Universidad del Bío Bío
Facultad de ciencias Empresariales
Departamento de Sistemas de Información

Profesor Guía: Claudio Gutiérrez Soto

Evaluación Empírica de Algoritmos en Línea para Clustering de Documentos

"Informe de Proyecto de Título"

Fecha: 13 de octubre de 2017

Alumnos: Rodrigo Amigo Plaza
Alex Vidal Arévalo

Ingeniería Civil en Informática

Resumen

La recuperación de la información tiene como principal objetivo satisfacer una necesidad de información del usuario, a partir de un conjunto de recursos de información. Con el avance del tiempo y las tecnologías, la información está disponibles en muchos tipos de formatos y en cantidades enormes lo que hace necesario la automatización de esta actividad. Con esto aparecen los Sistemas de recuperación de información, los cuales utilizan el clustering de documentos con el propósito de mejorar su eficacia. Los métodos de clustering más utilizados para su funcionamiento son el clustering estático y el clustering dinámico. La hipótesis de clustering confirma que los documentos relevantes aparecen en los mismos grupos cuando estos tienden a ser similares entre sí. En investigaciones realizadas anteriormente, se concluye que el clustering dinámico de documentos si proporciona mejoras.

Esta investigación tiene como principal objetivo mejorar la precisión de los documentos recuperados, aplicados en un contexto dinámico. Además de investigar la efectividad de diferentes algoritmos de clustering (Single link, Complete link y Average link). Para lograr estos objetivos se realizan dos tipos de experimentos, los cuales se llevan a cabo en dos conjuntos de documentos, denominados documentos antiguos y documentos nuevos. El primer tipo de experimentos son los relacionados con la precisión de los documentos, mientras que el segundo tipo de experimento consiste en analizar el comportamiento de los algoritmos de clustering (Single link, Complete link y Average link). Se debe mencionar que cada uno de estos experimentos fueron realizados en dos etapas, la primera etapa consiste en la extensión de los experimentos realizados en la tesis realizada anteriormente (Delia Moncada – Frederick Lara), y la segunda etapa consiste en un cambio de la metodología utilizada para simular los juicios de usuario aplicados a los documentos, con el fin mejorar la precisión y la efectividad.

Los resultados para los experimentos relacionados a la precisión obtenida con la recuperación de documentos antiguos y la recuperación donde se unen documentos antiguos con documentos nuevos, en la primera etapa, indican que es posible conseguir una mejora en la precisión a medida que va aumentando la cantidad de documentos nuevos. Mientras que en la segunda etapa se demuestra que además de conseguir una

mejora en la precisión a medida que aumenta la cantidad de documentos nuevos, la precisión aumenta en relación a la primera etapa.

Para los experimentos respecto a la efectividad de los algoritmos de clustering, donde se considera la cantidad de documentos relevantes al recorrer el cluster generado solo con documentos antiguos, y la cantidad de documentos relevantes visitados al recorrer el cluster generado con la unión de los documentos antiguos y nuevos. Estos clústeres fueron generados con cada uno de los algoritmos de clustering, tanto en la primera etapa como en la segunda, para poder analizar cuál de los tres algoritmos obtuvo mejores resultados, siendo en ambas etapas el Average link el más efectivo.

Índice General

Índice de Figuras

Índice de Tablas

CAPÍTULO 1: INTRODUCCIÓN

La Recuperación de la información (RI), es una actividad de recolección de recursos informáticos relevantes para satisfacer una necesidad de información del usuario, a partir de un conjunto de recursos de información. Esta actividad implica varias tareas como la búsqueda, organización, análisis y almacenamiento de la información. Esta necesidad de información es realizada a través de consultas, las cuales buscan representar las expectativas del usuario. Con el avance del tiempo y las nuevas tecnologías, se puede encontrar información en innumerables fuentes como imagen, audio, vídeo, archivos, entre otros. Uno de los formatos más comunes en los que se puede encontrar información es en un documento o conjuntos de documentos. Sumado a lo anterior, la relevancia como respuesta a la necesidad de información que presenta un documento o conjunto de documentos es brindada por el usuario, lo que se conoce como juicios de usuario. En este proceso el usuario clasifica los documentos asociados a una consulta como “relevantes” o “no relevantes”. La comunidad de RI ha denominado al conjunto de documentos, conjunto de juicios de usuarios y conjunto de consultas como colección de pruebas. Hoy en día la información se encuentra distribuida en millones de documentos, por lo cual es factible buscar formas que permitan mejorar el proceso de recuperación de información. Como respuesta a lo anterior surgen los Sistemas de Recuperación de Información (SRI), cuya tarea es facilitar la búsqueda de documentos que satisfagan la necesidad de información por parte del usuario.

Un SRI entrega un conjunto de documentos, los cuales normalmente se encuentran en una lista por una puntuación decreciente, estos documentos están relacionados con la consulta que ha sido enviada por el usuario. La puntuación mencionada anteriormente corresponde a la medida de similaridad entre los documentos y la consulta (también puede ser aplicada entre documentos y entre consultas).

No obstante, no todos los documentos que han sido recuperados son relevantes para el usuario. Por lo cual, el orden en que son mostrados los documentos recuperados al usuario adquiere gran importancia para el usuario, siendo el caso ideal cuando los documentos relevantes aparecen juntos y en la parte superior de la lista.

Esta propiedad compuesta por los juicios de usuario y la posición de los documentos recuperados es llamada *Precisión*[1]. Esta situación se puede ejemplificar con el siguiente caso: para un conjunto de documentos *D*, que posee un subconjunto de documentos relevantes *SDR*, se aplica una consulta *Q* utilizando los sistemas de recuperación de información *SRI1* y *SRI2*. Por un lado, *SRI1* entrega una respuesta *R1* en la cual se muestra el *SDR* en la parte superior de la lista. Por otro lado, *SRI2* provee una respuesta *R2* en la cual el *SDR* aparece en la parte inferior de la lista. De acuerdo a los resultados anteriores, *R1* presentaría una mayor precisión, puesto que, si bien ambos sistemas retornaron a *SDR*, *R1* los dejó en la parte superior de la lista. En términos generales, el objetivo que persigue todo SRI es entregar los documentos relevantes al principio y los menos relevantes al final, es decir, altos niveles de precisión, lo cual es sumamente complejo y costoso.

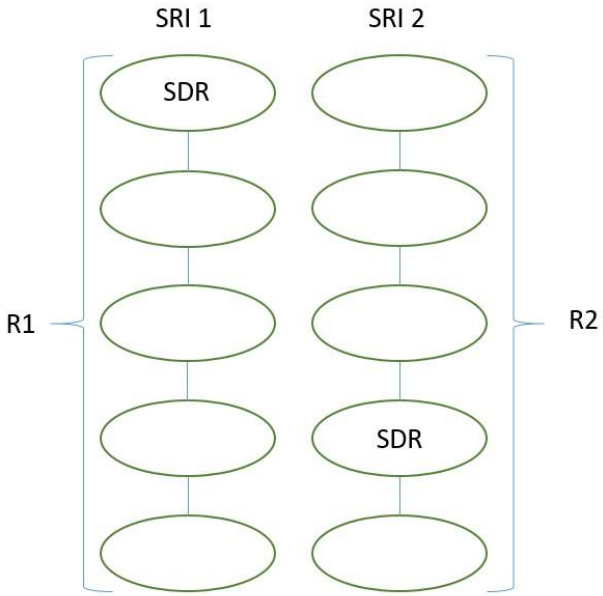

Figura Nº1: Representación gráfica del funcionamiento de dos SRI a una misma consulta.

En la actualidad, los SRI son cada vez más utilizados, siendo los Motores de Búsqueda Web los más masivos. Los motores de búsqueda web responden a millones de consultas día a día, en colecciones que contienen billones de documentos. Estos números están en permanente aumento, en consecuencia, la acción de encontrar documentos relevantes para una determinada consulta de un usuario resulta ser una tarea bastante compleja.

[1] La precisión es la cantidad de documentos relevantes recuperados dividido entre el total de documentos recuperados

En la literatura se puede encontrar una vasta cantidad de contribuciones asociadas a la RI, que abordan diferentes perspectivas como funciones de *matching*[2], *indexación*[3], modelos formales y retroalimentación relevante.

Las estructuras de datos comúnmente utilizadas son el clustering y los archivos invertidos. Los archivos invertidos utilizan palabras claves permitiendo obtener una respuesta eficiente, mientras que el clustering crea grupos de documentos que sean similares. El propósito de usar clustering en IR es mejorar la *eficiencia*[4] (tiempo) y la *efectividad*[5] (precisión) de los SRI. En términos generales, se pueden identificar dos categorías de clustering: clustering estático y clustering dinámico. Por un lado, se tiene al clustering estático, como el método tradicional de aplicar el método de cluster sobre una colección de documentos. En cambio, por el otro lado está en el clustering dinámico, donde se incluye información desde la consulta hacia el clustering de documentos. El clustering de consulta, es un tipo de clustering dinámico, el cual se dedica a encontrar consultas similares en un clúster. La similaridad que pueden tener los documentos, se encuentra dada por la intersección entre los términos que poseen los documentos. Las funciones de similaridad de Jaccard o coseno son las más utilizadas actualmente para medir la similaridad entre documentos (Salton y McGill, 1983).

Como fue mencionado anteriormente, se puede trabajar en recuperación de la información utilizando clustering estático. Sin embargo, hoy en día existe un crecimiento exponencial de documentos, lo cual hace que cada vez sea más necesario la implementación de estos algoritmos en un contexto dinámico (clustering dinámico). Si bien esta temática fue abordada en un proyecto anterior (Frederick Lara - Delia Moncada, 2016), en el desarrollo de este proyecto, los experimentos cuentan con dos etapas. En la primera etapa se extienden los experimentos realizados en el proyecto anterior, mientras que en la segunda etapa se implementa un algoritmo que permita simular los juicios de usuario bajo otras condiciones, para así obtener mejoras en los resultados de precisión y efectividad.

2 La función de matching pretende estimar si un documento es relevante de acuerdo a la consulta enviada por un usuario

3 El proceso de indexación corresponde a la representación y el almacenamiento de documentos

4 Capacidad de realizar o cumplir adecuadamente una función utilizando la menor cantidad de recursos

5 Capacidad de cumplir con los objetivos propuestos de forma satisfactoria sin tomar en cuenta los recursos utilizados

1.2 Objetivos Generales

Implementar un algoritmo de clustering en línea el cual, considerando los términos relevantes de los documentos que se encuentran en un clúster, permitiendo re-agrupar los clusters en razón al aumento o modificación del número de documentos (de los cuales, se desconoce la relevancia en el tiempo) con el objetivo de mejorar la precisión.

1.3 Objetivos Específicos

1.- Analizar los conceptos involucrados en recuperación de la información y de los algoritmos de clustering. Investigar sus principales características y comprender su problemática.

2.- Revisar una tesis realizada previamente (Frederick Lara - Delia Moncada, 2016), la cual servirá de base para extender los experimentos.

3.- Implementar y analizar el comportamiento de, al menos, tres algoritmos de clustering en línea (Single link, Complete link, Average link).

4.- Desarrollar un algoritmo que permita realizar los experimentos bajo otras condiciones.

5.- Obtener resultados experimentales, considerando como parámetros el número de documentos en los clústeres, el número de términos relevantes (para los documentos que ya están en el cluster), el número de documentos nuevos entre otros.

6.- Evaluar y comparar resultados obtenidos con el fin de determinar conveniencia del algoritmo en función de precisión y cantidad de documentos.

1.4 Descripción de capítulos

Este informe posee 6 capítulos (incluyendo el actual). Los cuáles serán descritos brevemente a continuación.

Capítulo 2

En este capítulo se realiza una introducción a los conceptos teóricos envueltos en la Recuperación de la Información (RI) y los Sistemas de Recuperación de la Información (SRI). Realizando un énfasis en los elementos más importantes para la realización de esta investigación.

Capítulo 3

En este capítulo se realiza una explicación del como los algoritmos son utilizados en la RI. Además, se realiza una breve descripción de los algoritmos de clustering jerárquico, los cuales serán utilizados a lo largo de esta investigación.

Capítulo 4

En este capítulo, se describe el funcionamiento que poseen los algoritmos en línea, también se analizara parte de la estructura del framework utilizado en la realización de este proyecto, poniendo énfasis en el proceso de creación de consultas y documentos y en la simulación de los juicios de usuario.

Capítulo 5

En este capítulo se muestran los resultados obtenidos para cada uno de los algoritmos implementados en las dos etapas experimentales que contempla esta investigación. Se analizan e interpretan los resultados para poder realizar las conclusiones en torno a estos.

Capítulo 6

En este capítulo se presentan las conclusiones generales de la investigación, así como las contribuciones que esta aporta.

CAPÍTULO 2: RECUPERACIÓN DE LA INFORMACIÓN

La recuperación de la información (RI), es una rama de las ciencias de la computación, que se refiere a la organización, estructuración, análisis, almacenamiento y búsqueda de la información. Existen varias definiciones para recuperación de la información en la literatura. Baeza-Yates y Ribeiro-Neto (1999) señala lo siguiente: "... la recuperación de la información de algún modo debe interpretar el contenido de los elementos de la información (documentos) en una colección, y clasificarlos según el grado de relevancia para la consulta del usuario. Esta interpretación del contenido de un documento implica la extracción de información sintáctica y semántica del texto del documento …".

Del párrafo anterior, hay tres aspectos importantes que deberían ser considerados de esa definición. En primer lugar, la existencia de un usuario que posee una necesidad de información, una colección de documentos con la cual este requerimiento es comparado, y finalmente la lista de documentos brindada por un sistema de recuperación de la información como respuesta a la consulta. En consecuencia, un sistema de recuperación de la información busca proporcionar un grupo de documentos que responda la necesidad de información del usuario. Un usuario, quien posee una necesidad de información, expresa esa necesidad a través de una consulta, la cual es ingresada en el sistema. Buscando procesar esta consulta, el SRI ejecuta una representación interna de ella. Una vez formada la representación de la consulta, es asociada con la colección de documentos. Como resultado, una lista de documentos clasificados es entregado al usuario. Por lo tanto, el usuario puede evaluar la lista final de documentos y si los documentos no le resultan relevantes, el usuario puede reformular su consulta o ingresar una nueva.

Varios modelos de RI han sido desarrollados, en los cuales es posible encontrar no sólo una descripción de cómo los documentos y consultas son representados, sino que también el método en que se realiza el matching entre las consultas-documentos. Los modelos más comunes en RI son el Booleano, El Espacio Vectorial (Salton, 1971; Salton et al., 1975), el Probabilístico (Robertson et al., 1981), y el Lógico (Van Rijsbergen, 1986).

El modelo de Espacio Vectorial representa los documentos y las consultas a través de vectores, basado en la frecuencia de términos en los documentos. La medida de Similaridad entre una consulta y un documento corresponde al coseno entre el término de la consulta

y los términos del documento. Este modelo es la base tanto para el trabajo experimental en el clustering de documentos como también para el desarrollo de esta tesis.

Los primeros sistemas de recuperación de información aparecieron con el propósito de apoyar la automatización de búsquedas de material en librerías por parte de los usuarios. Como respuesta al exponencial aumento de la información disponible mediante formato electrónico, los sistemas de recuperación de la información ampliaron su espectro a otros ámbitos. En consecuencia, existe una amplia gama de investigaciones basada en los SRI que puede ser encontrada en la literatura.

Algunas investigaciones se ocupan de los fragmentos provenientes de la información recuperada (Salton et al., 1993; Liu y Croft, 2002). Otras investigaciones basadas en XML, con técnicas de estudio que permiten la recuperación de segmentos de información desde datos estructurados (Fuhr et al.,2005; Fuhr y Lalmas, 2007). Así la World Wide Web se ha convertido en el medio más popular por la gente al momento de buscar información. Los SRI han sido desarrollados en internet como Motores de Búsqueda Web (WSEs por sus siglas en inglés), los cuales indexan una gran cantidad de información y proporciona una forma simple de acceder a la información por parte de los usuarios.

Actualmente, se han realizado una gran cantidad de trabajos que apuntan a explotar los atributos que caracterizan las páginas web, como la estructura HTML, la popularidad de las páginas Web, y estructuras de hipervínculos entre otros (Bharat y Henzinger, 1998; Kleinberg, 1999). En Broder (2002), se presenta una taxonomía la cual clasifica las consultas web en tres grandes categorías, navegacional, informacional y transaccional. En líneas generales, la categoría navegacional se refiere como los usuarios pueden encontrar una página web o documento en particular. En la segunda categoría, la información de las consultas provee al usuario con información relacionada a un tema en particular. En este caso el usuario puede estar interesado en más de un solo documento. Por último, las consultas transaccionales abordan los servicios de localización con los que cada usuario debe interactuar.

