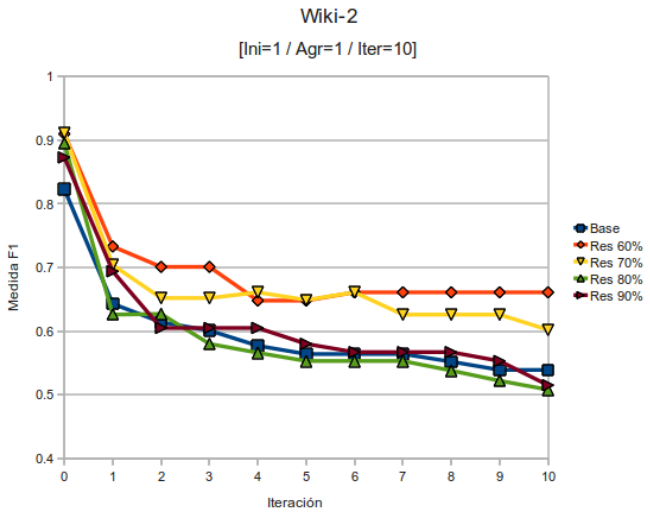

Figura 5.28: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

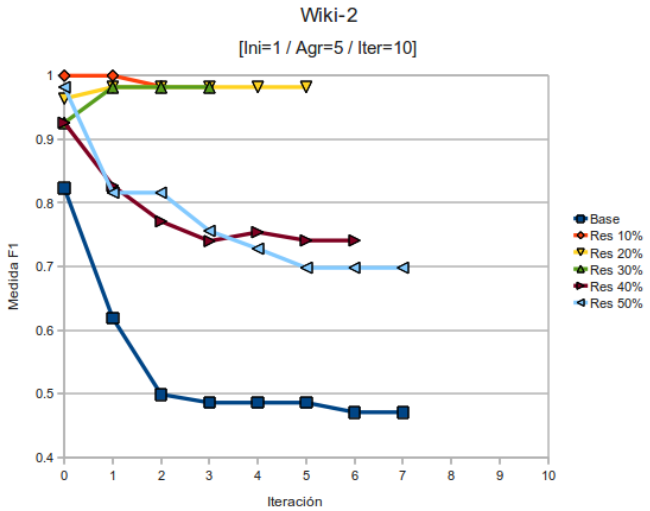

Figura 5.29: Medida F1 de la evaluación del sistema semisupervisado en cada una de las 10 iteraciones (Iter=10) con el corpus Wiki (el conjunto no etiquetado está formado por los documentos *no derivados*) usando resúmenes de entre 10 % y 50 % . El conjunto etiquetado inicial contiene 1 documento de cada categoría (Ini=1) y se pueden agregar 5 documentos por clase en cada iteración (Agr=1).

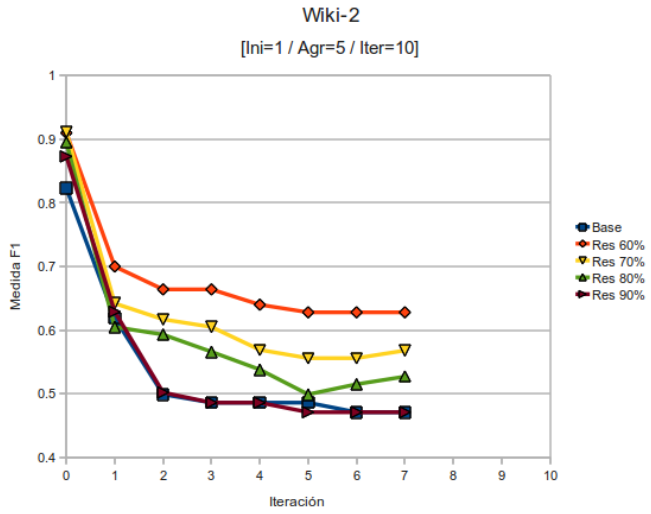

Figura 5.30: Continuación de los resultados de la figura anterior para resúmenes entre 60 % y 90 %.

Sin embargo, cabe señalarse una vez más el impacto de los resúmenes en el sistema ya que en los casos en que se emplean resúmenes de 10 %, 20 % y 30 % la evaluación se mantiene estable mientras los otros casos, incluido el baseline muestra una caída.

5.4. Discusión y Conclusiones

Como puede observarse en los resultados obtenidos, el algoritmo propuesto permite utilizar información de los documentos no etiquetados para mejorar el desempeño del clasificador. Aunque este comportamiento es constante y se manifiesta en mayor o menor medida en todos los experimentos realizados, está sujeto a una serie de condiciones que restringen sus efectos. Entre ellas pueden mencionarse, además de la naturaleza de los documentos que ya ha sido discutida en el capítulo anterior, la cantidad de documentos no etiquetados de los que se dispone, el número y la calidad de los documentos inicialmente etiquetados y la homogeneidad del clúster que forman, la proporción de información contenida en los documentos de entrenamiento que es relevante para la clase y la calidad del clasificador inicial.

Para discutir estas condicionantes del sistema conviene hacer una revisión del algoritmo propuesto. La entrada del sistema son los conjuntos etiquetado y no etiquetado. La salida es un clasificador que es evaluado con el conjunto de prueba más

un conjunto de documentos confiables seleccionados a partir de la clasificación sobre el conjunto no etiquetado realizada por el clasificador entrenado con el conjunto etiquetado y que cumplen las condiciones establecidas por el criterio de selección.

Primero. La calidad del clasificador inicial y en consecuencia de la clasificación realizada sobre el conjunto no etiquetado está directamente relacionada con la calidad de los documentos del conjunto etiquetado inicial. Por esta razón, dichos documentos son elegidos (o etiquetados) manualmente. Si dichos documentos representan adecuadamente a la categoría a la que pertenecen, entrenarán un clasificador con un buen desempeño (*bueno* en la medida que el número de documentos etiquetados lo permita) y la clasificación de los documentos no etiquetados también será buena. Esto se refleja en la evaluación del clasificador inicial (iteración 0).

Si se ha hecho una clasificación suficientemente buena del conjunto no etiquetado, el criterio de selección de documentos confiables elegirá a los k documentos que guardan una menor distancia hacia el clúster formado por los documento etiquetados, pertenecientes a su categoría, con los que el clasificador fue entrenado. Esta condición garantiza que los clústers de cada categoría se mantengan compactos en el espacio sobre el cuál son representados. Así, la salida del sistema es un conjunto de documentos que son muy cercanos a aquellos con los cuales el clasificador inicial ha sido entrenado. Estos documentos se retiran del conjunto no etiquetado y se agregan al conjunto etiquetado con el que se entrenará al clasificador de la siguiente iteración.

Segundo. Como se vio en el capítulo anterior, reemplazar los documentos del conjunto de entrenamiento (conjunto etiquetado) por sus resúmenes, puede verse como una selección de atributos. Esto se debe a que los resúmenes mantienen la información relevante de los documentos mientras desechan información que puede considerarse como ruido. Naturalmente, esta separación o filtrado de información está en función del tamaño de resúmenes empleados y de la proporción de información útil-no útil que contienen los documentos. La selección de atributos mediante resúmenes automáticos permite conservar los atributos que presentan una mayor *densidad de frecuencia* en el documento original, de manera que la distribución de los pesos de los atributos de los resúmenes es cercana a la del documento del cual fue extraído. Esta condición favorece el uso de resúmenes en el conjunto de entrenamiento y documentos completos en el conjunto de prueba ya que, a pesar de pertenecer a la misma categoría, la distribución de los atributos de los resúmenes del conjunto de prueba puede ser distinta a aquella que corresponde a los resúmenes del conjunto de entrenamiento de modo que al

mantener los documentos completos en el conjunto de prueba se evita este problema. Con este argumento se justifica la observación de que en todos los casos el clasificador inicial muestra una mejor evaluación cuando se usan resúmenes automáticos independientemente del tamaño que cuando se emplean documentos completos.