2.1 Representando documentos y consultas

Generalmente, los documentos son transformados por un sistema de recuperación de información desde su forma original a una representación interna propia, este proceso es llamado indexación. El propósito de este proceso es de proveer una representación de la

información lo más acertada posible. Para lograr este objetivo, un grupo de atributos de indexación son asignados a cada documento. La característica más relevante para un documento corresponde a una lista de palabras, la cual permite diferenciar entre ellas. Esto son comúnmente conocidos como *términos*. En efecto, un documento no es sólo representado por una lista de términos, sino que también se tiene acceso por términos que pertenecen a su lista. En orden de obtener el mayor nivel de representación que permita recuperar frases unitarias, o el uso de lingüística, semántica y conocimiento basado en métodos; deberían verse involucradas características complejas en el proceso de indexación, (Lewis y Jones, 1996). En esta investigación, la representación de los documentos es mediante una lista de términos, la cual es extraída de los documentos.

Antes de la indexación existe un proceso de normalización. El objetivo de este proceso es entregar únicamente los términos relevantes. Por ejemplo, las palabras con alta frecuencia en los documentos, llamadas *Stop-Word*, no serán consideradas en la indexación (Rijsergen, 1979). Stop-Word son las palabras conocidas como preposiciones (A, antes, de, desde, entre otras) y artículos (El, la, los, las). La ventaja principal de este proceso es reducir el tamaño del texto en aproximadamente cincuenta por ciento. Otro proceso realizado antes de la indexación consiste en remover sufijos de las palabras restantes del texto. Para lograr esto, se aplica un algoritmo de derivación para reducir las palabras a una raíz común. Por ejemplo, si se tienen las palabras “cardiovascular” y “cardiología”, el algoritmo de derivación reducirá estas palabras a “cardio”, la cual estará en el vocabulario de los términos indexados.

Buscando obtener los términos que son representativos de un documento contra otros documentos que pertenecen a la colección (en el vocabulario de los términos indexados), es necesario centrarse en la presencia de términos que no son frecuentes, a diferencia de los términos frecuentes que aparecen en todos los documentos. Para lograr este objetivo, el concepto de ponderación de la frecuencia inversa de un documento (IDF) fue introducido en (Jones, 1972). Así, el peso de un término en un documento es aumentado si aparece más a menudo en ese documento. Por el contrario, el peso de un término en un documento disminuye si éste aparece frecuentemente en otros documentos. Por lo tanto, para una colección de documentos que posee *N* documentos, si el término *i* ocurre en n_i documentos, entonces el peso *idf* de un término es dado por *log(N/ni).* Una función de ponderación del término corresponde a la combinación de los pesos de *tf* e *idf*, el cual comúnmente se conoce como peso *tf-idf* (Salton, 1971):

$$W_{ij} = \frac{\log(freq_{ij}+1)}{\log(lenght_{ij})} \; log \frac{N}{n_i}$$

W_{ij} = peso *tf - idf* del termino *i* en el documento *j.*

$freq_{ij}$ = frecuencia del termino *i* en el documento *j.*

$length_j$ = largo (en términos) de documento *j.*

N = número de documentos en la colección.

n_i = número de documentos que el termino *i* está asignado.

Una perspectiva de varios esquemas de ponderación como también medidas de evaluación en el dominio de RI son entregas por Salton y Buckley (1987). Es importante destacar que, en los experimentos relacionados a la simulación, el proceso de derivación ha sido omitido. Esto debido que cada experimento es construido en un entorno ideal. En consecuencia, el proceso de derivación no es necesario debido a que cada término es único, por lo tanto, no existe una raíz común entre los términos. Del mismo modo, los términos llamados stop-Word también han sido descartadas, porque los términos usados se asumen, son representativos.

Tres grandes categorías de estructuras de datos son usadas por los sistemas de recuperación de información para almacenar términos, índices lexicográficos (índices que son ordenados), estructuras de archivos en clusters, e índices basados en *hashing*[6]. No obstante, la estructura de datos más usada por los motores de búsqueda web corresponde al archivo invertido (Ouksel, 2002). Esta estructura corresponde a una lista, que contiene términos representativos para una colección de documentos, así el término que pertenece a la consulta es asociado con palabras clave (términos) que están en la lista. Por consiguiente, es factible localizar inmediatamente todos los documentos de la colección que poseen esta palabra clave (Rijsbergen, 1979).

2.2 Operaciones de Consultas

El objetivo principal para un SRI es brindar apoyo al usuario en la búsqueda de los

6 Una función de Hash es una caja negra que tiene como entrada una llave y como salida una dirección

documentos que satisfagan su necesidad de información. Los requerimientos de información son expresados de modo que puedan ser entendidos por el SRI. El modo de expresar estos requerimientos es denominado "*consulta".*

La formulación de una consulta puede ser realizada por algunos IRS, mediante el uso de operadores booleanos. Un ejemplo de este tipo de consultas puede ser:

((Simulación Y Algoritmos) NO Aleatorios)

La gran desventaja para SRI booleanos, es que sus resultados son poco intuitivos para los usuarios poco experimentados, por lo tanto, la formulación de una consulta ad-hoc pueden ser no efectivos (Sparck Jones and Willett, 1997)). En consecuencia, estos sistemas pueden ser reemplazados por sistemas que entreguen la formulación de la consulta mediante el lenguaje natural sin utilizar operadores específicos. Estos sistemas se basan en el mejor-match o en la búsqueda de similaridad, la cual es calculada para cada documento y para cada consulta.

Del mismo modo que los documentos, las consulta son procesadas antes de ser indexadas. Esto significa, que el procesamiento léxico y el peso ponderado son ejecutados por los SRI.

2.3 Matching entre documentos y consultas

Mediante el modelo booleano, un SRI encuentra el subconjunto de documentos de una colección de documentos, en el cual cada documento que pertenece a este subconjunto posee al menos un término de la consulta ingresada al SRI, los cuales son entregados sin clasificar. Por el otro lado, es posible encontrar sistemas que poseen mejores métodos de comparación y provean un resultado acorde a una puntuación de relevancia sobre la pertinencia del documento, respecto a la consulta. Así un SRI brinda una lista clasificada de documentos, ordenados de forma descendente al usuario.

En el modelo espacio vectorial (Salton y McGill, 1986), los documentos y las consultas son representados como vectores en un espacio multidimensional. La representación espacial corresponde a los términos indexados de la colección de documentos.

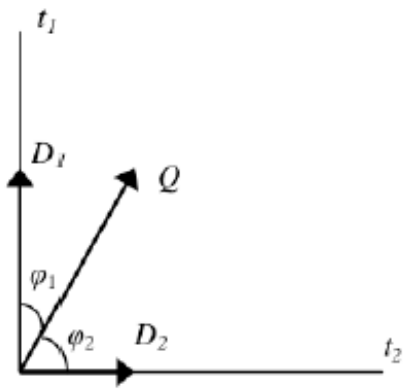

Figura Nº2: Representación de los documentos y una consulta utilizando el modelo espacio vectorial

En este modelo un documento y una consulta son representados. En el ejemplo, el vector de espacio es mostrado en dos dimensiones. Así, la dimensión corresponde a los términos t1 y t2. Los documentos D1 y D2 son representados en el espacio usando cada peso de los términos del documento como coordenadas. Los pesos corresponden al peso de la frecuencia de un término: D1 = t1, t1 y D2 = t2. Por otro lado, la consulta Q es representada como Q = t1, t2, t2. Además, los ángulos entre los vectores D1 - Q y D2 - Q son presentados en la Figura Nº2.

En este espacio, las medidas para cuantificar la similaridad entre las consultas y documentos pueden ser definidas. El mejor match entre una consulta en particular y una colección de documentos corresponde al documento más cercano respecto a la consulta de acuerdo a la medida de similaridad. En la literatura de la RI se puede encontrar un amplio rango de fórmulas que tratan la medición de distancias. Información más detallada sobre ellos puede ser encontrada en libros como (Rijsbergen, 1979; Salton y McGill, 1986).

Una forma simple de comparar un documento con respecto a una consulta es contar el número de términos en común entre ellos. De este modo, se puede asumir que ambos son representados como vectores de longitud n (donde n es el número de términos en la colección). Así, tenemos la siguiente medida:

$$\text{sim}(D,Q)=\sum_{i=1}^{n} D_i\, Q_i \qquad (2.1)$$

Esta fórmula es conocida como función de matching de *nivel de coordinación.* Por otro lado, otras medidas de similaridad pueden ser normalizadas de acuerdo al largo del documento y la consulta.

La medida que más utilizan por los SRI corresponde a la distancia del coseno. La razón de su popularidad recae en la interpretación geométrica del modelo vectorial.

$$\text{sim}(D, Q) = \frac{\sum_{ni=1} D_i Q_i}{\sqrt{\sum_{i=1}^{n} (D_i)^2 \sum_{i=1}^{n} (Q_i)^2}} \qquad (2.2)$$

La medida del coseno es la función que provee los ángulos entre los vectores del documento y la consulta (revisar la ecuación 2.2), y cuyo rango de valores se encuentran entre 0 (los vectores del documento y la consulta forman un ángulo de 90°, es decir, son diferentes) y 1 (los vectores de los documentos y la consulta forman un ángulo de 0°, es decir, son iguales). En la Figura Nº2, la similaridad entre los documentos D1 y D2 es 0, porque el ángulo entre los dos vectores es 90°. Esto significa que ellos no poseen términos en común.

2.4 Evaluación de los SRI

Existe una gran cantidad de metodologías que se encargan de la evaluación de los SRI. La tarea de evaluación implica una complejidad peculiar, esto debido a que envuelve diferentes puntos de diferentes áreas de investigación como cognición, estadísticas, diseño experimental, diseño de sistema, interacción humano-computador, entre otras. En esta sección, se presenta un resumen de las evaluaciones con énfasis en los aspectos centrales que constituyen esta investigación.

Existen variados aspectos que pueden ser evaluados en el proceso de RI. Estos aspectos deben estar relacionados, con la velocidad de un SRI, las interfaces y el nivel de interacción del usuario final, el formato de la información presentada al usuario, entre otros. Sin embargo, un punto relevante de esta tesis corresponde a la evaluación del número de documentos relevantes que entrega un SRI en respuesta a la consulta de un usuario. Las medidas de efectividad más utilizadas corresponden a *precisión* y *recuperación*. Precisión

representa la fracción de documentos recuperados que son relevantes, mientras que la recuperación corresponde a la proporción de documentos relevantes que son recuperados. De acuerdo a como se muestra en la Figura Nº3, precisión y recuperación pueden definirse como:

$$Precisión = \frac{(numero\ de\ documentos\ relevantes) AND (documentos\ recuperados)}{(numero\ de\ documentos\ recuperados)}$$

$$Recuperación = \frac{(numero\ de\ documentos\ relevantes) AND (documentos\ recuperados)}{(numero\ de\ documentos\ relevantes)}$$

Por lo general, estas medidas son expresadas entre los rangos 0 y 1, aunque pueden también ser expresadas como porcentaje (Chowdhury, 2010).

A partir de las definiciones previamente mencionadas, es necesario saber el número total de documentos relevantes en una colección de documentos para calcular la recuperación. Sin embargo, no siempre es posible calcular esta medida, esto debido a que involucra no solo una enorme cantidad de esfuerzo, sino que también de tiempo. Por otro lado, colecciones que permitan calcular esta medida pueden ser comercializadas o de libre disposición. Este tipo de colección de documentos puede ser encontrado junto a un grupo de consultas y juicios de usuario (la relevancia de un documento en relaciona una consulta). Por medio del uso de estas colecciones, los investigadores de RI tienen la oportunidad de evaluar sus propuestas con resultados empíricos y comparar sus resultados con los obtenidos por otros sistemas y propuestas.

En la Figura Nº3 se presenta un gráfico que muestra el balance entre recuperación-precisión (R-P). Esto significa que precisión y recuperación se encuentran relacionados de manera inversa. Cada vez que la precisión es incrementada, la recuperación disminuye y lo mismo ocurre de forma inversa. Esto se encuentra representado en gran parte de libros y manuales de RI.

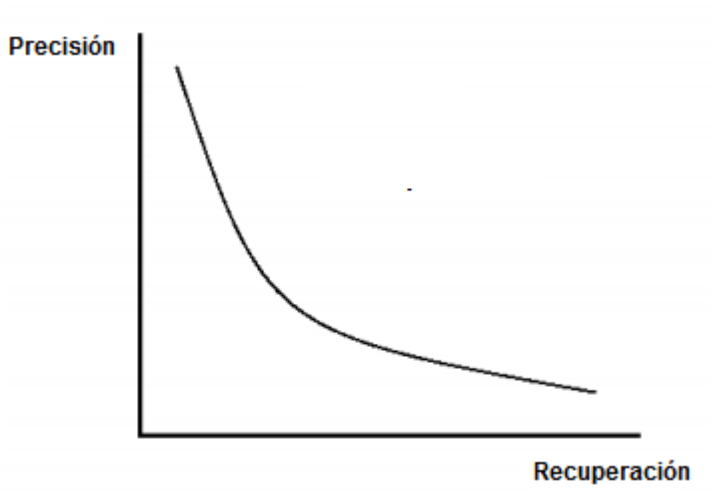

Figura Nº3: Grafico de recuperación vs precisión.

Dado el constante incremento de grandes colecciones, especialmente con la aparición de las Conferencias de Recuperación de Texto (TREC, por sus siglas en inglés) (Harman, 1993), que contienen miles y millones de documentos (que involucra el uso de muchos gigabytes de espacio en disco), se ha vuelto imposible proveer un juicio exhaustivo sobre la pertinencia de los documentos (Tombros et al., 2002).

Con el objetivo de abordar este problema, una técnica muy comúnmente usada y conocida en estos casos es denominada *pooling* (Harman, 1993). El centro de esta técnica se encuentra en la combinación de documentos de mayor clasificación de muchos SRI, donde una consulta en particular es ingresada en cada uno de ellos. De este modo, los juicios de usuario corresponden a una combinación de los conjuntos provenientes de los SRI. Podemos deducir que esta técnica es efectiva, cuando los documentos relevantes recuperados son representativos de todos los documentos relevantes disponibles en los SRI.

El tipo de evaluación que se estableció previamente, se basa en juicios proporcionados por algunos jueces expertos, basados en la relevancia actual (o algorítmico). En (Schamber et al., 1990; Barry 1994), la relevancia corresponde a un concepto multidimensional, el cual implica solo una dimensión. Inspirado por este trabajo, han sido presentadas otras dimensiones sobre el concepto de relevancia mediante la evaluación de metodologías con respecto a la utilidad entregada por los SRI. (Borlund y Ingwersen, 1997; Reid, 2000).

Es importante destacar, que las conclusiones extraídas de los resultados empíricos de las investigaciones de RI, son un problema que involucra un gran número de factores, como la elección de medidas adecuadas de rendimiento, la validación estadística de los resultados empíricos, entre otros. Para abordar estos problemas, se ha propuesto una metodología para obtener argumentos científicos de los experimentos de RI (Keen, 1992).

2.5 Elección de una medida

En el clustering de documentos, existe una gran variedad de medidas que se pueden emplear, debido a esto surge la necesidad de escoger una medida apropiada para el clustering. Van Rijsbergen (1979), advierte contra el uso de cualquier medida que no esté normalizada por la longitud de los vectores del documento bajo comparación. Además, Van Rijsbergen, Sneath y Soakl (1973)), realizaron una observación, la cual indica que las diferentes medidas de asociación y distancia son monótonas unas con otras. De lo anterior se puede concluir que, si un método de clustering que depende sólo del ordenamiento de los valores de similaridad, tendría resultados similares para todas estas medidas.

Willett (1983) comprobó la necesidad de normalizar las medidas de similaridad por la longitud de los vectores de documentos. En este estudio, utiliza tres conjuntos de documentos, cuatro medidas de similaridad (producto interno, coeficiente de Tanimoto, distancia del coseno y coeficiente de superposición) y cinco esquemas de ponderación de término. En los resultados obtenidos en estos experimentos, se puede apreciar la poca efectividad de las jerarquías obtenidas con medidas normalizadas. Finalmente, la medida que obtuvo una mejor tasa de efectividad de recuperación, fue la medida de distancia del coseno.

Griffiths et al. (1984) presentó pruebas adicionales de lo inadecuado que era emplear medidas no normalizadas. Se utilizó la distancia de Hamming y el coeficiente de Dice para medir las relaciones entre documentos. En ambos casos fueron medidos en una serie de experimentos. Los resultados que obtuvieron al emplear la distancia de Hamming fueron notoriamente menores a los que se obtuvieron en el coeficiente de Dice, resultados que se mantuvieron en todos los escenarios experimentales. A pesar de esto, los autores no reconocieron la importancia de la normalización.

Kirriemuir y Willet (1995), aplicaron cuatro métodos de clustering y cinco medidas de similaridad y distancia, métodos como Jaccard, distancia del coseno y distancia euclidiana eran parte de estos métodos. Los resultados de estos experimentos demostraron que el método de distancia del coseno y Jaccard fueron los más efectivos.

Rorvig (1999) realizó investigaciones en las que buscaba comparar diversas medidas de similaridad entre documentos. Tenía un mayor interés en investigar como un conjunto de medidas de similaridad actuaría como parte de una interfaz de recuperación de información visual. Se seleccionaron cinco temas TREC y los conjuntos de documentos utilizados para cada tema comprendían entre 421 y 586 documentos TREC.

Para cada tópico, se definió separar los documentos relevantes de los no relevantes, y de ese modo poder evaluar la calidad de las medidas de similaridad. En esta investigación fueron incluidas cinco medidas de similaridad (Asimetría, distancia del Coseno, Dice, Jaccard y Superposición), de las cuales se obtuvieron como resultado dos medidas que poseían un mayor éxito, estas medidas son la medida de la distancia del coseno y superposición en la estructura de la recuperación. Comúnmente la distancia del coseno y la distancia euclidiana son las medidas más utilizadas al momento de querer obtener la proximidad entre documentos en un espacio vectorial de documentos. Jones y Furmas (1987) estudiaron una gran cantidad de medidas de similaridad, incluyendo la de la distancia del coseno. De este estudio concluyeron que la comparación de documentos, a partir de su ángulo en un espacio vectorial, se asemeja a una comparación que se basa en su contenido tópico. Esto ocurre porque se expresa a través de relaciones de términos dentro del documentos.

Dubin (1996) indicó que las medidas angulares eran más sensibles a los pesos relativos de los atributos, mientras que las medidas de distancia lo son a los pesos absolutos. De otro modo, la distancia euclidiana ha sido criticada por Willet (1988), debido a que consideraba que dos documentos podrían ser muy similares, incluso si no comparten términos en común.

Luego de analizar algunas de las propuestas de diversos autores, se determina la distancia del coseno como la medida a utilizar en esta investigación. Esto debido a que es considerada por muchos autores como una de las mejores medidas de similaridad de documentos.

Por otra parte, la distancia euclidiana para el clustering ha sido criticado por Willett (1988), señalando que con esta medida dos documentos pueden considerarse muy similares, incluso si no comparten términos en común.

Después de observar las medidas propuestas por distintos autores, se determinó que para este proyecto se utilizará la medida de distancia del coseno, ya que varios autores la consideran como una de las mejores medidas de similaridad de documentos. A continuación, se dará a conocer más detalladamente esta medida.

2.6 Medida de distancia del coseno

Actualmente la medida de distancia del coseno es la más utilizada por los SRI. Esta se compone por cada término que se encuentran en todos los documentos, sin repetición.

El ejemplo propuesto por Martínez (2004), explica cómo es realizado el cálculo de la distancia del coseno.

Si nuestro SRI contiene los siguientes cuatro documentos:

D1: el río Danubio pasa por Viena, su color es azul.

D2: el caudal de un río asciende en invierno.

D3: el río Rhin y el río Danubio tienen mucho caudal.

D4: si un río es navegable, es porque tiene mucho caudal.

La matriz obtenida de términos y documentos dentro del modelo del Espacio Vectorial se muestra en la Tabla Nº1:

	Rio	Danubio	Viena	color	azul	caudal	invierno	Rhin	navegable
D1	1	1	1	1	1	0	0	0	0
D2	1	0	0	0	0	1	1	0	0
D3	2	1	0	0	0	1	0	1	0
D4	1	0	0	0	0	1	0	0	1

Tabla Nº1: Representación de matriz de términos y documentos en el espacio vectorial

Para la creación de esta tabla se realiza el proceso de derivación, con el propósito de eliminar las preposiciones, artículos, entre otros. que se encuentran en los documentos.

Sparck-Jones (1997), valoró la discriminación de un término en relación a otro. Esta generalidad que puede poseer un término de estar dentro de un conjunto de documentos y no en un documento en particular, y se pensó en valorar de mejor forma aquellos términos que son encontrados en una menor cantidad de documentos en comparación a los que aparecen en una gran cantidad de documentos, esto a raíz de que un término que se encuentra en una gran cantidad de documentos es poco útil al momento de querer representar el contenido de un documento. El valor de discriminación es medido mediante la frecuencia inversa de documentos *idf*. De este modo, al momento de construir la matriz con los términos y documentos son consideradas las siguientes definiciones:

n = Número de términos distintos en la colección de documentos

$\boldsymbol{tf_{ij}}$ = Número de veces que se repite el término *tj* en el documento D_i *(*Frecuencia que posee un término o tf).

$\boldsymbol{df_j}$ = Número de documentos en que se encuentra el término *tj*.

$\boldsymbol{idf_j}$ = el $\log(\frac{d}{df_j})$. donde d es el número total de documentos (frecuencia inversa del documento).

A cada término de cada documento se le asigna un peso, el cual se basa en la frecuencia en que aparecen en el conjunto de documentos, así como en un documento en particular. El peso de un término en un documento aumenta si este aparece con más frecuencia en un documento. Y el peso de un término disminuye si este aparece con más frecuencia en todos los demás documentos. El peso de un término es distinto a cero sólo si el término aparece en el documento. Para un conjunto grande de documentos que tiene numerosos documentos pequeños, es normal encontrar documentos en los que sus vectores contengan mayormente ceros.

Para medir el peso(d) que posee un término en un documento, se describe como la combinación de la frecuencia de término (*tf*), y la frecuencia inversa del documento (*idf*). Para obtener el valor de la *j*-esima entrada del vector que corresponde al documento *i*, se emplea la siguiente fórmula: $d_{ij} = tf_{ij} \; x \; idf_j$. La obtención de las frecuencias inversas de los términos en los documentos y utilización de esta fórmula sobre la matriz del ejemplo resultaría en la siguiente matriz de pesos.

Cálculo de frecuencias inversas:

idf(Río) = Log (4/4) = log (1) = 0

idf(Danubio) = Log (4/2) = log (2) = 0,301

idf (Viena) = Log (4/1) = log 4 = 0,602

idf (color) = Log (4/1) = log 4 = 0,602

idf (azul) = Log (4/1) = log 4 = 0,602

idf (caudal) = Log (4/3) = log 1,33 = 0,124

idf (invierno) = Log (4/1) = log 4 = 0,602

idf (Rhin) = Log (4/1) = log 4 = 0,602

idf (navegable) = Log (4/1) = log 4 = 0,602

A continuación, en la Tabla Nº 2, podemos observar un ejemplo de la matriz de peso de documentos con términos:

	Rio	Danubio	Viena	color	azul	caudal	invierno	Rhin	Navegable
D1	0	0,301	0,602	0.602	0.602	0	0	0	0
D2	0	0	0	0	0	0,124	0,602	0	0
D3	0	0,301	0	0	0	0,124	0	0,602	0
D4	0	0	0	0	0	0,124	0	0	0,602

Tabla Nº2: Matriz de peso de documentos con términos.

Una vez se calculan los pesos para cada término, se realiza el cálculo de la similaridad entre cada uno de los documentos (D1, D2, D3, D4). Para calcular la similaridad se utiliza la medida de distancia del coseno. El modo más simple de obtener la similaridad es con el producto escalar de los vectores (es decir, multiplicar los componentes de cada vector y suman sus resultados). Continuando el ejemplo se obtienen los siguientes resultados de similaridad:

Cálculo de similaridad:

Sim (D1, D2) = 0*0 + 0,301*0 + 0,602*0 + 0,602*0 + 0,602*0 + 0*0,124+0*0,602+ 0*0 + 0*0 = 0

Sim (D1, D3) = 0*0 + 0,301*0,301 + 0*0,602 + 0*0,602 + 0*0,602 +0*0,124 + 0*0 + 0*0,602 +

0*0 = 0,09

Sim (D1, D4) = 0*0 +0*0,301 + 0*0,602 + 0*0,602 + 0*0,602 + 0*0,124 +0*0+ 0*0 + 0*0,602 = 0

Sim (D2, D3) = 0*0 + 0*0,301 + 0*0 + 0*0 + 0*0 + 0,124*0,124 + 0*0,602 + 0*0,602 + 0*0 =

0,15376

Sim (D2, D4) = 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0,124 * 0,124 + 0*0,602 + 0*0 + 0*0,602 =

0,15376

Sim (D3, D4) = 0*0 + 0*0,301 + 0*0 + 0*0 + 0*0 +0,124*0,124 + 0*0 +0*0,602 + 0*0,602 =

0,15376

Para estos cálculos se establece que la similaridad entre los mismos documentos es 1, es decir D1 y D2 es lo mismo a D2 y D1. Por lo que no es necesario volver a calcular su resultado. Con estos valores obtenidos, se construye la matriz de similaridad, la cual se muestra en la Tabla Nº3.

	D1	D2	D3	D4
D1	1	0	0,09	0
D2	0	1	0,15376	0,15376
D3	0,09	0,15376	1	0,15376
D4	0	0,15376	0,15376	1

Tabla Nº3: Matriz de similaridad.

CAPÍTULO 3: CLUSTERING DE DOCUMENTOS PARA RI

3.1 Clustering de documentos en RI

El clustering ha sido realizado por los humanos desde hace mucho tiempo (WIllet,1988), el análisis de cluster o clustering es una técnica estadística multivariable que intenta agrupar conjuntos de objetos de modo que los objetos de un mismo grupo sean similares y estén en un mismo espacio, que usualmente es multidimensional. Los grupos de objetos se componen de tal forma que los objetos en el mismo cluster sean similares unos a otros y disímiles a los objetos de otros clusters (Gordon,1987). Las técnicas de análisis de cluster han sido ampliamente aplicadas en múltiples campos investigativos como las ciencias médicas, ciencias sociales, ciencias terrestres y ciencias de la ingeniería, entre otras (Anderberg, 1973). Actualmente, esta tarea se realiza completamente automatizada, gracias a las ventajas en tecnologías de la computación (Willet, 1988). Además, la aplicación del análisis de cluster se ha aplicado en RI no solo para el clustering de términos, sino que también para el clustering de documentos. El clustering de documentos es aplicado basado en los términos compartidos entre documentos. El clustering de términos entrega una representación de un grupo de términos que pertenece a un documento o consulta.

La similaridad entre los documentos es utilizada como base sobre la cual opera el clustering de documentos. Se puede medir la similaridad entre un par de documentos mediante la cantidad de términos que poseen en común. Uno de los primeros investigadores que sugieren el uso de clustering automático para RI fue Good (1958), esto según Van Rijsbergen (1979). La forma más común de aplicar el clúster de documentos es de forma estática, en donde se posee una colección completa de documentos, antes de realizar la consulta (clustering estático). Una manera distinta para agrupar los documentos, es la de agrupar los documentos que han sido recuperados por el SRI al momento de responder una consulta (clustering dinámico) (Preece,1973). Puede ocurrir, que en el clustering dinámico los grupos de documentos resultantes sean distintos al momento de realizar consultas distintas. Existen dos grandes tipos de clustering utilizados en los RI al momento de realizar el clustering de documentos, los cuales son jerárquicos y aglomerativo.

Comúnmente en el clustering aglomerativo, los documentos se encuentran representados mediante un vector en un espacio de n-dimensiones, en el cual *n* corresponde al número de términos que componen el vocabulario de indexación de la base de datos. Así, dando una colección de *N* documentos, el clustering aglomerativo crea *k* cluster que son mutuamente distintos entre sí, el valor de *K* es especificado a priori o se determina cómo parte del método de clustering. Suelen poseer bajos requisitos computacionales los métodos particionales, usualmente en el orden de 0(*N*) a 0(*N log N*) (en tiempo) para el clustering de N documentos (Willet, 1988). Como consecuencia, en los inicios de la investigación en el clustering de documentos, los métodos particionales fueron favorecidos, dado que ofrecían el potencial de aumentar la eficiencia de un SRI (Rocchio, 1966; Salton, 1971). La idea base para esta metodología es la de elegir una partición inicial de documentos, posteriormente se irán agregando objetos a los cluster para obtener una mejor partición (Anderberg, 1973) (. por ejemplo, número de clusters, tamaño del cluster) y poder lograr una solución óptima (Salton y Wong, 1978; Willett, 1988). Los primeros experimentos dieron como resultado que la efectividad de las búsquedas basadas en particiones de documentos es significativamente menor a las búsquedas basadas en archivos no agrupados (Salton, 1971).

Según la literatura de IR, existen muchas aplicaciones para el clustering jerárquico, que se han realizado considerando el uso de términos simple. Por el contrario, enfoques recientes se encargan de la representación de documentos a través de unidades frasales utilizando diferentes niveles de análisis lingüístico. Este tipo de clustering ha sido ampliamente aceptado en la comunidad de RI (Willet, 1988), debido a que entrega una base teórica sólida. Normalmente, cada documento *"D" es* representado como un vector $D= \{d_1, d_2, d_3, d_n\}$, donde *d* es el término y *n* es el número de términos que componen el vocabulario de indexación de la colección de documentos. Todos los términos que componen el vocabulario de indexación son utilizados en la representación de la indexación (Rijsbergen, 1979). Antes del clustering de documentos, se realizan algunos procesos como el de derivación o normalización, entre otros. Luego, se obtiene un peso relativo sobre la importancia de cada término, considerando toda la colección de documentos para incrementar la efectividad (Salton y Buckley, 1987). Una vez obtenida la representación apropiada del conjunto de documentos para el cluster, es necesario tener una medida acorde al grado de similaridad para todos los posibles pares de documentos que pertenecen a este conjunto. Para lograr esto, un gran número de medidas que cuantifican el parecido entre los objetos, pueden ser aplicadas para poder entregarles una categorización.

Existen cuatro principales clases de medidas: asociación, disimilitud, probabilístico y coeficiente de correlación (Sneatch y Sokal, 1973). La mayor parte de la literatura se enfoca en asociación y disimilitud, mientras el uso de coeficiente de correlación y probabilístico en documentos de clustering es limitado.

Es importante destacar que, en esta investigación, los experimentos son restringidos a unidades de indexación de un solo término. Se empleará el método de clustering jerárquico, debido a que es apropiado a nuestra problemática y, además, es ampliamente aceptado por la comunidad de IR (Willett, 1988), debido a que entrega una base teórica sólida.

3.2 Método de clustering jerárquico

El método de clustering jerárquico proporciona una clasificación en una estructura de árbol o dendograma, la cual está compuesta por objetos (en este caso, documentos), donde los objetos de un mismo cluster son muy similares entre sí. Al mismo tiempo están en clusters más grandes que contienen menos objetos similares.

Suponiendo tenemos un conjunto de documentos el cual está representado por X, que se agrupan $X = \{x_1, x_2, x_3, \ldots, x_n\}$. Los documentos son representados por vectores n-dimensionales, en el que cada dimensión es un término de indexación.

El clustering de X en m conjuntos de documentos se puede definir como $R = \{C_1\ C_2, \ldots, C_m\}$, de modo que se cumplan las siguientes condiciones:

- *Cada cluster* C_i debe contener al menos un documento $C_i \neq \emptyset,\ i = 1, \ldots, m$
- Todos los clústeres unidos corresponden al conjunto $X = U_{i=1}^{m} \quad C_i = X$
- Dos clústeres no tienen documentos en común: $C_i \cap C_j = \emptyset,\ i \neq j,\ i, j = 1, \ldots, m$

Se dice que un agrupamiento R_1 que contiene k clústeres está anidado en el agrupamiento R_2 que contiene clústeres $r<k$, si cada cluster en R_1, y al menos un cluster de R_1 es un subconjunto propio de R_2 (Theodoridis y Koutroumbas, 1999). Por ejemplo, el cluster $R_1 = \{\{x_1, x_3\}, \{x_4\}, \{x_2, x_5\}\}$ está anidado en $R_2 = \{\{x_1, x_3, x_3\}, \{x_2, x_5\}\}$. Por otra parte, R_1 no está anidado dentro de $R_3 = \{\{x_1, x_4\}, \{x_3\}, \{x_2, x_5\}\}$ (Theodoridis y Koutroumbas, 1999).

Los métodos jerárquicos se pueden separar en dos categorías, aglomerativos y divisivo. El método aglomerativo proporciona series de (N-1) uniones, para una colección de N

documentos, donde los resultados del clustering se forman desde la parte inferior a la superior de la estructura. En el método divisivo, una agrupación inicial única, se divide en grupos más pequeños de documentos consecutivamente (Rijsbergen, 1979). Normalmente el método divisivo entrega como resultado clasificaciones que solo contienen una característica única, donde los documentos que pertenecen a un cluster específico deben contener términos específicos para poder pertenecer a esta clasificación (Sneath and Sokal, 1973); (Rijsbergen, 1979); (Gordon, 1987). Por otro lado, en el método aglomerativo, no es necesario tener términos específicos para pertenecer a una clasificación. Es importante señalar que el método aglomerativo han sido ampliamente utilizados en RI, y que el clustering jerárquico aglomerativo (HACM, en inglés) aún son utilizados en este campo (Willet, 1988).

Los métodos aglomerativos se pueden distinguir en dos diferentes teorías, los basados en teoría de matrices, y los basados en teoría de grafos. Bajo la premisa de que los métodos basados en matrices son los más utilizados, serán los utilizados en esta investigación.

Generalmente el procedimiento genérico que realizan los métodos jerárquicos aglomerativos es (Murtagh, 1983):

1.- Determinar la similaridad entre los documentos.