Otra característica derivada de esta condición es que, al emplear los atributos de los resúmenes para representar a los documentos que el clasificador procesará, el criterio de selección de documentos confiables emplea los mismos atributos, lo que ocasiona que sea relativamente más simple seleccionar documentos confiables cuando se emplean resúmenes que cuando no. Esto puede justificar el hecho de que en algunos experimentos el caso base mejora muy lentamente debido a que se agregan muy pocos documentos en cada iteración.

Tercero. Si la condición de que el clúster formado por los documentos etiquetados usados en el entrenamiento se mantenga poco disperso se cumple, el resultado de la evaluación en cada iteración tiende a mejorar. Esta condición puede modificarse principalmente debido a la relajación del criterio de selección de documentos confiables ya que esto significa dejar pasar a documentos que, aunque pertenecientes a la categoría, mantienen una distancia mayor respecto al clúster provocando que se disperse. Cuando esto ocurre, el desempeño del clasificador puede mejorar si la categoría no es muy homogénea como en el caso del corpus Meter; sin embargo, si la categoría es muy homogénea y se permite agregar documentos no muy cercanos al clúster, el clasificador puede desviarse.

El número de documentos etiquetados iniciales también afecta a esta condición pues, como se ve en el ejemplo de R8, el sistema tiene un muy buen desempeño cuando se utiliza únicamente 1 documento por categoría, lo que significa que a lo largo de la iteraciones, documentos muy cercanos al mismo son hallados e incorporados al conjunto etiquetado como puede verse cuando se relaja el criterio de selección permitiendo agregar hasta 5 y hasta 10 documentos por clase en cada iteración y manteniendo el comportamiento del sistema. En tanto que, cuando se emplean 5 documentos etiquetados por categoría el desempeño del sistema no mejora y tiende a decaer.

Finalmente, a partir de los experimentos realizados y el estudio del comportamiento del sistema, se pueden enunciar algunas conclusiones:

- **Incorporar resúmenes automáticos en un sistema de clasificación semisupervisada mejora el desempeño del sistema.** En todas las situaciones

reportadas, el uso de resúmenes automáticos permite mejorar el desempeño del sistema semisupervisado en mayor o menor medida dependiendo de las condiciones del sistema y con respecto al caso base en el que no se emplean resúmenes automáticos.

- **Los resúmenes de menor tamaño permiten hacer una mejor clasificación.** Como se mostró en el capítulo anterior, el efecto de los resúmenes es mayor cuando su extensión está por debajo de cierto umbral. En el caso del sistema semisupervisado, cuanto menores son los resúmenes mayor es la confianza que se tiene en la clasificación y los documentos seleccionados como confiables que se integrarán en la siguiente iteración guardarán una relación más cercana con el conjunto etiquetado inicial por lo que el nuevo clasificador mantendrá un buen desempeño. La misma conclusión se puede obtener al observar que usando resúmenes con una extensión cercana al documento completo (80 % - 90 %) el desempeño del sistema tiende a ser como en el caso base.

- **El conjunto etiquetado inicial es determinante en el desempeño del sistema.** Debido a que el criterio de selección de nuevos documentos etiquetados está en función de la calidad de los clústers formados por los documentos de cada categoría en el conjunto etiquetado, un clúster bien formado favorece el buen funcionamiento del sistema, esta condición es trivial cuando se emplea únicamente un documento por categoría como conjunto etiquetado inicial pero cobra importancia cuando el número de documentos por clase es mayor. Particularmente, en el caso del corpus Meter donde cada categoría estaba formada por varias subcategorías, cuantas más de estas fueron abarcadas por los documentos etiquetados, mejor fue el desempeño del sistema.

- **El criterio de selección de documentos confiables puede relajarse para integrar una cantidad mayor a 1 documento en cada iteración pero existe una cota superior para evitar que el clasificador se desvíe.** Los experimentos demostraron que cuando se integra un documento de cada categoría al conjunto etiquetado se tiene un mayor control sobre éste. Sin embargo, los experimentos en que se permite integrar hasta 5 documentos muestran un comportamiento semejante en un número menor de iteraciones además de que los clasificadores correspondientes llegan a obtener evaluaciones superiores. Los experimentos que permiten agregar hasta 10 documentos al conjunto etiquetado

en cada iteración muestran que existe una cota superior en el número de documentos que pueden ser agregados en cada iteración, ya que además de decaer rápidamente, la evaluación de sus clasificadores es inferior.

- **Es necesario contar con una cantidad suficientemente grande de documentos no etiquetados.** Como se mostró en los experimentos, el tamaño del conjunto de documentos no etiquetados también determina qué tanto puede mejorarse el clasificador inicial ya que el sistema debe poder elegir los más cercanos a los documentos etiquetados en cada iteración con la finalidad de mantener los clúster de cada categoría lo más compactos posible para garantizar el buen desempeño del clasificador. En los casos en que no se cuenta con suficientes documentos no etiquetados, aquellos seleccionados por el criterio de selección suelen guardar una distancia considerable con el clúster de la categoría a la que pertenecen haciendo que este se disperse más en cada iteración.

Capítulo 6
Conclusiones

En este trabajo de tesis se ha desarrollado, implementado y evaluado un método semisupervisado para clasificación de documentos que emplea resúmenes automáticos. Mediante experimentos realizados en diferentes conjuntos de datos y modificando los parámetros del sistema se ha demostrado que el método propuesto permite entrenar un clasificador con mejor desempeño en comparación con no utilizar resúmenes automáticos, en pocas iteraciones y utilizando información de un número reducido de documentos.

En los experimentos realizados se mostró que el comportamiento del sistema depende de parámetros como el número y calidad de los documentos etiquetados que sirven de entrada al sistema, el grado de relajación del criterio de selección de documentos confiables y la cantidad de documentos integrados al conjunto etiquetado en cada iteración, la cantidad de documentos no etiquetados de los que se dispone para extraer información que mejore al clasificador inicial, el número de iteraciones realizadas y la extensión de los resúmenes empleados.

El número y calidad de los documentos etiquetados de los que se dispone es determinante para un sistema basado en *self-training* debido a que toda la información que se proporcione al clasificador será una función directa de la calidad del grupo que forman los documento de cada categoría. Así, si dichos documentos forman un grupo compacto en el espacio de atributos, el clasificador reconocerá adecuadamente a los documentos no etiquetados que pertenecen a la misma clase, el criterio de selección elegirá documentos altamente confiable y la categoría tendrá un crecimiento uniforme, situación que no posible si el grupo de documentos inicial está muy disperso.

El criterio de selección de documentos confiables actúa principalmente como un filtro que permite en mayor o menor medida que documentos del conjunto no etiquetado pasen a formar parte del conjunto con el que se entrena al clasificador de cada

iteración. El comportamiento del sistema es, hasta cierto grado independiente de la calidad del criterio de selección y recae más en la calidad de la clasificación realizada sobre el conjunto no etiquetado, sin embargo, un criterio confiable permite filtrar, entre los documentos etiquetados como pertenecientes a dicha categoría a aquellos que son más confiables descartando documentos dudosos favoreciendo el crecimiento de un conjunto etiquetado adecuado.