2.- Formar un cluster de los dos objetos más cercanos.

3.- Redefinir la similaridad entre el cluster nuevo con los demás objetos, sin alterar las demás similaridades.

4.- Repetir los pasos 2 y 3 hasta que todos los objetos se encuentren en el cluster.

Algunos métodos aglomerativos no aplican con exactitud el tercer paso mencionado anteriormente. En cada paso *t* del proceso de clustering, la matriz de similaridad *S*(*X*), que en un inicio es de tamaño *NxN*, es convertida en *(N-t) x (N-t)*. Así, la matriz en el paso *t,* es derivada de la matriz $S_{t-1}\,(X)$, eliminando las dos filas y columnas correspondientes a los nuevos documentos fusionados, y es añadida una nueva fila y columna que posee las nuevas similaridades formadas a partir de los documentos fusionados y con todos los documentos que no fueron afectados.

Al utilizar un método de clustering jerárquico, es posible representar sus resultados en forma de un dendograma (Jardine & Sibson, 1971) (Figura N°4). Un dendograma corresponde a un árbol con niveles numéricos que se encuentran asociados a sus ramas.

Los valores numéricos que se observan en la Figura N°4 corresponde al nivel de similaridad en los que se forman los clústeres. Es posible trazar una línea perpendicular al eje de similaridad en cualquier nivel de similaridad. De este modo, cada rama del árbol que se corta por la línea, representa un cluster que consiste en elementos en el subárbol con raíz en esa rama. Todos los documentos se encuentran en un solo cluster en el nivel más bajo e similaridad.

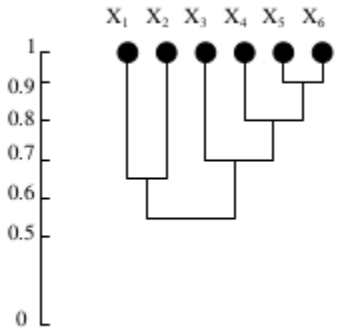

Figura N°4: Dendograma de similaridad.

Es posible encontrar una gran cantidad de lecturas, las cuales brindan una visión detallista en relación a los aspectos de eficiencia de los diversos métodos de clustering. Es importante destacar que la eficiencia es una propiedad del algoritmo que implementa el método de clustering (Jardine & Sibson, 1971). Van Rijsbergen (1979) señaló que en ocasiones es útil diferenciar el método cluster de su algoritmo, pero también mencionó que en el contexto de RI esta diferenciación es menos importante.

Cuatro algoritmos de clustering jerárquicos serán presentados a continuación, los cuales han sido investigados más ampliamente en investigaciones de RI. Los algoritmos corresponden a Single Link, Complete Link, Average Link y Ward.

3.3 Algoritmo Single Link

Al utilizar el algoritmo Single Link, la distancia entre dos clusters se encuentra determinada por un par de elementos, estos elementos (uno en cada cluster) son lo más cercanos entre

sí. La distancia más corta que se obtiene entre objetos en cada paso causa la unión de los cluster en que se encuentran dichos objetos.

$$\underset{x \in X, y \in Y}{D\,(X,\ Y)} = Min\ d(x, y)$$

Donde:

d (x, y) es la distancia entre los elementos $x \in X, y \in Y$.

X e *Y* son dos conjuntos de elementos (clusters)

Un inconveniente que tiene la utilización de este método, es llamado "fenómeno de encadenamiento", el cual suele producir clusters de forma alargada, en los cuales los elementos cercanos pertenecientes a un mismo cluster poseen distancias pequeñas, pero los elementos ubicados en los extremos opuestos de un cluster pueden poseer distancias mucho mayores entre sí que en relación a elementos ubicados en otros cluster. Esto puede traer dificultades al momento de definir las clases con las cuales se buscará dividir la información.

3.4 Algoritmo Complete Link

En el algoritmo Complete Link cada elemento comienza en su propio grupo. Luego, los clusters se combinan secuencialmente en grupos más grandes, hasta que todos los elementos terminan estando en el mismo grupo. En cada paso se combinan los grupos separados por la distancia más corta. Este valor corresponde al menor valor de distancia, del total de distancias más largas entre todos los clusters existentes.

La expresión del algoritmo Complete Link es:

$$\underset{x \in X, y \in Y}{D\,(X,\ Y)} = Max\ d(x, y)$$

En donde:

d (x, y) es la distancia entre los elementos $x \in X, y \in Y$.

X e *Y* son dos conjuntos de elementos (clusters)

El método Complete Link evita el inconveniente producido en Single Link llamado fenómeno de encadenamiento (clusters alargados), en donde los clusters formados mediante single link pueden ser forzados a estar juntos debido a un solo elemento que se encuentra cercano, incluso cuando muchos de los elementos en cada cluster se encuentran muy distanciados entre sí. En términos generales, Complete Link suele formar clusters compactos los cuales poseen un diámetro similar.

3.5 Algoritmo Average Link

En el algoritmo Average Link, la similaridad entre dos clusters corresponde a la similaridad entre todos los pares de documentos, en donde un documento se encuentra en un cluster y el otro documento es parte el otro cluster.

Como resultado, para la aplicación de Average Link, los clusters formados se basan en la similaridad promedio, y por lo cual no es posible deducir el máximo o mínimo de similaridad que existe entre los clusters (Voorhees, 1985). Como consecuencia, un gran rango de comparaciones realizadas en estudios por diferentes investigaciones, Sneath y Sokal (1973) apuntaron a que el método Average Link es el más usado de los métodos jerárquicos.

CAPÍTULO 4: ENTORNO DE LA INVESTIGACIÓN

La etapa experimental de esta investigación consta de dos tipos de experimentos, el primero se evalúa la precisión de un SRI mientras que en el segundo se evalúan los algoritmos de clustering aplicados a los resultados que entrega el mencionado SRI. Además, ambos experimentos se realizan bajo dos diferentes condiciones de funcionamiento del SRI.

4.1 Algoritmo en Línea

Como se menciona anteriormente, en la primera parte de los experimentos se evalúa la precisión de un SRI, esto se realizó a través de un algoritmo en línea. Para comprender los experimentos realizados de mejor forma se debe entender el funcionamiento de un algoritmo en línea y su relación con la precisión.

Un algoritmo en línea permite su ejecución sin la necesidad de tener todos los parámetros de entrada al inicio de su ejecución, es decir, los datos de entrada son recibidos mientras este se está ejecutando. No obstante, esta característica también puede ser una desventaja, debido a que, al no conocer toda la información al comienzo de su ejecución, se hace imposible encontrar su valor óptimo. Debido a esto que para saber si un algoritmo en línea es competitivo se utiliza una constante *c,* donde la *competitividad = c * Óptimo.* Esto se puede comprender fácilmente a través del conocido algoritmo en línea de *aprender a esquiar.*

Supongamos que aprender/arrendar el equipamiento cuesta *$x* al día y comprarlo cuesta *$y*. Además, *y=cx* para un entero $c \geq 1$. La persona no sabe si le gustará el nuevo deporte y al final de cada día decidirá si arrienda o compra. El Ó*ptimo* es cuando se sabe cuántos días se utilizará el equipamiento. Supongamos serán *t* días.

Si $tx < y$, conviene arrendar, y si $tx \geq y$ entonces conviene comprar. Mientras que si $tx = y \Rightarrow t = c$, comprar o arrendar tienen el mismo valor). Esto tal como es expuesto en la figura Nº5:

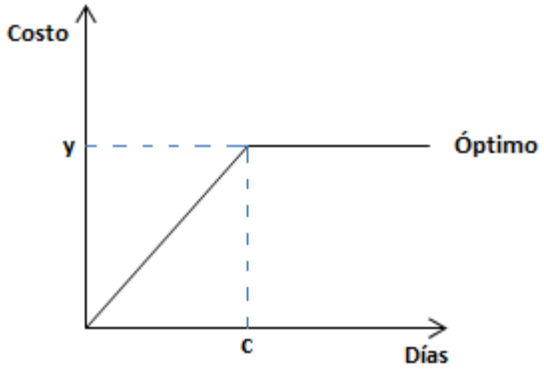

Figura Nº5: Grafico ejemplo *aprender a esquiar.*

Supongamos que arrendamos hasta que $c = \frac{y}{x}$ arriendos, y luego compramos si decimos seguir aprendiendo a esquiar (día c+1).

Esta situación se puede observar por medio de la Figura Nº6:

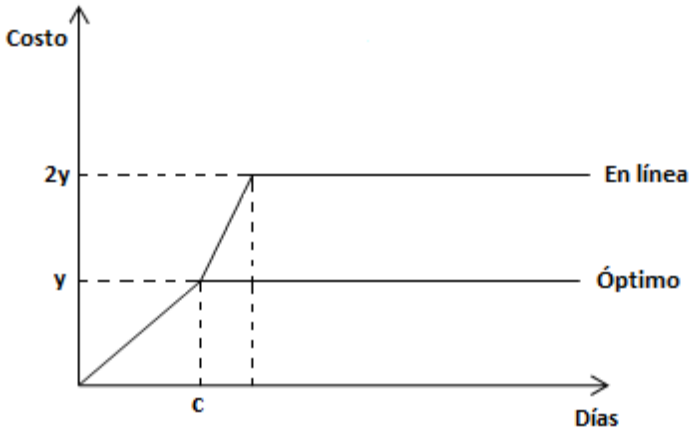

Figura Nº6: Grafico ejemplo optimo v/s en línea.

$$Competitividad = \begin{cases} 1, t \leq c \\ 2, t > c \end{cases}$$, con *t*=Días.

Ahora, supongamos que arrendamos hasta *K* días y luego compramos el equipo (Antes teníamos *K = c).*

En la figura Nº7 se muestra el “peor caso" de este ejemplo.

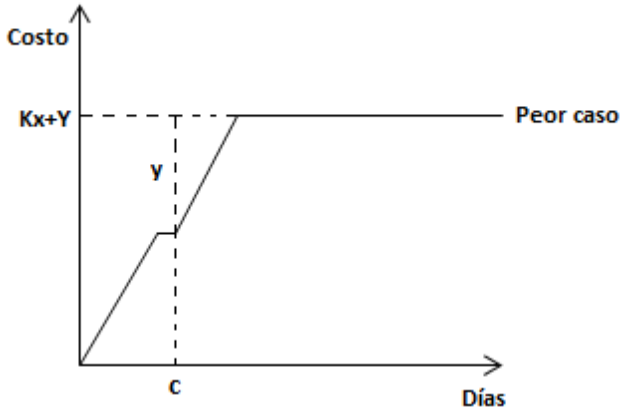

Figura Nº7: Grafico del "peor caso" para el ejemplo de *aprender a esquiar*.

Ningún valor de *K* puede tener competitividad 1. (Basta con escoger *t = k+1)*

$$Competividad = Max\left(\frac{kx+y}{tx}, \frac{kx+y}{y}\right)$$

k = 0, t = 1 => t = c ≥ 2

si kx ≤ y tenemos que

$$\frac{kx+y}{kx} \geq 2\ (k = t)$$

Y, si kx > y entonces

$$\frac{kx+y}{y} \geq 2$$

Entonces ningún algoritmo determinístico tiene *c < 2* => usar *kc* es óptimo. Pero un deportista "extremo" no es tan conservador, así que decide utilizar un algoritmo aleatorio y en ese caso la competitividad estará basada en el promedio sobre los algoritmos y no en el peor caso.

Sea A_i el algoritmo que arrienda *i-1* días y compra el *i-ésimo* día, sea π_i, la probabilidad de usar A_i *(i ≥ 1).*

Supongamos que la competitividad promedio de todas las A_i sea a los más *α*. Entonces tendremos que escoger la distribución *π* tal que:

Costo (en línea) ≤ αtx…t ≤ c

Costo (en línea) ≤ αy…t > c

$$Costo\ (en\ linea) = \sum_{i \geq 1} \pi_{\ i}\ costo\ (A_i) = \sum_{i \geq 1} \pi_{\ i}\ ((\ i-1)x + y)$$

Resolviendo el sistema de ecuaciones para la igualdad tenemos

$$\pi_i = \begin{cases} \dfrac{\alpha - 1}{c}\left(\dfrac{c}{c-1}\right)^i, & i = 1 \ldots c\ (Geométrica) \\ 0, & i > 0 \end{cases}$$

Pero

$$\sum_{i \geq 1} \pi_i = 1 => \ \alpha = \frac{1}{\left(1 + \left(\frac{1}{c-1}\right)\right)^c - 1} + 1$$

(En este caso *c* puede ser real y el tiempo es continuo)

Si c = 2 => α = $\frac{4}{3}$ < 2 (Mejor que el caso determinismo)

Si

$$c \to \infty, \left(1 + \frac{1}{c-1}\right)^c \ \to e$$

$$\alpha = \frac{e}{e-1} \approx 1.58$$

Así que

$$\alpha \ < 2\ \forall\ c$$

4.2 Pseudocódigo algoritmo en línea

En este proyecto se implementó un algoritmo en línea, el cual tiene como propósito mejorar la precisión de un conjunto de documentos. Con este algoritmo, se determina la cantidad de documentos nuevos necesarios para que exista un incremento en la precisión y se determine cuando rehacer el clúster. A continuación, se explicará el pseudocódigo utilizado para comprender de mejor forma su funcionamiento.

Definición de funciones:

- generarDocumentosAntiguos: Función que posee como parámetros de entrada el conjunto de términos antiguos, a partir del cual se genera el conjunto antiguo de documentos.
- generarDocumentosNuevos: Función que recibe como parámetro de entrada el conjunto antiguo de términos y el conjunto nuevo de términos. Los documentos nuevos generados están compuestos por un 66% de términos antiguos y un 33% de términos nuevos.
- generarConsultas: Función que recibe como parámetro de entrada el conjunto antiguo de documentos. Las consultas construidas a partir de este conjunto son retornadas por la función.
- aplicarConsultaADocumentos: Función que recibe como parámetro una consulta y un conjunto de documentos. Los documentos más similares a la consulta, ordenados del mas similar al menos similar son retornados por la función.
- distribucionZeta: Función que recibe como parámetro de entrada los documentos recuperados al realizar una consulta. Es retornada la relevancia para cada uno de los documentos antiguos recuperados, siendo 1 si es documento es relevante y un 0 si no es relevante.
- agregarDoc: Función que recibe como parámetro de entrada un documento relevante recuperado al aplicar una consulta y una lista de documentos relevantes. Se comprueba si el documento relevante se encuentra en la lista de documentos relevantes, en caso de no estarlo, es agregado.
- calcularPrecision: Función que recibe como parámetro la relevancia que poseen los documentos recuperados al aplicar una consulta sobre un conjunto de documentos. Es retornada la precisión calculada.

- extraerTerminos: Función que recibe como parámetro los documentos antiguos relevantes. Los términos que poseen los documentos relevantes son retornados por la función.
- porcentajeTerminosRelevantes: Función que recibe como parámetro de entrada un documento nuevo y la lista de términos relevantes. El porcentaje de términos relevantes presentes en el documento nuevo es retornado.
- algoritmoProbabilistico: Función que recibe como parámetro los documentos recuperados al ejecutar una consulta sobre la lista de documentos unidos. Es retornada por la funciona la relevancia para los documentos recuperados, en donde es retornado 1 si el documento es relevante y un 0 en caso de no ser relevante. El algoritmo mantiene la relevancia de los documentos antiguos (asignada por la distribución Zeta) y es aplicado sobre los documentos nuevos un algoritmo probabilístico que simula los juicios de usuarios.

Definición de conjuntos y variables donde:

- terminosAntiguos corresponde a un conjunto de términos antiguos.
- terminosNuevos corresponde a un conjunto de términos nuevos.
- documentosAntiguos corresponde a un conjunto en el cual se almacenan documentos antiguos, inicialmente documentosAntiguos = Ø.
- documentosNuevos corresponde a un conjunto donde son almacenados documentos nuevos, inicialmente documentosNuevos = Ø.
- consultas corresponde al conjunto donde se almacenan consultas, inicialmente consultas = Ø.
- documentosRecuperados corresponde a un conjunto donde se almacenan documentos antiguos recuperados al ejecutar una consulta sobre un conjunto de documentosAntiguos, inicialmente documentosRecuperados = Ø.
- relevanciaDocumentosAntiguos corresponde a un vector donde es almacenada la relevancia de los documentos recuperados al ejecutar una consulta sobre el conjunto documentosAntiguos, inicialmente relevanciaDocumentosAntiguos = Ø.
- documentosRelevantes corresponde a un subconjunto de documentosAntiguos, donde inicialmente documentosRelevantes = Ø.
- precisionAntiguaParcial corresponde al vector donde se almacena la precisión de los documentos antiguos para cada una de las consultas, inicialmente precisionAntiguaParcial = Ø.