Por otro lado, la cantidad de documentos que son seleccionados en cada iteración también es relevante para el desempeño del clasificador ya que determina en cierta medida el grado de desviación de los documentos etiquetados de cada categoría con respecto a los documentos iniciales. Si se desea que la categorías se mantengan tan compactas como en el conjunto inicial, pocos deben ser los documentos seleccionados en cada iteración, esto provoca, en la mayoría de los casos que el clasificador evolucione lentamente aunque de una forma más consistente. Si por el contrario, desea introducirse un poco de ruido para ampliar las categorías del conjunto etiquetado, un criterio más permisivo puede ser empleado, aunque esto abre la posibilidad de que el clasificador se desvíe más rápidamente. Evidentemente, un compromiso entre número de iteraciones y la mejoría del clasificador existe, corresponde a las condiciones de cada situación determinar en qué medida es preferible favorecer a una u otra.

Finalmente, la cantidad de documentos no etiquetado de los que se dispone también es una condición qué determina cuánto se puede mejorar al clasificador inicial debido a que es precisamente de los documentos no etiquetados de dónde se obtiene información para enriquecer al clasificador. De esta manera, resulta poco útil tener un buen sistema si no se tienen suficientes documentos no etiquetados de los cuáles extraer nueva información.

Aunque la discusión y conclusiones particulares de cada etapa del desarrollo del sistema han sido comentadas en la secciones finales de los últimos dos capítulos, algunas conclusiones generales derivadas de este trabajo pueden enunciarse como sigue.

- **Se ha realizado una exploración sobre el impacto de los resúmenes automáticos en clasificación supervisada** Este análisis tuvo como objetivo ampliar la información disponible en la literatura acerca del comportamiento de los resúmenes automáticos en sistemas de clasificación supervisada ya que los enfoques reportados se limitan en la forma de incorporar los resúmenes

siguiendo un solo esquema. En este estudio, se exploraron diversos esquemas para incorporar resúmenes automáticos en un sistema de clasificación lo que permitió determinar que el uso de resúmenes automáticos en el conjunto de entrenamiento pero no en los documentos que han de ser clasificados aumenta el efecto de los resúmenes.

- **Se demostró que los resúmenes automáticos pueden emplearse como una estrategia de selección de atributos** Debido a que usar resúmenes únicamente en el conjunto de prueba puede verse como una estrategia de selección de atributos, una exploración de sus alcances fue realizada. Se mostró que es particularmente útil en situaciones cuando el conjunto de entrenamiento cuenta con pocos documentos ya que a diferencia de otros métodos de selección de atributos, hace una selección utilizando únicamente información local de cada documento y no utiliza información estadística que dependa del tamaño del conjunto de datos. Esta condición permite incorporar Resúmenes Automáticos en el sistema semisupervisado.

- **Se propuso e implementó una arquitectura para incorporar resúmenes automáticos en un sistema de clasificación semisupervisado** El análisis sobre el impacto de los resúmenes automáticos en clasificación supervisada determinó las condiciones en las que los resúmenes pueden aprovecharse en un sistema semisupervisado, a partir de ellas, se propuso la arquitectura para un sistema semisupervisado basado en *self-training* que incorpora resúmenes automáticos.

- **Se propuso un criterio de selección de documentos confiables** Los experimentos realizados demostraron que un clasificador que utiliza información sobre la ubicación espacial de los documentos y la separación entre diferentes categorías como *support vector machines* realza el efecto de los resúmenes. Debido a que el sistema semisupervisado requiere de un criterio para seleccionar los documentos confiables del conjunto no etiquetado que se integrarán al conjunto etiquetado, fue necesario desarrollar un criterio adecuado a las condiciones del clasificador. El criterio propuesto está basado en la medida de distancia del coeficiente *Silhouette* y un clasificador kNN y permite elegir documentos cercanos a los de entrenamiento para mantener homogéneo el conjunto de documentos de cada categoría.

- **Se desarrolló y evaluó un sistema que implementa el método de clasificación semisupervisado basado en *self-training* que utiliza resúmenes automáticos** Una vez que se conoció el efecto de los resúmenes automáticos en un sistema de clasificación supervisada, fue posible desarrollar un algoritmo semisupervisado para clasificación de documentos que incorpora resúmenes automáticos para mejorar su desempeño. Se propuso una arquitectura y se mostró que la implementación del algoritmo es efectiva. De igual modo se comprobó que los resúmenes automáticos permiten producir mejores clasificadores en pocas iteraciones y utilizando pocos documentos de entrenamiento.

6.1. Trabajo Futuro

En este trabajo se han explorado el impacto de los resúmenes automáticos en clasificación supervisada y la capacidad de los resúmenes automáticos para mejorar el desempeño de un sistema semisupervisado. Se ha desarrollado un algoritmo que incorpora resúmenes en clasificación supervisada y se ha demostrado que su impacto puede ser positivo. Algunas de las perspectivas para trabajo futuro incluyen:

- **Analizar el comportamiento del sistema en otros conjuntos de datos** Entre los conjuntos utilizados en los experimentos, R8 reunía las condiciones adecuadas para aplicar el sistema mientras los demás representaban características especiales como pocos documentos no etiquetados, categorías jerárquicas o elevada similitud entre sus documentos. Aunque el sistema mostró un buen comportamiento aún bajo dichas condiciones, su aplicación en un dominio más amplio, con suficientes documentos no etiquetados y categorías con traslape puede proporcionar información relevante.

- **El uso de criterios de selección distintos** Ya que el criterio propuesto fue desarrollado para ser usado junto al clasificador SVM, combinaciones distintas de clasificador-criterio de selección pueden ampliar los resultados obtenidos.

- **Usar documentos virtuales como conjunto etiquetado** Con *documentos virtuales* se hace referencia a documentos no extraídos del dominio sobre el que se aplicará el sistema sino construidos a partir de información disponible a priori. Dichos documentos podrían estar formados únicamente por grupos de

palabras distintivas de la categoría a la que el documento pertenecerá. Este enfoque debería funcionar pues la representación de los documentos como bolsas de palabras no distingue entre un documento real y uno artificial de esta naturaleza. Adicionalmente, este enfoque tendría la ventaja de no requerir documentos etiquetados sino únicamente una lista de palabras distintivas. Probablemente bajo este enfoque sería innecesario sumarizar el conjunto inicial pues carecería de sentido, no así con los documentos seleccionados en cada iteración.

- **Explorar la posibilidad de incorporar resúmenes automáticos en otros esquemas semisupervisados** Se ha mostrado que los resúmenes automáticos pueden mejorar el desempeño de un sistema de clasificación basado en Self-training. Una exploración de las posibilidad de incorporar resúmenes en otros esquemas como *co-training* o métodos basados en *expectation maximization* podría producir resultados importantes.

Bibliografía

[1] C. Apté, F. Damerau, and S. M. Weis. Toward Language Independent Automated Learning of Text Categorization Models. *SIGIR '94*, 1994.

[2] S. Aranganayagi and K. Thangavel. Clustering Categorical Data Using Silhouette Coefficiente as a Relocating Measure. *International Conference on Computational Intelligence and Multimedia Applications 2007*, 2007.

[3] P. Baxendale. Machine Index for Technical Literature, An Experiment. *IBM Journal of Research Development. 2(4) pp. 354-361. US.*, 1958.