- listaTerminosRelevantes corresponde a un conjunto donde se almacenan términos relevantes, donde inicialmente listaTerminosRelevantes = Ø.
- subconjuntoDocumentosNuevos corresponde a un subconjunto donde son almacenados los documentos nuevos que tienen una cantidad de términos relevantes >= 50%, inicialmente subconjuntoDocumentosNuevos = Ø.
- listaDocumentosUnidos corresponde al conjunto donde se almacenan documentos antiguos y documentos nuevos, inicialmente listaDocumentosUnidos = Ø.
- documentosRecuperadosUnidos corresponde a un conjunto donde son almacenados documentos antiguos y documentos nuevos, inicialmente listaDocumentosUnidos = Ø.
- documentosRecuperadosUnidos corresponde a un conjunto donde son almacenados documentos antiguos y documentos nuevos recuperados al ejecutar una consulta sobre el conjunto listaDocumentosUnidos, donde inicialmente documentoRecuperadosUnidos = Ø.
- relevanciaDocumentosUnidos corresponde a un vector donde se almacena la relevancia de los documentos recuperados al ejecutar una consulta sobre el conjunto listaDocumentosUnidos, inicialmente relevanciaDocumentosUnids = Ø.
- precisionNuevaParcial corresponde a un vector donde es almacenada la precisión de los documentos unidos para cada consulta, inicialmente precisionNuevaParcial = Sea precisionNuevaParcial un vector donde se almacena la precisión de los documentos unidos para cada consulta, inicialmente precisionNuevaParcial = ∅.
- i,j corresponden a variables enteras.
- porcentaje, precisionAntigua y precisionNueva corresponden a variables Reales.

```
1.  terminosAntiguos;
2.  terminosNuevos;
3.  documentosAntiguos <- generarDocumentosAntiguos(terminosAntiguos);
4.  documentosNuevos <- generarDocumentosNuevos(terminosAntiguos, term
   inosNuevos);
5.  consultas <- generarConsultas(documentosAntiguos);
6.  for i <- 1 to i <= consultas.length do
7.          documentosRecuperados <- aplicarConsultaADocumentos(consult
   as[i], documentosAntiguos);
8.          relevanciaDocumentosAntiguos <- distribucionZeta(documentos
   Recuperados);
9.          for j <- 1 to j <= documentosRecuperados.length do
10.                         if relevanciaDocumentosAntiguos[i] = 1 then
11.                 documentosRelevantes<-
   agregarDoc(documentosRecuperados[i], documentosRelevantes);
12.                         ifEnd
13.                 forEnd
14.                 precisionAntiguaParcial <- calcularPrecision(relevanc
   iaDocumentosAntiguos);
15.         forEnd
16.         listaTerminosRelevantes <- extraerTerminos(documentosRelevan
   tes);
17.         for i <- 1 to i <= documentosNuevos.length do
18.                 porcentaje <- porcentajeTerminosRelevantes(documentos
   Nuevos[i], listaTerminosRelevantes);
19.                 if porcentaje >= 0,5 then
20.                 subconjuntoDocumentosNuevos <- subconjuntoDocumentosN
   uevos U documentosNuevos[i];
21.                 ifEnd
22.         forEnd
23.         listaDocumentosUnidos <- documentosAntiguos U
   subconjuntoDocumentosNuevos;
24.         for i <- 1 to i <= consultas.length do
25.                 documentosRecuperadosUnidos <- aplicarConsultaADocume
   ntos(consultas[i], listaDocumentosUnidos);
26.                 relevanciaDocumentosUnidos <- algoritmoProbabilistico
   (documentosRecuperadosUnidos);
27.                 precisionNuevaParcial <- calcularPrecision(relevancia
   DocumentosUnidos);
28.         forEnd
29.         precisionAntigua <- 0;
30.         precisionNueva <- 0;
31.         for i <- 1 to i <= precisionAntiguaParcial.length do
32.                 precisionAntigua <- precisionAntigua + precisionAntig
   uaParcial[i];
33.                 precisionNueva <- precisionNueva + precisionNuevaParc
   ial[i];
34.         forEnd
35.         precisionAntigua <- precisionAntigua / precisionAntiguaParci
   al.length;
36.         precisionNueva <- precisionNueva / precisionNuevaParcial.len
   gth;
37.                 if precisionAntigua > precisionNueva then
38.                         //conviene rehacer el cluster
39.                 else
40.                         //no conviene rehacer el cluster
41.         ifElseEnd
```

4.3 Framework de simulación

Desde nuestro punto de vista, normalmente los mismos documentos relevantes pueden ser utilizados para responder consultas similares. Dicho esto, asumimos que la mayoría de los documentos relevantes para una consulta antigua pueden ser relevantes para una nueva consulta similar, y los anteriores juicios de usuario podrían ser utilizados para responder esta consulta.

En esta sección se presenta un framework de simulación (Gutiérrez-Soto, 2016), el cual permite simular colecciones tradicionales de RI, las cuales se forman por un conjunto de documentos, un conjunto de consultas y juicios de usuarios para los documentos del conjunto de documentos. Por lo tanto, el principal objetivo de este framework es entregar un entorno ideal, que permita evaluar bajo diferentes enfoques, el comportamiento de la aplicación de una misma consulta al conjunto de documentos antiguos y al conjunto de documentos nuevos. El funcionamiento del framework se basa en las leyes de las ciencias de la información de Bradford y Zipf, las que se utilizan en diversos campos de investigación.

4.4 Creación de consultas y documentos

Para la creación de consultas y documentos se emplea el framework propuesto por Gutiérrez-Soto (2015). Para lograr la simulación de un ambiente de algoritmos en línea, se realiza la construcción de dos conjuntos de documentos. El primer conjunto corresponde al a los documentos antiguos, el cual simula los documentos ya existentes en el cluster, mientras que el segundo conjunto corresponde a los documentos que van apareciendo a través del tiempo.

Tanto el conjunto de documentos antiguos como el de documentos nuevos requieren un listado de términos para ser formados. Para eso se cuenta con dos listas de términos, una lista que contiene los términos antiguos y otra con los términos nuevos. Cada termino está compuesto por un mínimo de 3 y un máximo de 7 letras, y cada letra se escoge utilizando una distribución uniforme. Además, cada termino es único y la intersección de términos entre ambas listas es vacío. El conjunto de documentos antiguos está compuesto con la

lista de términos antiguos, cada documento está formado por un mínimo de 15 y un máximo de 30 términos. En cambio, el conjunto de documentos nuevos está formado con la lista de términos antiguos y la lista de términos nuevos. Cada documento nuevo se forma con un 66% de términos antiguos y un 33% de términos nuevos, e igualmente que los documentos antiguos la cantidad de términos varía entre 15 y 30. De este modo, si un documento tiene 24 términos, 16 de estos términos corresponden a la lista de términos antiguos y 8 corresponden a la lista de términos nuevos.

Para crear la lista de consultas, se escoge un documento antiguo para cada consulta utilizando una distribución uniforme. Los términos que conforman la consulta se eligen del documento antiguo bajo la misma distribución, además una consulta está compuesta por una cantidad de 8 términos.

4.5 Simulación de juicios de usuario

Para simular los juicios de usuario, que son los que determinan si un documento es o no relevante ante una consulta determinada, se utiliza la distribución Zeta. Esta distribución entrega un enfoque discreto de la ley de Bradford. Esta ley dice que, entre la producción de artículos de revistas, hay un número heterogéneo de artículos donde los artículos relevantes están en pocas revistas, mientras que un número de artículos relevantes se difunden en una gran cantidad de revistas. Aplicando esto, para una consulta dada quiere decir que los documentos más relevantes deben estar al comienzo de la lista, debido a que su similaridad con la consulta es más alta, mientras que unos pocos documentos relevantes deben estar distribuidos en la parte inferior en la lista de documentos. Esto con el fin de simular este escenario, la distribución Zeta se utiliza para determinar los documentos relevantes para un conjunto de documentos antiguos. Además de lo anteriormente mencionado, en este proyecto se busca simular un contexto en línea, esto significa que se van agregando documentos a través del tiempo, lo cual hace necesario tener un conjunto de documentos nuevos. De este modo se van agregando documentos del conjunto de documentos nuevos, al conjunto de documentos antiguos, creando así, un nuevo conjunto que contiene documentos nuevos y documentos antiguos, al aplicar una consulta una consulta a este nuevo conjunto de documentos, se recuperan los documentos que hicieron match con la consulta. Para los documentos recuperados que pertenecen al conjunto de

documentos antiguos, se mantiene la relevancia calculada anteriormente por la distribución Zeta. Mientras que para los documentos recuperados pertenecientes al conjunto de documentos nuevos se hace necesario implementar un algoritmo que permita calcular la relevancia para estos documentos.

La primera etapa de los experimentos, se utiliza el algoritmo probabilístico implementado por (Delia Moncada y Frederick Lara, 2016). Este algoritmo considera dos criterios. En primer lugar, se tomó en cuenta la posición en la que se encuentran los documentos nuevos en la lista de documentos recuperados, esto debido a que la posición en que se encuentran los documentos recuperados en la lista indican su nivel de match con la consulta, en donde los que se encuentran ubicados en los primeros lugares poseen un mayor match, de este modo, un documento que se encuentra en la posición *i* posee una mayor probabilidad de ser relevante que uno ubicado en la posición i+1. En segundo lugar, se considera la cantidad de términos relevantes presentes en el documento, ya que entre más términos relevantes tenga el documento nuevo, existe una mayor probabilidad, de que este documento sea relevante.

En la segunda etapa de los experimentos, se implementa un algoritmo del tipo Monte Carlo, el cual tiene como propósito mejorar los valores de precisión obtenidos con el algoritmo utilizado en la primera etapa. Este algoritmo utilizado en la segunda etapa de experimentos, además de tomar en cuenta los criterios del algoritmo utilizado en la primera etapa, considera los términos relevantes que poseen las consultas que se repiten con mayor frecuencia.

CAPÍTULO 5: EXPERIMENTOS

5.1 Entorno experimental

En los experimentos realizados en esta investigación se utiliza el framework de simulación implementado por Gutiérrez-Soto (2015). Este framework se utiliza para simular un contexto dinámico, el cual consta de dos conjuntos de documentos, el primero contiene el conjunto de documentos que están agrupados en un cluster, se denominara como *"documentos antiguos"*, el segundo corresponde a los documentos que se van apareciendo constantemente con el paso del tiempo, denominados "*documentos nuevos"*, esto con el fin de evitar pérdida de información (por ejemplo, documentos no considerados). Además, se construye una lista con términos antiguos, y otra con los términos nuevos, ambas con 1400 términos cada una. A partir de estos términos se genera la lista de documentos antiguos (con la lista de términos antiguos) y la lista de documentos nuevos (con la lista de términos antiguos y nuevos). Las consultas fueron construidas a partir del conjunto de documentos antiguos. Cabe destacar, que además estos experimentos fueron realizados con 2100 términos nuevos y 2100 términos antiguos. La construcción de los términos, documentos y consultas se explica en profundidad en la sección 4.4.

5.2 Procedimiento a utilizar

Una vez obtenida la relevancia para los documentos antiguos recuperados es posible construir una lista con todos los términos relevantes (se consideran relevantes todos los términos que pertenecen a un documento relevante). Esta lista de términos relevantes será utilizada para construir el conjunto de documentos nuevos, como cada documento de este conjunto contiene un 66% de términos antiguos y un 33% de términos nuevos, se podrá utilizar la lista de términos relevantes antiguos para saber cuántos términos antiguos de cada documento nuevo son relevantes.

Debido a que se conoce la cantidad de términos relevantes presentes en cada documento del conjunto de documentos nuevo, es posible generar una lista con todos los documentos nuevos que posean más de un 50% de términos relevantes. Posteriormente esta lista se

une con el conjunto de documentos antiguos formando así una nueva lista de documentos que contiene a todo el conjunto de documentos antiguos y a la lista de documentos nuevos que contienen el 50% de términos relevantes. Esta nueva lista se denomina "*lista de documentos unidos".*

Una vez generada la lista de documentos unidos, se aplica la misma consulta que se aplicó al conjunto de documentos antiguos. Luego de haber aplicado esta consulta, se obtiene una nueva lista de documentos recuperados que hicieron la mayor cantidad de match con la consulta, esta nueva lista contiene tanto documentos antiguos como documentos nuevos. Para los documentos antiguos que se encuentran en la lista de documentos se mantiene la relevancia obtenida con la distribución Zeta, en cambio para obtener la relevancia de los documentos nuevos se utilizan ambos algoritmos probabilísticos mencionados anteriormente.

Después de obtener la relevancia para el conjunto de documentos antiguos y la lista de documentos unidos, se necesita calcular la precisión de estos. Esto tiene como objetivo determinar si añadir documentos nuevos al conjunto de documentos antiguos logra mejorar la precisión.

Documentos: Corresponde a los documentos más similares (utilizando medida de distancia del coseno) con la consulta. En la tabla Nº4, el documento 1 (D1) es la más similar a la consulta.

Z: Indica si los documentos son o no relevantes

X: Es la cantidad de documentos relevantes parciales divididos por la cantidad de documentos parciales.

n: corresponde a la cantidad de documentos relevantes acumulados.

m: Corresponde a la cantidad de documentos (30).

Documentos	Z	X
D1	1	1/1
D2	0	½
D3	1	2/3
D4	0	2/4
D5	0	2/5
D6	1	3/6
. .		. .
Dm	0	n/m

Tabla Nº4: Procedimiento para calcular la precisión.

$$Precision = \frac{\sum_{i=1}^{30} x_i}{m}$$

Una vez realizado el cálculo de la precisión para ambos conjuntos de documentos, se debe analizar si es que la precisión obtenida para la lista de documentos unidos mejoro o no respecto a la precisión obtenida para el conjunto de documentos antiguos. En el caso de que la precisión mejore se almacenan dos listas de documentos relevantes. En la primera lista se deben almacenar todos los documentos relevantes del conjunto de documentos antiguos y en la segunda listas se deben almacenar todos los documentos relevantes de la lista de documentos unidos.

5.3 Etapas experimentales

Esta sección se divide en dos etapas experimentales, en la primera etapa se extienden los experimentos realizados en la tesis de (Delia Moncada – Frederick Lara, 2016), luego en la segunda etapa se realizan cambios en el algoritmo probabilístico que determina los juicios de usuario, esto utilizando un algoritmo del tipo Monte Carlo, con el objetivo de mejorar los valores de precisión y eficacia entregados en la primera etapa.

Para ambas etapas experimentales existen dos tipos de experimentos, el primero consiste en los experimentos relacionados a la precisión de dos conjuntos de documentos, el primer conjunto de documentos corresponde a documentos antiguos y el segundo conjunto de

documentos corresponde a documentos antiguos y nuevos. Este experimento tiene como objetivo determina la existencia de una mejora en la precisión al aplicar las consultas a ambas listas de documentos, para así, determina la conveniencia para rehacer un clúster. El segundo experimento muestra los resultados de clustering donde se espera determinar con cuál de los algoritmos de clustering resulta más conveniente rehacer el cluster.

5.4 Experimentos de precisión

Como se mencionó anteriormente, los experimentos de precisión se dividen en dos etapas, en la primera se mostrarán los resultados de precisión asociados al algoritmo probabilístico. En la segunda parte se encontrarán los resultados de precisión luego de aplicar el algoritmo de Monte Carlo al algoritmo probabilístico.

Es importante destacar que, para cada uno de estos experimentos, fueron realizadas un total de diez pruebas. Esto con el fin de obtener un promedio para analizar de mejor manera su comportamiento. También, para cada experimento se encuentra una tabla donde se muestra los siguientes valores:

Precisión Antigua: Corresponde al promedio de la precisión calculada con la relevancia de los documentos recuperados al aplicar la consulta al conjunto de documentos antiguos.

Precisión Nueva: Corresponde al promedio de la precisión calculada con la relevancia de los documentos recuperados al aplicar la consulta al conjunto de documentos unidos.

Además, para cada experimento se encuentra un gráfico, donde en las ordenadas se representa la mejora de precisión y en las abscisas el aumento de documentos.

5.4.1 Experimentos de precisión primera etapa

Experimento Nº1: En el primer experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de

términos es de 1400. En la Tabla Nº5 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº8.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
700	700	0,3352712	0,5504101
700	1400	0,3369326	0,6574356
700	2100	0,3362888	0,710401
700	3500	0,3344499	0,7489703

Tabla Nº5: Resultados del primer experimento de precisión de la primera etapa.