[4] A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. *COLT: Proceedings of the Workshop on Computational Learning Theory.*, 1998.

[5] S. Brin and L. Page. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems, 30 (1 - 7).*, 1998.

[6] A. M. d. J. Cardoso-Cachopo. *Improving Methods for Single-label Text Categorization.* PhD thesis, Universidade Técnica de Lisboa, Instituto Superior Técnico, July 2007.

[7] J. M. Conroy and D. P. O'Leary. Text Summarization Via Hidden Markov Models. *In Preceedings of SIGIR'01 pp. 406-407. NY, US.*, 2001.

[8] D. Das and A. F. Martins. A Survey on Text Summarization. *Language Technologies Institute. Carnagie Mellon University.*, 2007.

[9] Y. Gong and X. Liu. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. *In SIGIR '01: Proceedings of the 24th Annual Inter-*

national ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 19 - 25., 2001.

[10] R. Guzmán-Cabrera, M. Montes-y Gómez, L. Villaseñor Pineda, and P. Rosso. Using the Web as Corpus for Self-Training Text Categorization. *Facultad de Ingeniería Mecánica, Eléctrica y Electrónica. Universidad de Guanajuato. INAOEP.*, 2008.

[11] G. Haffari. A Survey on Inductive Semi-supervised Learning. 6 Mar. 2006.

[12] A. Hotho, A. Maedche, and S. Staab. Text Clustering Based on Good Agregations. *Institute AIFB, University of Karlsruhe. Karlsruhe Germany.*, 2003.

[13] E. Hovy. *Text Summarization. In The Oxford Handbook of Computational Linguistics*. Oxford University Press, 2004.

[14] X. Jiang, X. Fan, and Z. Wang. Improving the Performance of Text Categorization Using Automatic Summarization. *Internartional Conference on Computer Modeling and Simulation. IEEE.*, 2009.

[15] C. K. Jiang Xiao-yu, Fan Xiao-zhong. Chinese Text Classification Based on Summarization Technique. *Third International Conference on Semantics, Knowledge and Grid (SKG 2007) pp.362-365,*, 2007.

[16] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of ECML-98, 10th European Conference on Machine Learning. LNCS 1398, pp. 137 - 142*, 1998.

[17] G. K. Kanji. *100 Statistical Tests*. SAGE Publications, 2006.

[18] S. Ker and J.-N. Chen. A Text Categorization Based on Summarization Technique. *Proceedings of ACL-2000 Workshop on Recent Avances in Natural Language Processing and Information Retrieval. pp 79-83.*, 2000.

[19] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[20] Y. Ko, J. Park, and J. Seo. Improving Text Categorization Using the Importance of Sentences. *Information Proceesing and Management 40. pp. 65 - 79*, 2004.

[21] S. J. Ko Youngjoong, Park Jinwoo. Automatic Text Categorization Using the Importance of Sentences. *Proceedings of the 19th International Conference on Computational Linguistics, Volume 1. pp 1 - 7*, 2002.

[22] K. J. Kolcz A., Prabakarmurthi V. Summarization as a Feature Selection in Text Categorization. *In H. Paques, L. Liu & D. Grossman. Proceedings of CIMKM-01 pp. 365-370. NY. US.*, 2001.

[23] Z. Kozareva, B. Bonev, and A. Montoyo. Self-training and Co-training Applied to Spanish Named Entity Recognition. *MICAI 2005, LNAI 3789, pp. 770 - 779.*, 2005.

[24] J. Kupiec, J. Pedersen, and F. Chen. A Trainable Document Summarizer - Step 1: Sentence Compression. *Proceedings of SIGIR'95 pp. 68-73. NY, US.*, 1995.

[25] C. Lanquillon. Learning from Labeled and Unlabeled Documents: A Comparative Study on Semi-supervised Text Classification. *PKDD 2000, LNAI 1910, pp. 490 - 497*, 2000.

[26] H. P. Luhn. The Automatic Creation of Automatic Abstracts. *IBM Journal of Research Develpment. 2(2) pp. 159-165. US.*, 1958.

[27] A. McCallum and K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. *School of Computer Science, Carnagie Mellon University, Pittsburgh PA. US.*, 1998.

[28] R. Mihalcea. Graph Based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. *Proceedings of Fourth Colloquium on Implementation of Contraint and Logic Programming Systems. Saint-Malo. France.*, 2004.

[29] R. Mihalcea and S. Hassan. Using the Essence of Text to Improve Document Classification. *Proceedings of RANLP. Borevetz Bulgaria.*, 2005.

[30] K. Nigam, A. McCallum, and S. Thrun. Learning To Classify Text From Labeled and Unlabeled Documents. *School of Computer Science, Carnagie Mellon University. Pittsburgh PA. US.*, 1998.

[31] D. R. Radev, E. Hovy, and K. McKeown. Introduction to the Special Issue on Summarization. *Computer Linguistics 28. pp. 399-408.*, 2002.

[32] M. Saravanana, R. P. C. Reghu, and S. Raman. Summarization and Categorization of Text Data in High Level Data Cleaning for Information Retrieval. *Applied Artificial Intelligence, Volume 19, No. 18, 461 - 47*, 2005.

[33] F. Sebastiani. Machine Learning in Automated Text Categorization. *Istitutoo di Elaborazione dellÍnformazione, Consiglio Nazionale delle Richerche, Italy*, 2002.

[34] F. Sebastiani. Text Categorization. *Dipartimento di Matematica Pura e Applicata, Universita di Padova. Padova Italia.*, 2005.

[35] M. Seeger. Learning with Labeled and Unlabeled Data. *Institue for Adaptive and Neural Computation, University of Edinburgh*, 13 Feb. 2001.

[36] D. Shen, Q. Yang, and Z. Chen. Noise Reduction Through Summarization for Web-page Classification. *department of CS and Technology, Hong Kong University of Science and Technology. HK.*, 2007.

[37] T. Solorio. Using Unlabeled Data to Improve Classifier Accuracy. Master's thesis, Instituto Nacional de Astrofísica Óptica y Electrónica, Aug. 2002.

[38] K. Svore, L. Vanderwende, and C. Burges. Enhancing Single-Documet Summarization by Combining RankNet and Third-Part Sources. *In Preceedings of EMNLP-CoNLL. pp. 448-457. US.*, 2007.

[39] V. Vapnic. *The Nature of Statistical Learning Theory.* Springer, New York., 1995.

[40] E. Villatoro-Tello. Generación Automática de Resúmenes de Multiples Documentos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, Feb. 2007.

[41] J.-N. Vittaut, M.-R. Amini, and P. Gallinari. Learning Classification with Both Labeled and Unlabeled Data. *ECML, LNAI 2430, pp. 468-479.*, 2002.

[42] B. Wang, B. Spencer, and C. X. Ling. Semi-supervised Self-training for Sentence Subjectivity Classification. *Canadian AI 2008, LNAI 5032, pp. 344 - 355.*, 2008.

[43] Z. Yabin, T. Shaohua, and L. Zhiyuan. Text Classification Based on Transfer Learning and Self-Training. *Fourth Internation Conference on Natural Computation, IEEE Computer Society*, 2008.

[44] Y. Yang and J. Pedersen. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the ICML-97, 14th International Conference on Machine Learning. pp. 412-420. Nashville US.*, 1997.

[45] D. Zhi-Hong, T. Shi-Wei, and Y. Dong-Qing. A Comparative Study on Feature Weight in Text Categorization. *APWeb 2004, LNCS 3007, pp. 588 - 597*, 2004.

[46] X. Zhu. Semisupervised Learning Literature Survey. *Computer Science TR, University of Winsconsin, Madison US.*, 2008.

Apéndices

Apéndice

Resultados de los Experimentos: R8

Las tablas mostradas en este y los siguientes apéndices corresponden a los datos que muestran las gráficas del Capítulo 5: Clasificación Semisupervisada con Resúmenes Automáticos. En ellas se muestran los resultados de la evaluación del Sistema Semisupervisado para Clasificación de Documentos que incorpora Resúmenes Automáticos. La medida de evaluación empleada es la Medida F1 (ver sección 2.1.5) y debido a que los corpora utilizados son multiclase los resultados de cada categoría se han promediado usando Micropromedios para obtener un solo índice del desempeño del clasificador.

Cada una de las siguientes secciones corresponde a uno de los conjuntos de datos empleados, en el pie de cada tabla se indica el número de la gráfica correspondiente.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.349	0.519	0.477	0.43	0.437	0.432
1	0.369	0.662	0.507	0.505	0.449	0.461
2	0.379	0.674	0.493	0.565	0.54	0.516
3	0.375	0.665	0.543	0.658	0.571	0.551
4	0.376	0.613	0.541	0.662	0.577	0.531
5	0.377	0.627	0.562	0.708	0.61	0.558
6	0.38	0.618	0.579	0.734	0.617	0.602
7	0.379	0.623	0.637	0.731	0.629	0.622
8	0.379	0.62	0.653	0.725	0.639	0.649
9	0.379	0.586	0.667	0.728	0.645	0.677
10	0.379	0.583	0.67	0.674	0.65	0.72
11	0.385	0.58	0.679	0.673	0.644	0.728
12	0.384	0.592	0.689	0.657	0.646	0.733
13	0.385	0.554	0.685	0.658	0.649	0.75
14	0.386	0.548	0.686	0.654	0.653	0.755
15	0.381	0.549	0.687	0.652	0.686	0.76
16	0.381	0.542	0.696	0.65	0.702	0.752
17	0.381	0.498	0.693	0.628	0.71	0.737
18	0.381	0.493	0.688	0.625	0.72	0.684
19	0.381	0.494	0.707	0.627	0.724	0.702
20	0.381	0.499	0.708	0.623	0.729	0.721

Tabla A.1: Colección: R8. Número de Documentos Iniciales:1. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.3

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.349	0.38	0.374	0.353	0.353
1	0.369	0.411	0.383	0.379	0.374
2	0.379	0.438	0.388	0.389	0.382
3	0.375	0.468	0.399	0.39	0.377
4	0.376	0.487	0.404	0.387	0.383
5	0.377	0.493	0.408	0.391	0.383
6	0.38	0.539	0.408	0.387	0.383
7	0.379	0.544	0.421	0.387	0.383
8	0.379	0.551	0.429	0.389	0.381
9	0.379	0.566	0.433	0.391	0.381
10	0.379	0.591	0.432	0.385	0.382
11	0.385	0.605	0.433	0.387	0.383
12	0.384	0.613	0.433	0.386	0.386
13	0.385	0.62	0.432	0.387	0.386
14	0.386	0.637	0.432	0.39	0.39
15	0.381	0.64	0.434	0.391	0.389
16	0.381	0.644	0.435	0.391	0.39
17	0.381	0.664	0.436	0.387	0.389
18	0.381	0.658	0.436	0.387	0.388
19	0.381	0.658	0.439	0.387	0.385
20	0.381	0.644	0.438	0.387	0.385

Tabla A.2: Colección: R8. Número de Documentos Iniciales:1. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.4

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.349	0.519	0.477	0.43	0.437	0.432
1	0.39	0.727	0.676	0.627	0.621	0.578
2	0.39	0.748	0.729	0.735	0.687	0.532
3	0.39	0.743	0.675	0.775	0.755	0.546
4	0.39	0.741	0.685	0.762	0.738	0.609
5	0.389	0.736	0.67	0.763	0.713	0.635
6	0.385	0.725	0.658	0.764	0.722	0.63
7	0.385	0.732	0.658	0.764	0.722	0.642
8	0.385	0.721	0.664	0.763	0.716	0.637
9	0.384	0.719	0.665	0.764	0.709	0.645
10	0.38	0.726	0.659	0.763	0.706	0.651
11	0.381	0.715	0.664	0.764	0.71	0.65
12	0.38	0.703	0.661	0.763	0.712	0.654
13	0.379	0.697	0.659	0.762	0.708	0.655
14	0.379	0.686	0.665	0.763	0.687	0.652
15	0.379	0.683	0.661	0.763	0.68	0.652
16	0.379	0.685	0.667	0.763	0.682	0.649
17	0.379	0.692	0.663	0.763	0.683	0.65
18	0.379	0.689	0.657	0.763	0.689	0.65
19	0.379	0.686	0.658	0.763	0.69	0.652
20	0.379	0.695	0.66	0.763	0.689	0.66

Tabla A.3: Colección: R8. Número de Documentos Iniciales:1. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.5

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.349	0.38	0.374	0.353	0.353
1	0.39	0.476	0.422	0.408	0.388
2	0.39	0.455	0.452	0.406	0.388
3	0.39	0.452	0.443	0.404	0.387
4	0.39	0.452	0.443	0.404	0.388
5	0.389	0.452	0.442	0.402	0.385
6	0.385	0.452	0.442	0.402	0.382
7	0.385	0.453	0.442	0.402	0.382
8	0.385	0.455	0.441	0.402	0.382
9	0.384	0.451	0.441	0.394	0.381
10	0.38	0.454	0.441	0.393	0.378
11	0.381	0.447	0.441	0.393	0.377
12	0.38	0.437	0.441	0.389	0.377
13	0.379	0.433	0.441	0.388	0.377
14	0.379	0.433	0.441	0.388	0.377
15	0.379	0.433	0.441	0.387	0.376
16	0.379	0.433	0.441	0.384	0.376
17	0.379	0.433	0.441	0.384	0.373
18	0.379	0.431	0.441	0.384	0.372
19	0.379	0.431	0.441	0.384	0.372
20	0.379	0.431	0.438	0.384	0.372

Tabla A.4: Colección: R8. Número de Documentos Iniciales:1. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.6

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.349	0.519	0.477	0.43	0.437	0.432
1	0.388	0.732	0.658	0.67	0.645	0.627
2	0.385	0.739	0.667	0.639	0.653	0.702
3	0.383	0.746	0.715	0.604	0.644	0.621
4	0.382	0.754	0.68	0.584	0.647	0.625
5	0.367	0.728	0.672	0.584	0.611	0.675
6	0.368	0.707	0.581	0.58	0.612	0.524
7	0.368	0.508	0.577	0.574	0.612	0.524
8	0.368	0.427	0.536	0.574	0.612	0.523
9	0.367	0.429	0.405	0.573	0.6	0.523
10	0.367	0.435	0.474	0.58	0.601	0.523
11	0.367	0.443	0.471	0.585	0.6	0.522
12	0.367	0.443	0.458	0.586	0.601	0.522
13	0.367	0.455	0.445	0.586	0.535	0.522
14	0.367	0.459	0.436	0.586	0.536	0.522
15	0.367	0.46	0.433	0.586	0.536	0.521
16	0.368	0.46	0.432	0.586	0.536	0.521
17	0.367	0.462	0.429	0.586	0.536	0.521
18	0.367	0.463	0.43	0.586	0.536	0.521
19	0.368	0.458	0.43	0.586	0.536	0.521
20	0.367	0.44	0.431	0.586	0.536	0.521