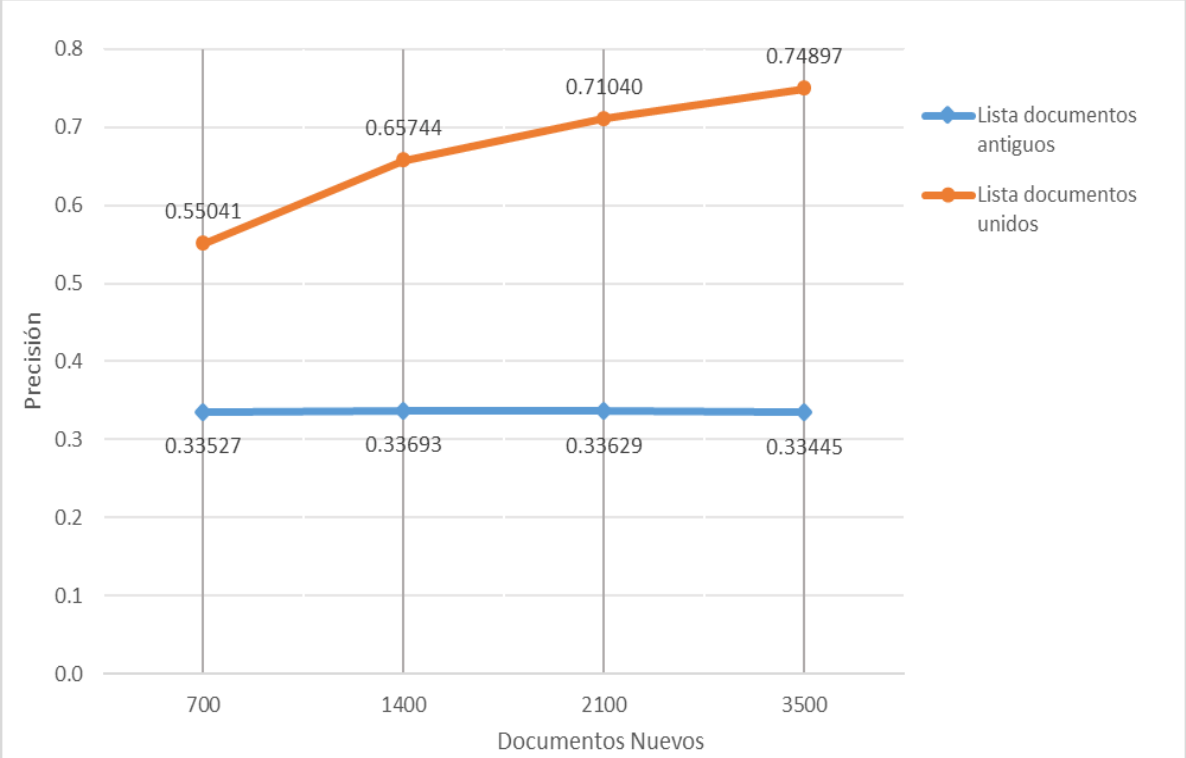

Figura Nº8: Gráfico del primer experimento de precisión de la primera etapa.

Experimento Nº2: En el segundo experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº6 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº9.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
1400	700	0,3363133	0,4386003
1400	1400	0,336783	0,4967937
1400	2100	0,3333048	0,5494354
1400	3500	0,331805	0,5877046

Tabla Nº6: Resultados del segundo experimento de precisión de la primera etapa.

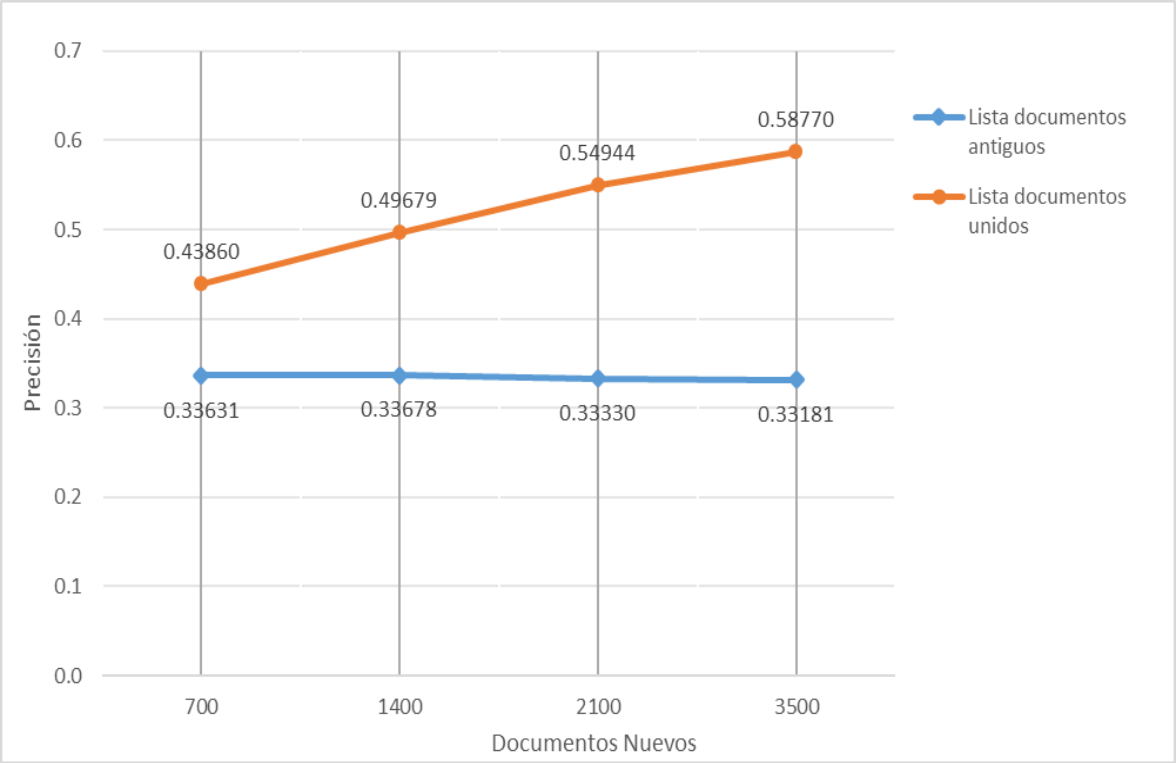

Figura Nº9: Gráfico del segundo experimento de precisión de la primera etapa.

Experimento Nº3: En el tercer experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº7 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº10.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
2100	700	0,3339704	0,3919826
2100	1400	0,3363242	0,4310496
2100	2100	0,3382678	0,4495592
2100	3500	0,3334236	0,4726842

Tabla Nº7: Resultados del tercer experimento de precisión de la primera etapa.

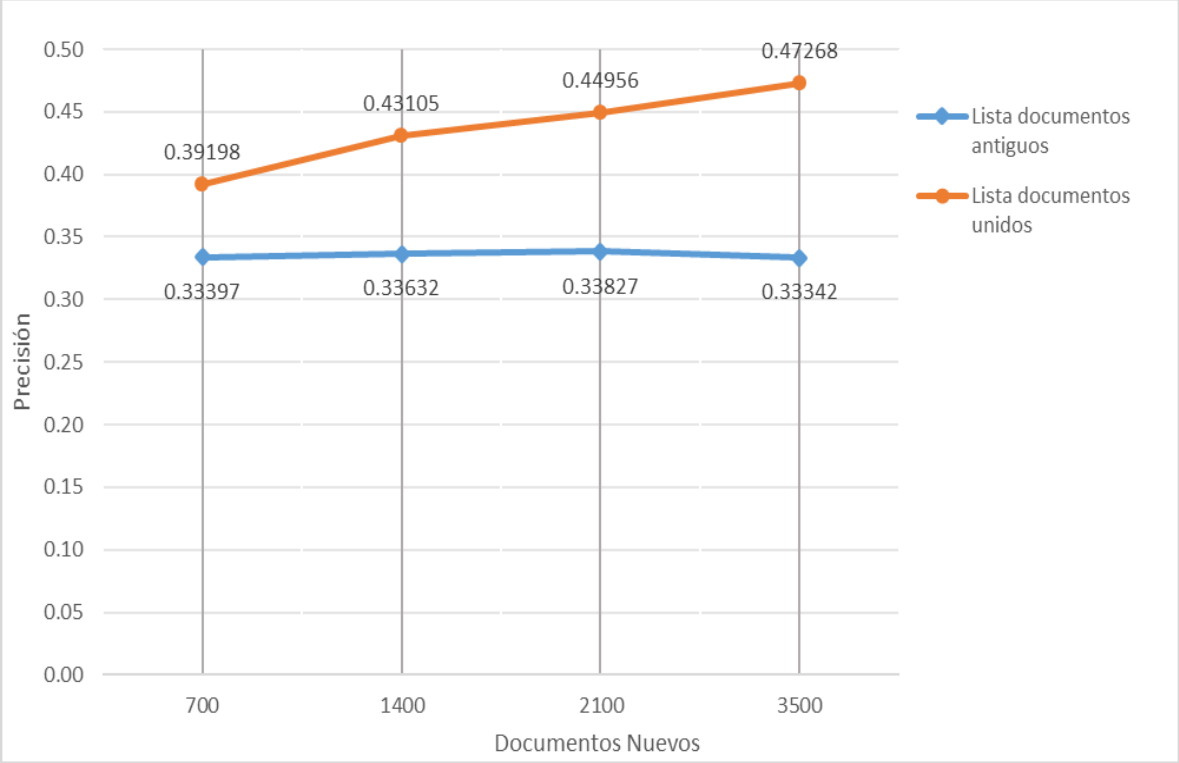

Figura Nº10: Gráfico del tercer experimento de precisión de la primera etapa.

Experimento Nº4: En el cuarto experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº8 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº11.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
3500	700	0,3404357	0,360393
3500	1400	0,3341964	0,366371
3500	2100	0,3358703	0,3715892
3500	3500	0,3358216	0,3809708

Tabla Nº8: Resultados del cuarto experimento de precisión de la primera etapa.

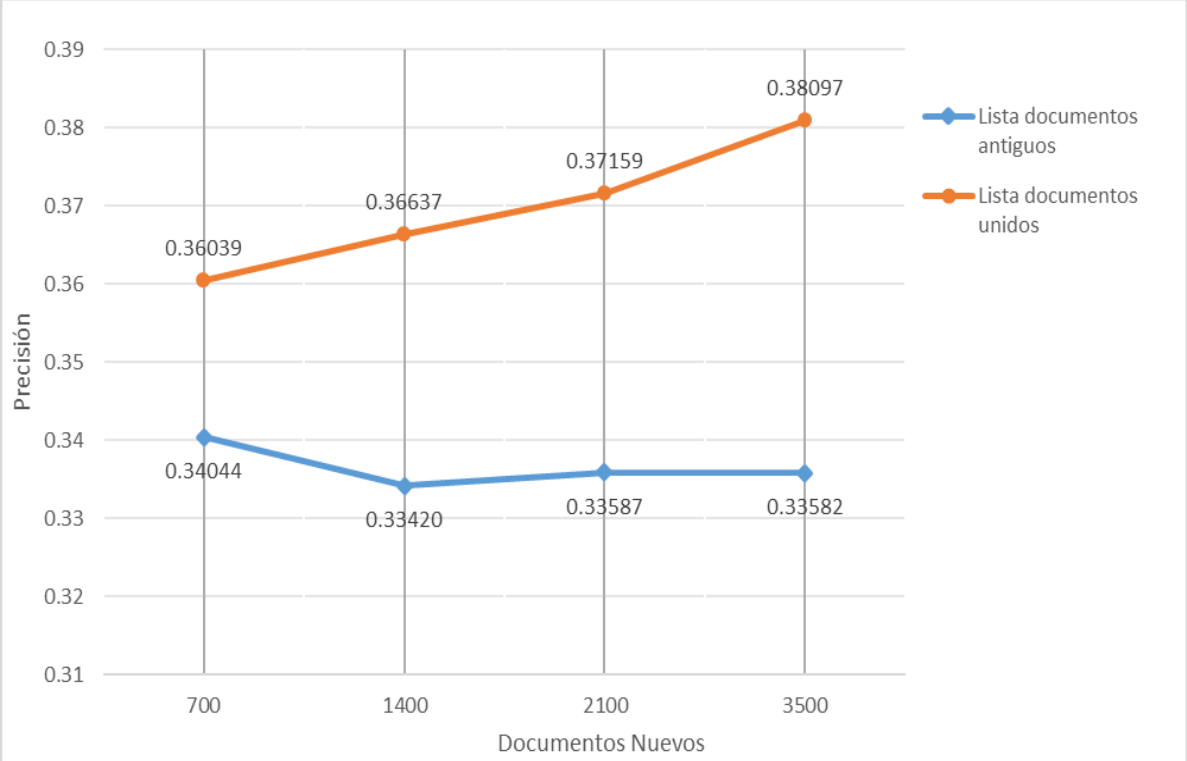

Figura Nº11: Gráfico del cuarto experimento de precisión de la primera etapa.

Experimento Nº5: En el quinto experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 2100. En la Tabla Nº9 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº12.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
700	700	0,3432065	0,4967619
700	1400	0,3542008	0,5799531
700	2100	0,3503717	0,6790887
700	3500	0,346915	0,696848

Tabla Nº9: Resultados del quinto experimento de precisión de la primera etapa.

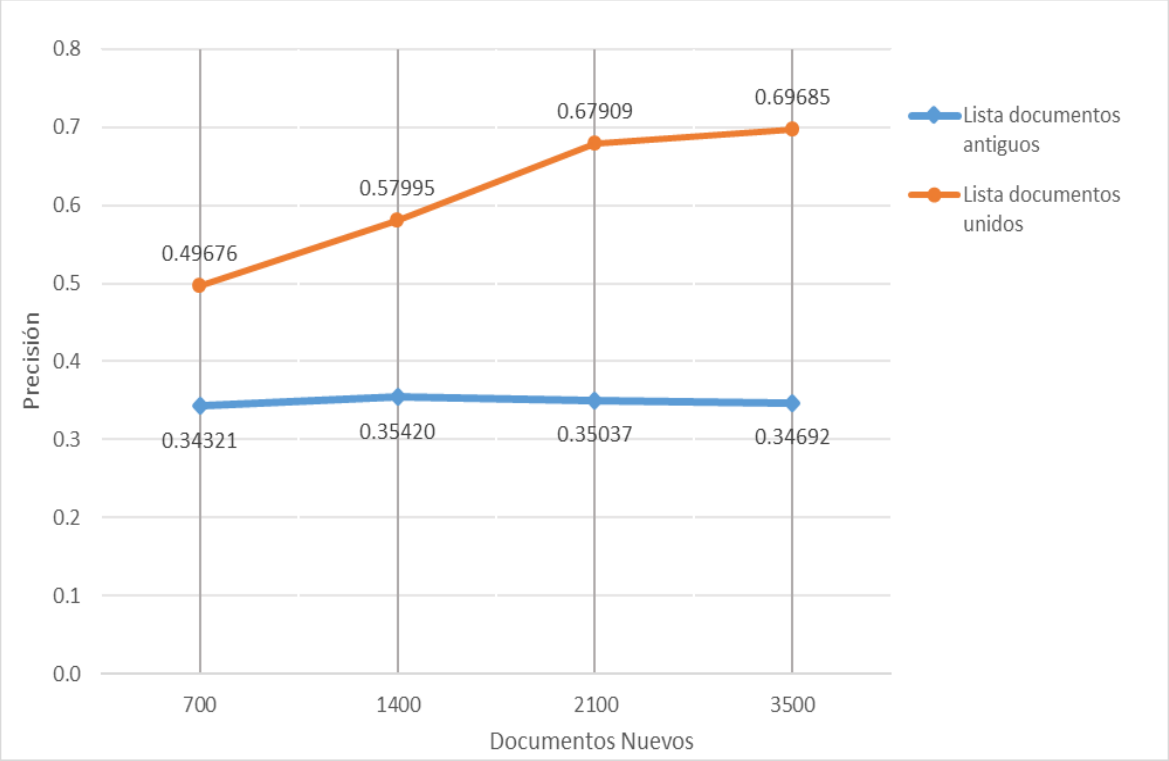

Figura Nº12: Gráfico del quinto experimento de precisión de la primera etapa.

Experimento Nº6: En el sexto experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 2100. En la Tabla Nº10 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº13.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
1400	700	0,3370268	0,4241249
1400	1400	0,3400808	0,4939214
1400	2100	0,3404267	0,5394049
1400	3500	0,3338448	0,5843974

Tabla Nº10: Resultados del sexto experimento de precisión de la primera etapa.