Tabla A.5: Colección: R8. Número de Documentos Iniciales:1. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 10. Figura Correspondiente: 5.7

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.349	0.38	0.374	0.353	0.353
1	0.388	0.509	0.456	0.423	0.389
2	0.385	0.461	0.453	0.401	0.385
3	0.383	0.462	0.423	0.403	0.384
4	0.382	0.462	0.423	0.4	0.382
5	0.367	0.463	0.423	0.4	0.371
6	0.368	0.462	0.424	0.399	0.372
7	0.368	0.45	0.422	0.399	0.372
8	0.368	0.45	0.423	0.399	0.372
9	0.367	0.45	0.422	0.399	0.371
10	0.367	0.45	0.422	0.398	0.371
11	0.367	0.45	0.422	0.399	0.371
12	0.367	0.448	0.423	0.399	0.371
13	0.367	0.449	0.422	0.399	0.371
14	0.367	0.449	0.423	0.399	0.371
15	0.367	0.449	0.423	0.399	0.371
16	0.368	0.449	0.422	0.399	0.371
17	0.367	0.449	0.423	0.399	0.371
18	0.367	0.448	0.421	0.399	0.371
19	0.368	0.448	0.421	0.399	0.371
20	0.367	0.449	0.422	0.399	0.371

Tabla A.6: Colección: R8. Número de Documentos Iniciales:1. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 10. Figura Correspondiente: 5.8

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.478	0.797	0.717	0.672	0.698	0.671
1	0.48	0.793	0.749	0.697	0.726	0.701
2	0.453	0.778	0.744	0.683	0.736	0.63
3	0.449	0.772	0.75	0.678	0.746	0.638
4	0.45	0.767	0.758	0.653	0.706	0.613
5	0.441	0.763	0.756	0.646	0.708	0.615
6	0.44	0.75	0.753	0.639	0.717	0.606
7	0.44	0.734	0.748	0.655	0.721	0.61
8	0.44	0.694	0.74	0.674	0.723	0.611
9	0.44	0.679	0.741	0.681	0.737	0.599
10	0.44	0.668	0.74	0.667	0.732	0.602
11	0.44	0.658	0.738	0.67	0.754	0.587
12	0.44	0.645	0.731	0.664	0.764	0.585
13	0.44	0.626	0.729	0.668	0.77	0.582
14	0.44	0.622	0.727	0.677	0.772	0.582
15	0.44	0.618	0.727	0.679	0.736	0.584
16	0.44	0.612	0.726	0.675	0.733	0.579
17	0.44	0.595	0.726	0.682	0.729	0.582
18	0.44	0.574	0.726	0.68	0.726	0.579
19	0.44	0.569	0.727	0.681	0.717	0.575
20	0.44	0.566	0.719	0.683	0.725	0.576

Tabla A.7: Colección: R8. Número de Documentos Iniciales:5. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.9.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.478	0.596	0.545	0.512	0.493
1	0.48	0.616	0.558	0.52	0.493
2	0.453	0.604	0.576	0.519	0.466
3	0.449	0.636	0.566	0.523	0.467
4	0.45	0.62	0.567	0.527	0.467
5	0.441	0.632	0.566	0.528	0.464
6	0.44	0.647	0.568	0.522	0.464
7	0.44	0.58	0.55	0.52	0.464
8	0.44	0.589	0.552	0.525	0.465
9	0.44	0.569	0.549	0.5	0.466
10	0.44	0.567	0.551	0.504	0.468
11	0.44	0.566	0.546	0.507	0.467
12	0.44	0.565	0.544	0.499	0.463
13	0.44	0.567	0.543	0.502	0.46
14	0.44	0.567	0.54	0.497	0.457
15	0.44	0.582	0.538	0.497	0.457
16	0.44	0.591	0.539	0.498	0.457
17	0.44	0.594	0.538	0.5	0.457
18	0.44	0.607	0.541	0.499	0.457
19	0.44	0.625	0.536	0.5	0.457
20	0.44	0.629	0.534	0.499	0.457

Tabla A.8: Colección: R8. Número de Documentos Iniciales:5. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.10.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.478	0.797	0.717	0.672	0.698	0.671
1	0.452	0.797	0.7	0.684	0.755	0.626
2	0.445	0.781	0.659	0.687	0.699	0.625
3	0.445	0.779	0.658	0.708	0.704	0.629
4	0.445	0.766	0.664	0.687	0.699	0.629
5	0.445	0.766	0.65	0.736	0.697	0.629
6	0.445	0.763	0.651	0.735	0.697	0.63
7	0.445	0.755	0.646	0.615	0.687	0.627
8	0.445	0.74	0.634	0.627	0.687	0.626
9	0.445	0.737	0.609	0.682	0.698	0.571
10	0.445	0.732	0.601	0.675	0.664	0.572
11	0.445	0.755	0.598	0.702	0.636	0.572
12	0.445	0.75	0.587	0.704	0.629	0.572
13	0.445	0.755	0.577	0.712	0.624	0.572
14	0.445	0.754	0.575	0.712	0.572	0.572
15	0.445	0.754	0.571	0.713	0.571	0.572
16	0.445	0.746	0.564	0.713	0.549	0.572
17	0.445	0.747	0.57	0.708	0.552	0.572
18	0.445	0.75	0.57	0.64	0.549	0.572
19	0.443	0.751	0.568	0.634	0.543	0.572
20	0.443	0.745	0.57	0.64	0.547	0.572

Tabla A.9: Colección: R8. Número de Documentos Iniciales:5. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.11.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.478	0.596	0.545	0.512	0.493
1	0.452	0.588	0.558	0.497	0.462
2	0.445	0.589	0.521	0.478	0.46
3	0.445	0.595	0.519	0.467	0.452
4	0.445	0.587	0.522	0.466	0.452
5	0.445	0.585	0.522	0.465	0.452
6	0.445	0.57	0.52	0.465	0.452
7	0.445	0.568	0.515	0.465	0.452
8	0.445	0.541	0.515	0.465	0.452
9	0.445	0.52	0.515	0.465	0.452
10	0.445	0.516	0.515	0.465	0.452
11	0.445	0.516	0.515	0.464	0.452
12	0.445	0.516	0.515	0.465	0.452
13	0.445	0.516	0.515	0.465	0.452
14	0.445	0.516	0.515	0.465	0.452
15	0.445	0.516	0.515	0.465	0.452
16	0.445	0.516	0.515	0.465	0.452
17	0.445	0.516	0.513	0.465	0.452
18	0.445	0.516	0.512	0.465	0.452
19	0.443	0.516	0.505	0.46	0.452
20	0.443	0.515	0.502	0.458	0.452