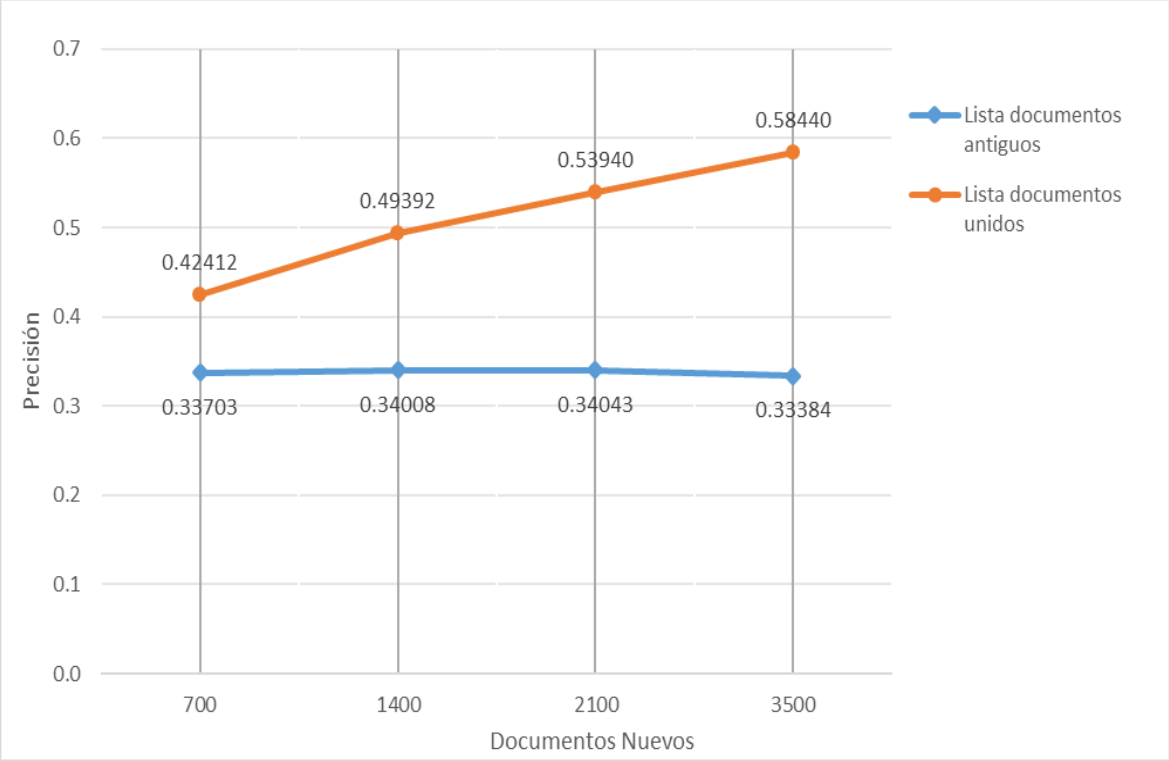

Figura Nº13: Gráfico del sexto experimento de precisión de la primera etapa.

Experimento Nº7: En el séptimo experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 2100. En la Tabla Nº11 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº14.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
2100	700	0,3379662	0,4044717
2100	1400	0,3296276	0,4166494
2100	2100	0,3401104	0,4688024
2100	3500	0,3309999	0,5065928

Tabla Nº11: Resultados del séptimo experimento de precisión de la primera etapa.

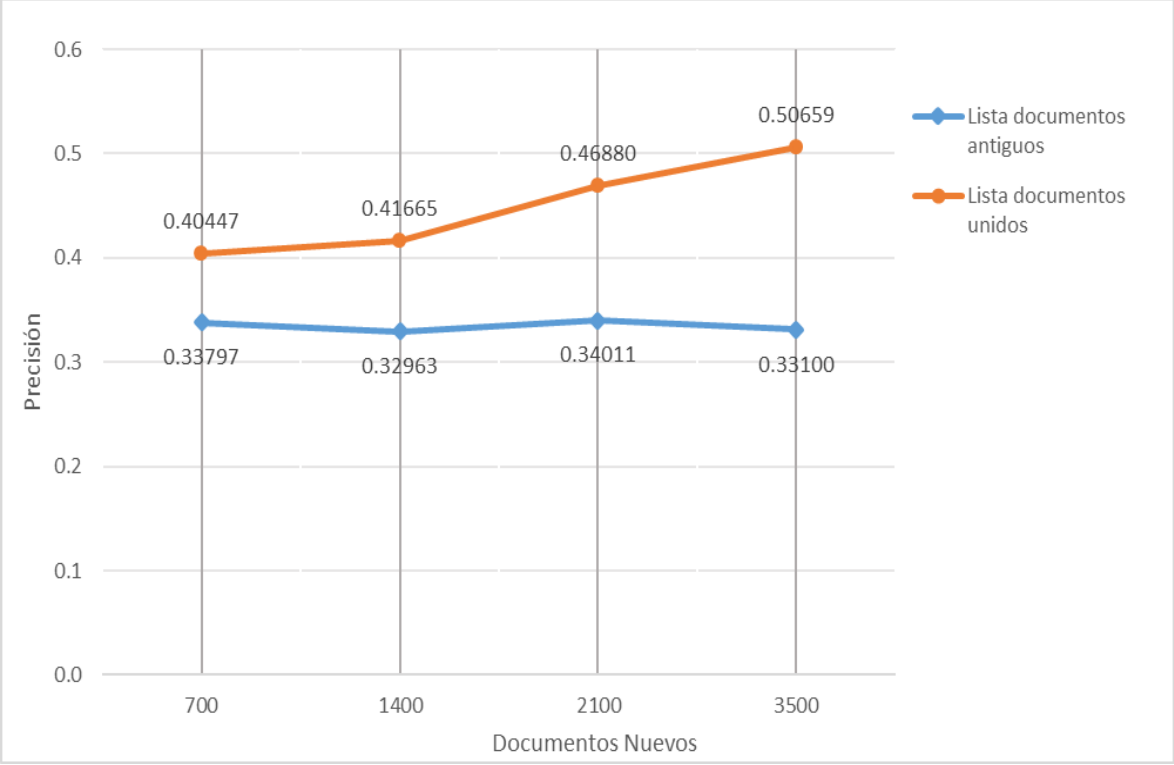

Figura Nº14: Gráfico del séptimo experimento de precisión de la primera etapa.

Experimento Nº8: En el octavo experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 2100. En la Tabla Nº12 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº15.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
3500	700	0,3324918	0,3561018
3500	1400	0,3349694	0,3741413
3500	2100	0,3371241	0,3863214
3500	3500	0,3343259	0,4133978

Tabla Nº12: Resultados del octavo experimento de precisión de la primera etapa.

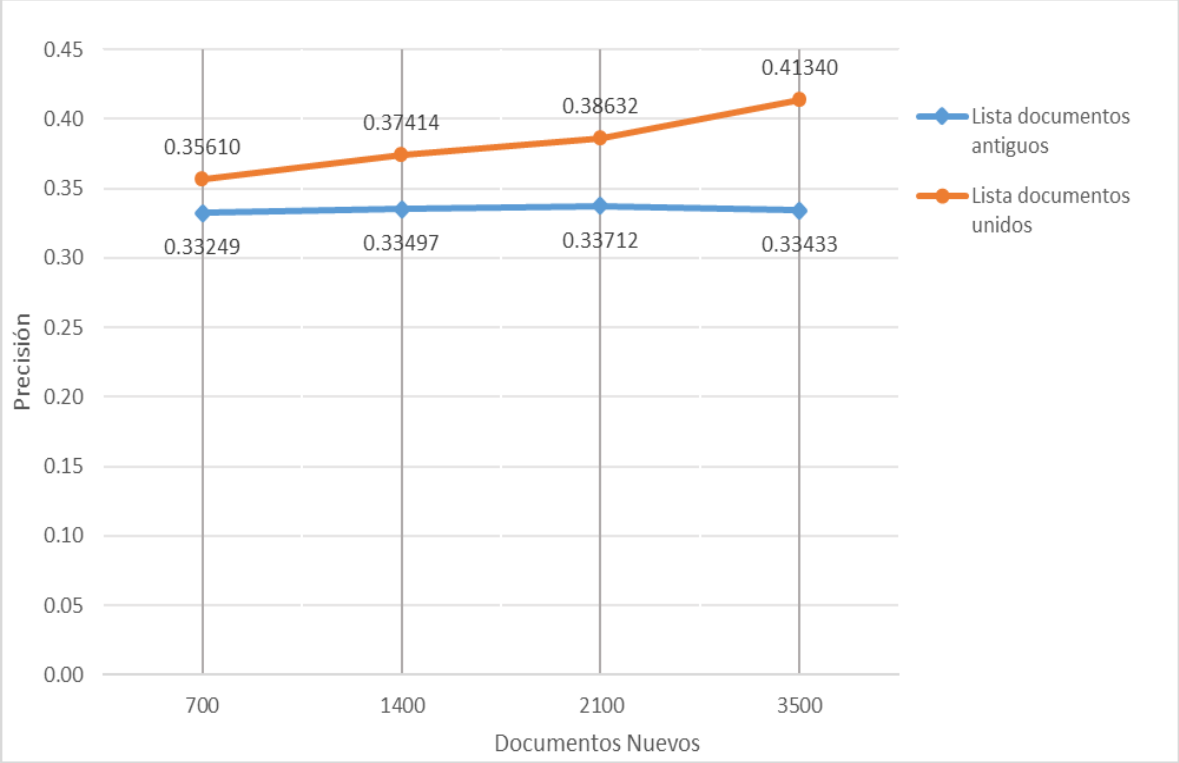

Figura Nº15: Grafico del octavo experimento de precisión de la primera etapa.

5.4.2 Experimentos de precisión segunda etapa

Experimento Nº1: En el primer experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº13 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº16.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
700	700	0.3336572	0.5753214
700	1400	0.3345751	0.6811099
700	2100	0.3353508	0.7330230
700	3500	0.3430986	0.7903188

Tabla Nº13: Resultados del primer experimento de precisión de la segunda etapa.

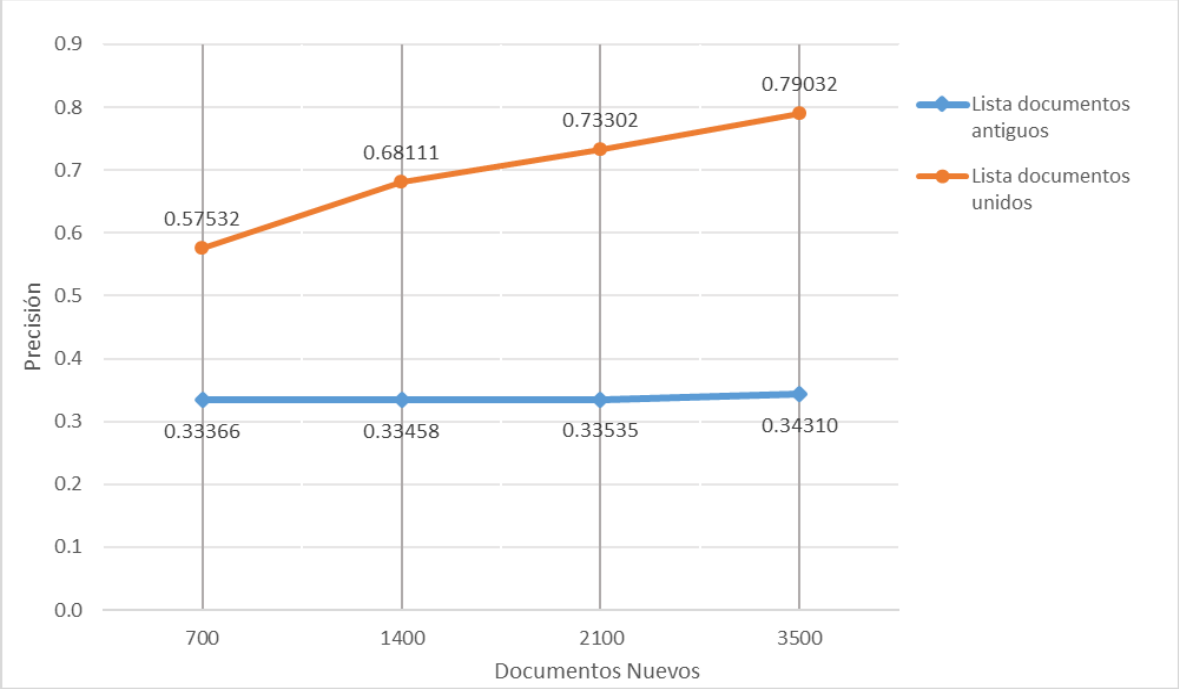

Figura Nº16: Gráfico del primer experimento de precisión de la segunda etapa.

Experimento Nº2: En el segundo experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº14 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº17.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
1400	700	0.3325603	0.4531100
1400	1400	0.3329403	0.5296754
1400	2100	0.3368594	0.5853388
1400	3500	0.3309420	0.6216598

Tabla Nº14: Resultados del segundo experimento de precisión de la segunda etapa.

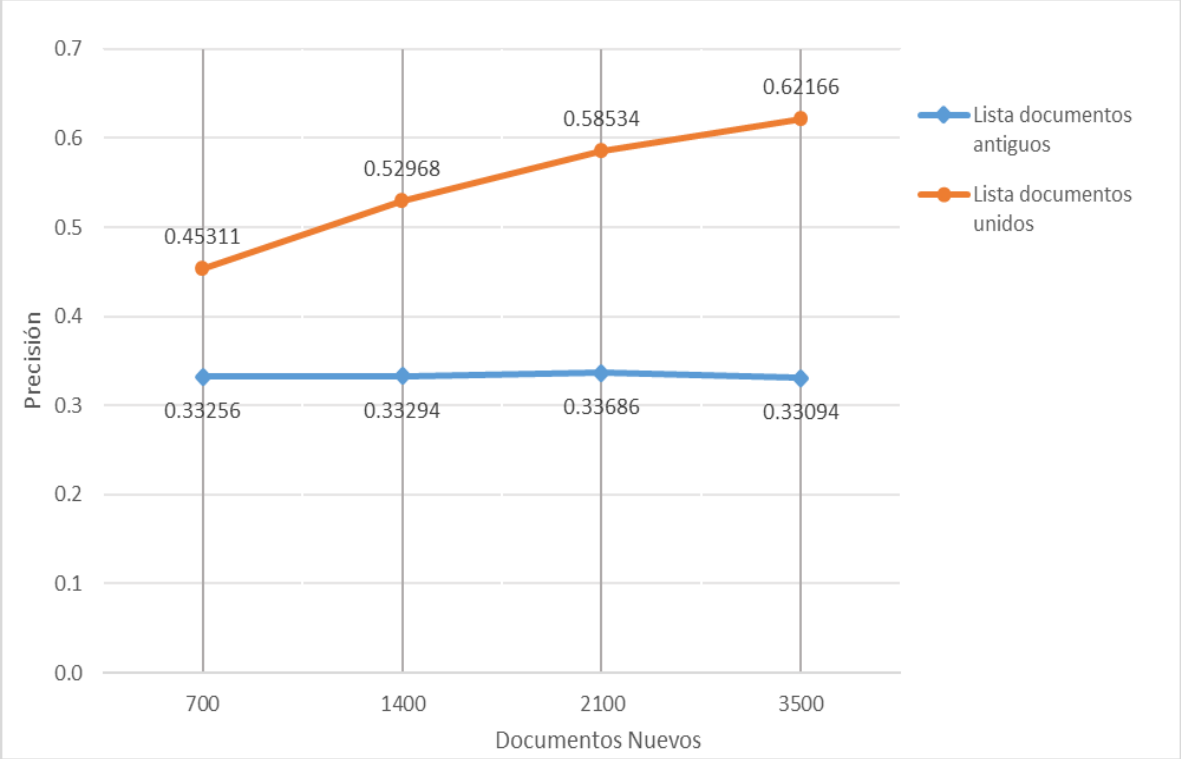

Figura Nº17: Gráfico del segundo experimento de precisión de la segunda etapa.

Experimento Nº3: En el tercer experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº15 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº18.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
2100	700	0.3321230	0.3951036
2100	1400	0.3293049	0.4454000
2100	2100	0.3316850	0.4768160
2100	3500	0.3341962	0.5211741

Tabla Nº15: Resultados del tercer experimento de precisión de la segunda etapa.

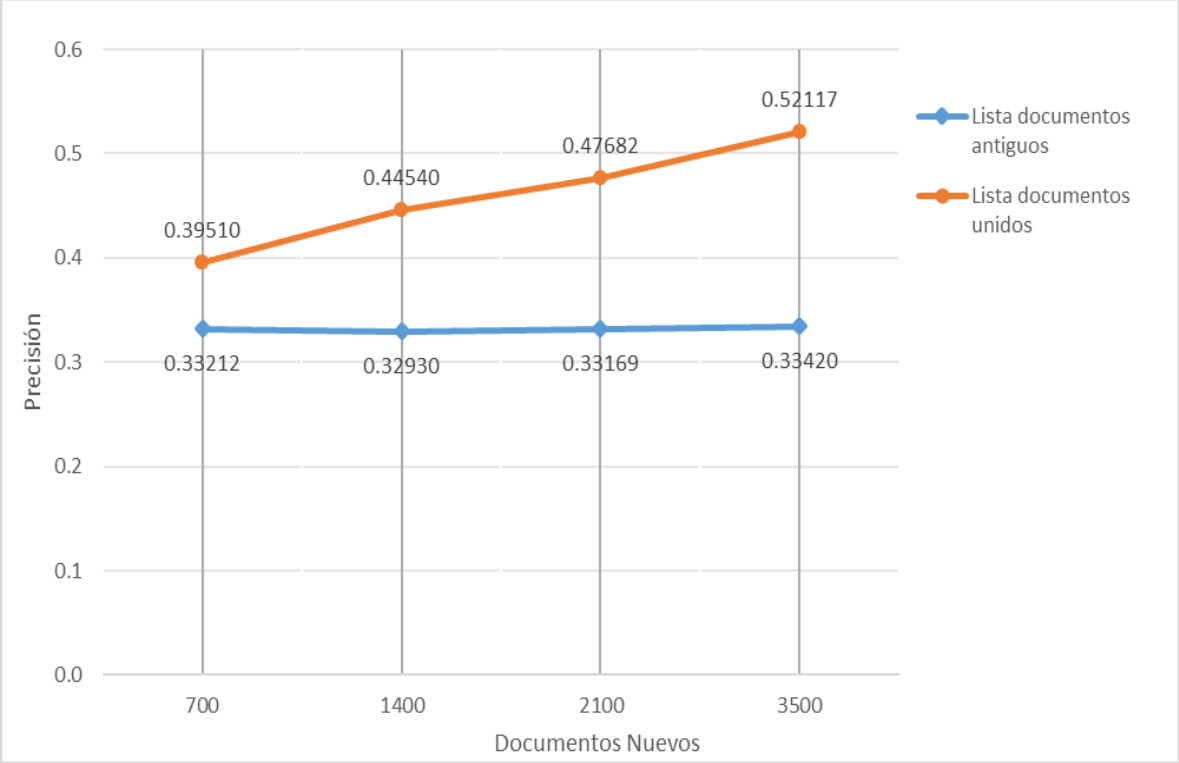

Figura Nº18: Gráfico del tercer experimento de precisión de la segunda etapa.