Tabla A.10: Colección: R8. Número de Documentos Iniciales:5. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.12.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.478	0.797	0.717	0.672	0.698	0.671
1	0.453	0.751	0.758	0.711	0.688	0.622
2	0.447	0.645	0.749	0.715	0.693	0.626
3	0.446	0.633	0.733	0.716	0.693	0.624
4	0.447	0.639	0.718	0.721	0.653	0.622
5	0.447	0.59	0.718	0.733	0.646	0.617
6	0.447	0.613	0.717	0.751	0.644	0.617
7	0.447	0.607	0.72	0.752	0.644	0.58
8	0.447	0.699	0.723	0.752	0.627	0.575
9	0.447	0.734	0.724	0.752	0.627	0.575
10	0.447	0.717	0.724	0.752	0.627	0.575
11	0.447	0.712	0.721	0.752	0.627	0.575
12	0.447	0.713	0.72	0.752	0.627	0.575
13	0.447	0.713	0.719	0.752	0.627	0.575
14	0.447	0.713	0.717	0.752	0.627	0.575
15	0.447	0.714	0.717	0.752	0.627	0.575
16	0.447	0.732	0.716	0.752	0.627	0.575
17	0.447	0.725	0.716	0.752	0.627	0.575
18	0.447	0.724	0.714	0.752	0.627	0.575
19	0.446	0.713	0.713	0.752	0.627	0.575
20	0.447	0.723	0.712	0.752	0.627	0.575

Tabla A.11: Colección: R8. Número de Documentos Iniciales:5. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 10. Figura Correspondiente: No.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.478	0.596	0.545	0.512	0.493
1	0.453	0.572	0.516	0.477	0.463
2	0.447	0.578	0.511	0.463	0.454
3	0.446	0.575	0.511	0.46	0.454
4	0.447	0.544	0.511	0.461	0.455
5	0.447	0.538	0.511	0.461	0.455
6	0.447	0.538	0.511	0.461	0.455
7	0.447	0.538	0.511	0.46	0.454
8	0.447	0.519	0.51	0.459	0.454
9	0.447	0.519	0.51	0.458	0.453
10	0.447	0.519	0.51	0.458	0.452
11	0.447	0.519	0.51	0.458	0.453
12	0.447	0.519	0.51	0.458	0.453
13	0.447	0.519	0.51	0.459	0.454
14	0.447	0.519	0.51	0.457	0.455
15	0.447	0.519	0.51	0.458	0.455
16	0.447	0.519	0.51	0.459	0.455
17	0.447	0.519	0.51	0.458	0.455
18	0.447	0.519	0.51	0.458	0.455
19	0.446	0.519	0.51	0.458	0.455
20	0.447	0.519	0.504	0.458	0.455

Tabla A.12: Colección: R8. Número de Documentos Iniciales:5. Número de Iteraciones:20. Número Máximo de Documentos Agregados por Iteración: 10. Figura Correspondiente: No.

Apéndice

Resultados de los Experimentos: Desastres Naturales

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.159	0.522	0.414	0.433	0.421	0.314
1	0.177	0.603	0.529	0.387	0.308	0.473
2	0.352	0.557	0.551	0.341	0.398	0.326
3	0.216	0.547	0.459	0.31	0.342	0.243
4	0.121	0.547	0.45	0.285	0.344	0.205
5	0.121	0.526	0.441	0.288	0.344	0.205
6	0.121	0.518	0.445	0.297	0.328	0.196
7	0.121	0.544	0.445	0.293	0.328	0.195
8	0.121	0.539	0.437	0.291	0.319	0.187
9	0.121	0.553	0.434	0.285	0.31	0.196
10	0.121	0.544	0.434	0.28	0.31	0.187

Tabla B.1: Colección: Desastres Naturales. Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.13.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.159	0.307	0.233	0.242	0.22
1	0.177	0.331	0.239	0.233	0.223
2	0.352	0.227	0.351	0.271	0.186
3	0.216	0.212	0.312	0.253	0.169
4	0.121	0.194	0.151	0.244	0.169
5	0.121	0.194	0.131	0.202	0.159
6	0.121	0.186	0.121	0.202	0.15
7	0.121	0.186	0.121	0.194	0.159
8	0.121	0.186	0.121	0.194	0.159
9	0.121	0.177	0.121	0.194	0.159
10	0.121	0.177	0.121	0.184	0.15

Tabla B.2: Colección: Desastres Naturales. Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.14.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.159	0.522	0.414	0.433	0.421	0.314
1	0.33	0.641	0.51	0.466	0.437	0.447
2	0.186	0.659	0.496	0.377	0.416	0.359
3	0.168	0.662	0.465	0.37	0.394	0.354
4	0.168	0.622	0.446	0.358	0.363	0.305
5	0.159	0.591	0.434	0.373	0.354	0.268
6	0.141	0.582	0.413	0.389	0.379	0.288
7	-	0.624	0.407	-	0.383	0.305
8	-	-	0.416	-	-	
9	-	-	0.406	-	-	
10	-	-	-	-	-	

Tabla B.3: Colección: Desastres Naturales. Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.15.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.159	0.307	0.233	0.242	0.22
1	0.33	0.276	0.227	0.237	0.194
2	0.186	0.214	0.202	0.202	0.177
3	0.168	0.214	0.177	0.186	0.177
4	0.168	0.196	0.177	0.169	0.169
5	0.159	0.188	0.16	0.16	0.16
6	0.141	0.179	0.151	0.151	0.15
7	-	0.241	-	-	0.15
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-

Tabla B.4: Colección: Desastres Naturales. Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.16.

Apéndice

Resultados de los Experimentos: Meter

[ht]

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.773	0.826	0.814	0.811	0.799	0.788
1	0.798	0.827	0.819	0.598	0.801	0.8
2	0.81	0.829	0.805	0.485	0.794	0.799
3	0.804	0.83	0.787	0.527	0.797	0.799
4	0.792	0.857	0.754	0.597	0.795	0.793
5	0.79	0.857	0.723	0.493	0.789	0.791
6	0.788	0.867	0.632	0.451	0.788	0.783
7	0.786	0.884	0.57	0.399	0.788	0.78
8	0.771	0.879	0.519	0.421	0.784	0.778
9	0.764	0.882	0.537	0.402	0.784	0.778
10	0.764	0.882	0.523	0.371	0.78	0.778

Tabla C.1: Colección: Meter. Número de Documentos Iniciales:3. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.17.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.773	0.789	0.789	0.788	0.775
1	0.798	0.797	0.809	0.814	0.798
2	0.81	0.826	0.789	0.799	0.816
3	0.804	0.843	0.784	0.785	0.858
4	0.792	0.844	0.778	0.777	0.859
5	0.79	0.839	0.776	0.777	0.855
6	0.788	0.813	0.778	0.775	0.853
7	0.786	0.791	0.774	0.773	0.849
8	0.771	0.784	0.774	0.77	0.772
9	0.764	0.782	0.773	0.766	0.765
10	0.764	0.777	0.775	0.766	0.765

Tabla C.2: Colección: Meter. Número de Documentos Iniciales:3. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.18.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.773	0.826	0.814	0.811	0.799	0.788
1	0.857	0.811	0.858	0.775	0.859	0.854
2	0.782	0.87	0.871	0.835	0.82	0.779
3	0.764	0.89	0.867	0.787	0.804	0.781
4	0.763	0.891	0.867	0.785	0.806	0.784
5	0.76	0.891	0.874	0.78	0.804	0.785
6	0.759	0.891	0.872	0.763	0.804	0.788
7	0.759	0.897	0.835	0.762	0.807	0.79
8	0.76	0.884	0.845	0.762	0.806	0.798
9	0.751	0.89	0.842	0.764	0.804	0.799
10	0.751	0.887	0.847	0.764	0.805	0.805

Tabla C.3: Colección: Meter. Número de Documentos Iniciales:3. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.19.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.773	0.789	0.789	0.788	0.775
1	0.857	0.867	0.859	0.847	0.854
2	0.782	0.862	0.83	0.773	0.765
3	0.764	0.838	0.843	0.767	0.765
4	0.763	0.823	0.826	0.764	0.764
5	0.76	0.826	0.819	0.764	0.759
6	0.759	0.822	0.82	0.762	0.757
7	0.759	0.819	0.822	0.762	0.757
8	0.76	0.821	0.826	0.76	0.758
9	0.751	0.823	0.826	0.76	0.756
10	0.751	0.823	0.83	0.76	0.751

Tabla C.4: Colección: Meter. Número de Documentos Iniciales:3. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.20.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.773	0.826	0.814	0.811	0.799	0.788
1	0.878	0.809	0.886	0.888	0.873	0.872
2	0.789	0.8	0.767	0.822	0.829	0.812
3	0.78	0.795	0.736	0.82	0.821	0.803
4	0.78	0.794	0.713	0.784	0.826	0.773
5	0.776	0.789	0.649	0.778	0.828	0.774
6	0.77	0.788	0.617	0.778	0.832	0.774
7	0.767	0.786	0.55	0.77	0.812	0.772
8	0.763	0.786	0.55	0.77	0.818	0.772
9	0.762	0.784	0.574	0.771	0.823	0.77
10	0.762	0.784	0.584	0.769	0.826	0.77

Tabla C.5: Colección: Meter. Número de Documentos Iniciales:3. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 10. Figura Correspondiente: 5.21.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.773	0.789	0.789	0.788	0.775
1	0.878	0.896	0.886	0.896	0.872
2	0.789	0.831	0.8	0.793	0.784
3	0.78	0.818	0.794	0.784	0.778
4	0.78	0.812	0.795	0.783	0.778
5	0.776	0.81	0.794	0.779	0.774
6	0.77	0.805	0.781	0.775	0.77
7	0.767	0.806	0.784	0.773	0.767
8	0.763	0.791	0.77	0.77	0.763
9	0.762	0.791	0.772	0.768	0.761
10	0.762	0.792	0.772	0.768	0.76

Tabla C.6: Colección: Meter. Número de Documentos Iniciales:3. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 10. Figura Correspondiente: 5.22.

Apéndice

Resultados de los Experimentos: Wiki

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.325	1	0.948	0.948	0.668	0.584
1	0.193	0.842	0.973	0.949	0.775	0.794
2	0.131	0.896	0.973	0.949	0.874	0.842
3	0.263	0.974	0.973	0.923	0.843	0.771
4	0.215	1	0.973	0.947	0.842	0.771
5	0.28	1	0.973	0.947	0.843	0.788
6	0.28	0.973	0.973	0.923	0.896	0.788
7	0.365	0.947	0.973	0.923	0.869	0.711
8	0.365	0.947	0.973	0.898	0.869	0.685
9	0.365	0.974	0.973	0.898	0.871	0.685
10	0.365	0.974	0.973	0.898	0.875	0.685

Tabla D.1: Colección: Wiki (Partición 1). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.23.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.325	0.473	0.416	0.421	0.305
1	0.193	0.571	0.605	0.395	0.282
2	0.131	0.713	0.543	0.336	0.22
3	0.263	0.653	0.541	0.402	0.274
4	0.215	0.628	0.509	0.526	0.3
5	0.28	0.607	0.532	0.592	0.443
6	0.28	0.64	0.547	0.573	0.433
7	0.365	0.665	0.488	0.573	0.481
8	0.365	0.666	0.511	0.604	0.519
9	0.365	0.666	0.518	0.604	0.481
10	0.365	0.662	0.551	0.604	0.529

Tabla D.2: Colección: Wiki (Partición 1). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.24.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.325	1	0.948	0.948	0.668	0.584
1	0.552	1	1	0.974	0.919	0.92
2	0.602	0.974	0.947	0.92	0.947	0.947
3	0.555	0.974	0.973	0.947	0.921	0.947
4	0.531	-	0.973	0.947	0.871	0.894
5	0.531	-	-	0.92	0.871	
6	0.501	-	-	-	-	
7	-	-	-	-	-	
8	-	-	-	-	-	
9	-	-	-	-	-	
10	-	-	-	-	-	

Tabla D.3: Colección: Wiki (Partición 1). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.25.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.325	0.473	0.416	0.421	0.305
1	0.552	0.891	0.919	0.683	0.62
2	0.602	0.839	0.796	0.618	0.587
3	0.555	0.839	0.764	0.59	0.575
4	0.531	0.868	0.686	0.571	0.551
5	0.531	0.808	0.645	0.571	0.551
6	0.501	0.839	0.695	0.59	-
7	-	-	0.81	-	-
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-

Tabla D.4: Colección: Wiki (Partición 1). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.26.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.823	1	0.964	0.926	0.926	0.982
1	0.643	1	0.946	0.964	0.82	0.903
2	0.614	1	0.964	0.964	0.827	0.853
3	0.601	1	0.964	0.964	0.827	0.851
4	0.577	1	0.982	0.964	0.827	0.852
5	0.564	0.982	0.982	0.947	0.813	0.813
6	0.564	0.965	0.982	0.965	0.783	0.817
7	0.564	0.965	0.982	0.965	0.783	0.817
8	0.552	0.983	0.982	0.948	0.783	0.802
9	0.539	0.983	0.982	0.913	0.757	0.817
10	0.539	0.983	0.982	0.913	0.757	0.802

Tabla D.5: Colección: Wiki (Partición 2). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.27.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.823	0.91	0.911	0.895	0.873
1	0.643	0.733	0.704	0.626	0.694
2	0.614	0.701	0.652	0.627	0.605
3	0.601	0.701	0.652	0.58	0.605
4	0.577	0.648	0.661	0.566	0.605
5	0.564	0.648	0.649	0.553	0.58
6	0.564	0.661	0.661	0.553	0.567
7	0.564	0.661	0.626	0.553	0.567
8	0.552	0.661	0.626	0.538	0.567
9	0.539	0.661	0.626	0.522	0.553
10	0.539	0.661	0.602	0.508	0.515

Tabla D.6: Colección: Wiki (Partición 2). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 1. Figura Correspondiente: 5.28.

Iteración	Base	Res 10 %	Res 20 %	Res 30 %	Res 40 %	Res 50 %
0	0.823	1	0.964	0.926	0.926	0.982
1	0.619	1	0.982	0.982	0.827	0.816
2	0.499	0.983	0.982	0.982	0.771	0.816
3	0.486	-	0.982	0.982	0.74	0.756
4	0.486	-	0.982	-	0.754	0.728
5	0.486	-	0.982	-	0.741	0.698
6	0.471	-	-	-	0.741	0.698
7	0.471	-	-	-	-	0.698
8	-	-	-	-	-	
9	-	-	-	-	-	
10	-	-	-	-	-	

Tabla D.7: Colección: Wiki (Partición 2). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.29.

Iteración	Base	Res 60 %	Res 70 %	Res 80 %	Res 90 %
0	0.823	0.91	0.911	0.895	0.873
1	0.619	0.7	0.642	0.605	0.629
2	0.499	0.664	0.617	0.593	0.502
3	0.486	0.664	0.605	0.566	0.486
4	0.486	0.64	0.569	0.538	0.486
5	0.486	0.628	0.556	0.499	0.471
6	0.471	0.628	0.556	0.515	0.471
7	0.471	0.628	0.568	0.527	0.471
8	-	-	-	-	-
9	-	-	-	-	-
10	-	-	-	-	-

Tabla D.8: Colección: Wiki (Partición 2). Número de Documentos Iniciales:1. Número de Iteraciones:10. Número Máximo de Documentos Agregados por Iteración: 5. Figura Correspondiente: 5.30.