Experimento Nº4: En el cuarto experimento se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos nuevos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº16 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº19.

Documentos		Precisión	
Antiguos	Nuevos	Promedio Antiguo	Promedio Nueva
3500	700	0.3350212	0.3602268
3500	1400	0.3387138	0.3704918
3500	2100	0.3335869	0.3779018
3500	3500	0.3308057	0.3853206

Tabla Nº16: Resultados del cuarto experimento de precisión de la segunda etapa.

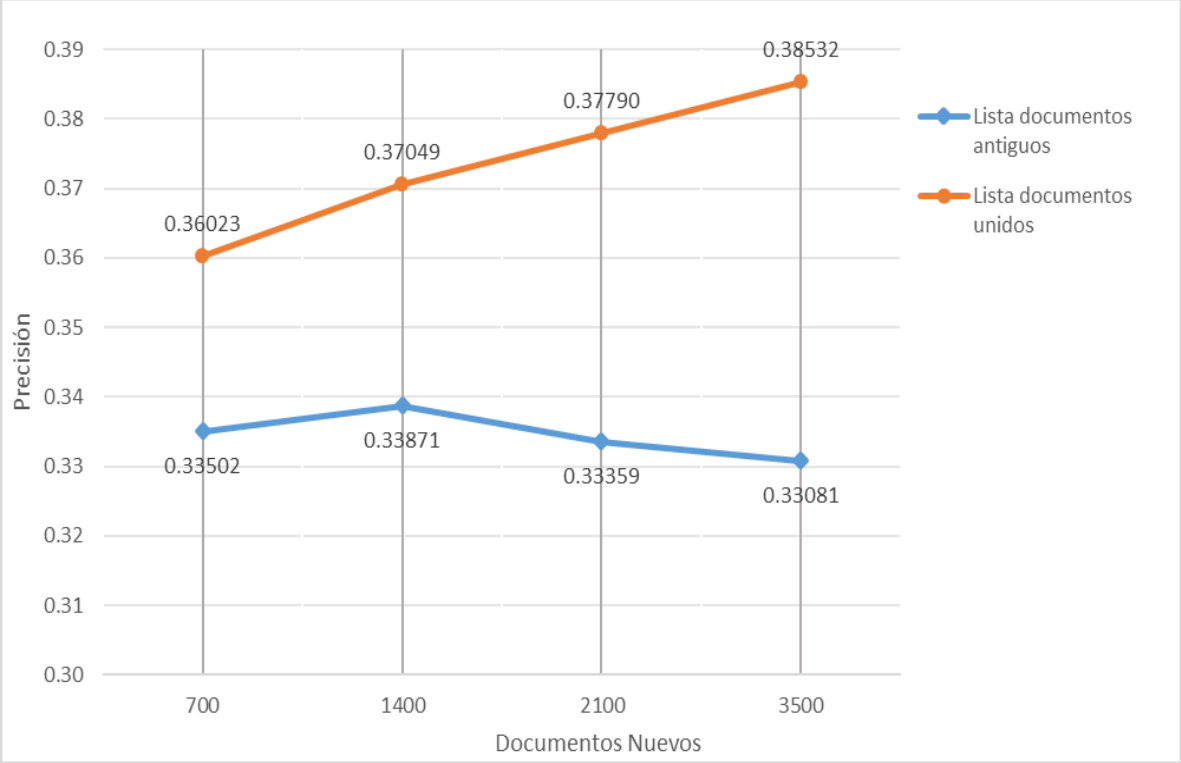

Figura Nº19: Gráfico del cuarto experimento de precisión de la segunda etapa.

5.5 Experimentos de Clustering

De acuerdo a los resultados obtenidos de los experimentos de la sección 5.4, se determina que es conveniente rehacer el clúster, debido a que los valores de precisión siempre mejoraron a medida que se iban incorporando nuevos documentos. Si bien los experimentos realizados demostraron que una mejora de precisión se producía con valores mayores o igual a 100 documentos nuevos, para los experimentos se fijó como mínima cantidad de documentos antiguos 700, puesto que era más notorio el aumento en la precisión.

Como se sabe que al agregar documentos nuevos se genera un aumento en la precisión, por lo que resulta conveniente, en estos casos, rehacer el clúster. Como se mencionó en la sección anterior, existen varios algoritmos de clustering, por lo cual, en esta segunda fase, se prueban los distintos algoritmos con el objetivo de determinar cuál brinda mejores resultados al rehacer el cluster. El atributo a medir en estos experimentos es la cantidad de documentos relevantes visitados al momento de recorrer los nuevos clústeres.

Previo al proceso de rehacer el cluster, es necesario conocer cuáles son los documentos relevantes del conjunto de documentos antiguos como también de la lista de documentos unidos, dado que estos son necesarios para evaluar cuantos documentos relevantes son visitados al momento de recorrer los clústeres. Los documentos relevantes tanto del conjunto de documentos antiguos y lista de documentos unidos será obtenida de la primera parte de los experimentos.

Para el funcionamiento de los algoritmos, además de contar con los documentos relevantes se debe construir una matriz de similaridad entre documentos, una para el conjunto de documentos antiguos y otra para la lista de documentos unidos, las matrices generadas serán utilizadas como entrada para los algoritmos de clustering implementados. Al momento de construir la matriz de similaridad entre documentos se utiliza la medida de distancia del coseno. Un ejemplo de cómo es construida una matriz de similaridad se encuentra en la sección 2.6 del capítulo 2.

Al igual que los experimentos de precisión, los experimentos de clustering se realizarán en dos etapas, en la primera etapa se mostrarán los resultados de clustering asociados al algoritmo probabilístico. En la segunda etapa se encontrarán los resultados de clustering luego de aplicar el algoritmo de Monte Carlo al algoritmo probabilístico.

Para cada algoritmo de clustering se encuentra una tabla donde se muestra los siguientes valores:

Doc. Antiguos: Corresponde a la cantidad de documentos relevantes visitados al recorrer el cluster generado con los documentos antiguos.

Doc. Unidos: Corresponde a la cantidad de documentos relevantes visitados al recorrer el cluster generado con la lista de documentos unidos (documentos antiguos y documentos nuevos).

Además, para cada experimento se encuentra un gráfico, donde en las ordenadas se representa la cantidad de documentos relevantes en el cluster y en las abscisas el aumento de documentos.

5.5.1 Experimentos de clustering primera etapa

5.5.1.1 Experimentos Single link

Experimento 1: En el primer experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº17 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº20.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
700	700	230	241
700	1400	223	224
700	2100	207	224
700	3500	216	200

Tabla Nº17: Resultados del primer experimento de Single link de la primera etapa.

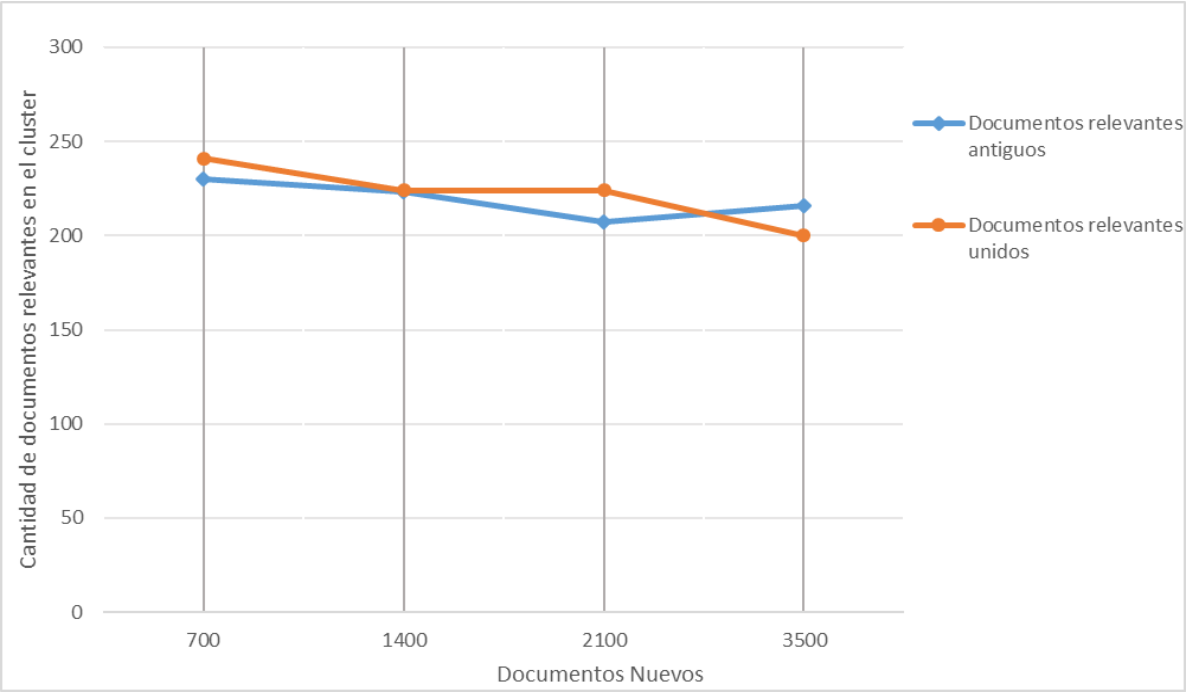

Figura Nº20: Gráfico del primer experimento de Single link de la primera etapa.

Experimento 2: En el segundo experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº18 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº21.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
1400	700	225	222
1400	1400	254	232
1400	2100	241	232
1400	3500	228	243

Tabla Nº18: Resultados del segundo experimento de Single link de la primera etapa.

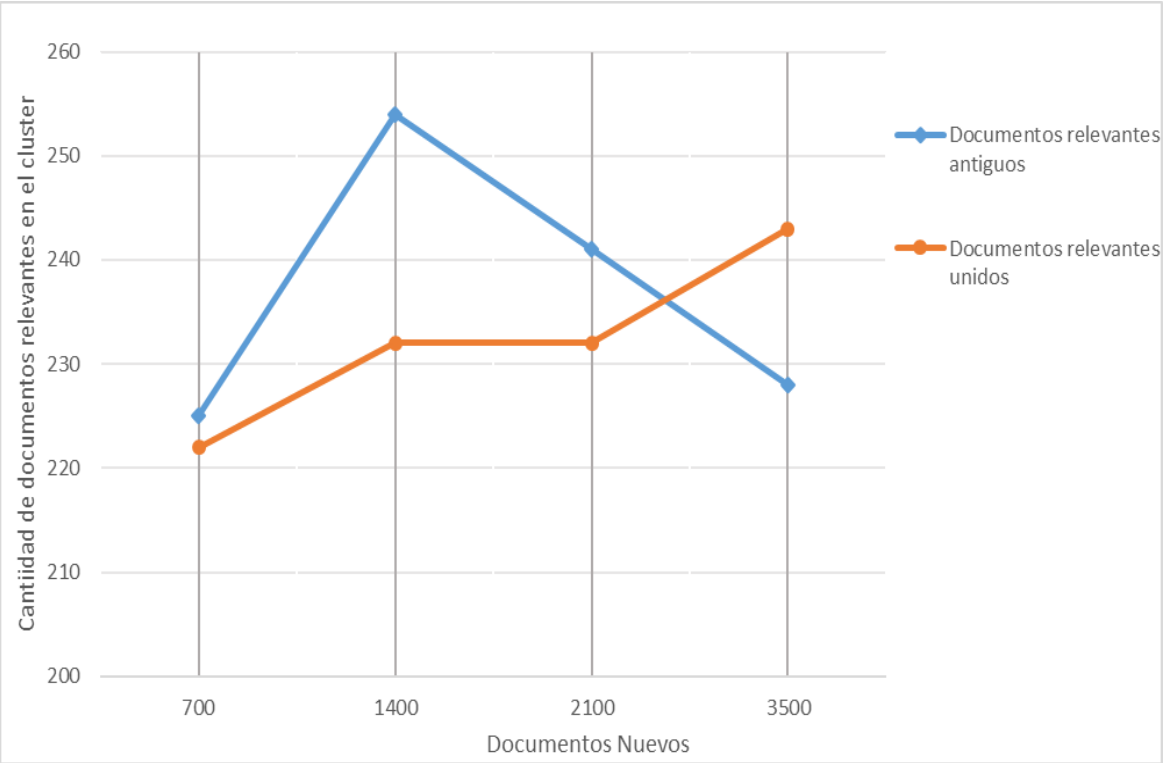

Figura Nº21: Gráfico del segundo experimento de Single link de la primera etapa.

Experimento 3: En el tercer experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº19 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº22.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
2100	700	235	245
2100	1400	224	201
2100	2100	203	231
2100	3500	226	223

Tabla Nº19: Resultados del tercer experimento de Single link de la primera etapa.

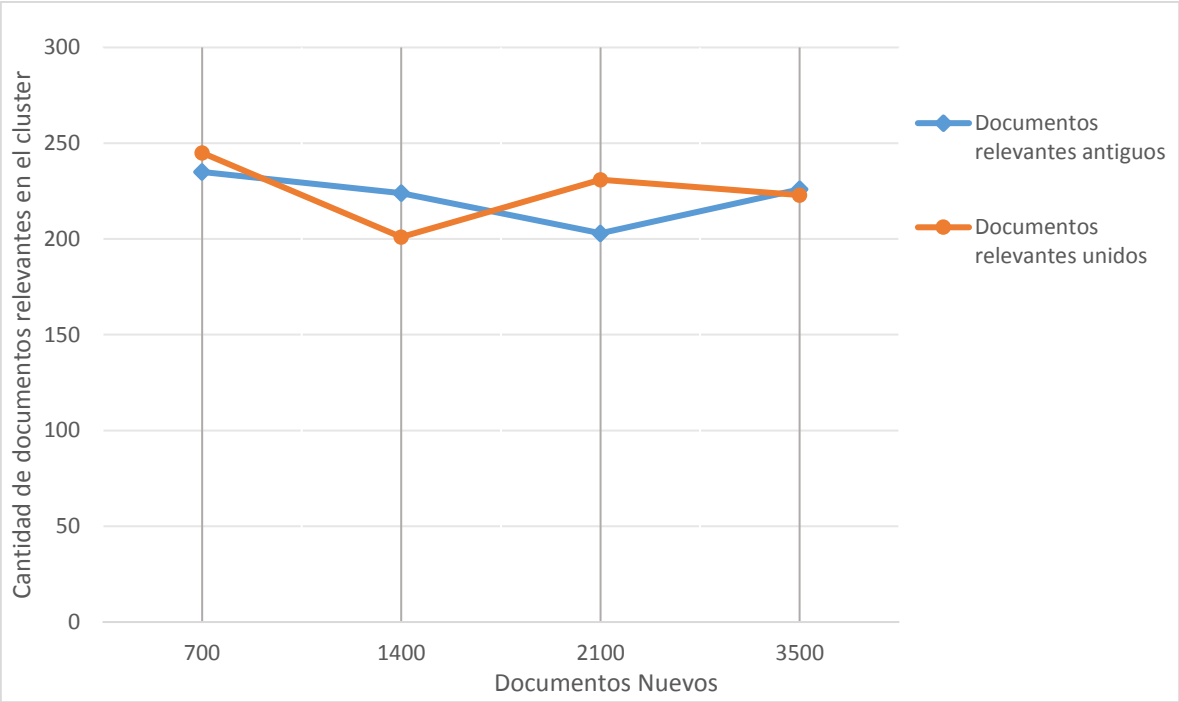

Figura Nº22: Gráfico del tercer experimento de Single link de la primera etapa.

Experimento 4: En el cuarto experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº20 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº23.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
3500	700	206	196
3500	1400	226	207
3500	2100	234	236
3500	3500	236	233

Tabla Nº20: Resultados del cuarto experimento de Single link de la primera etapa.

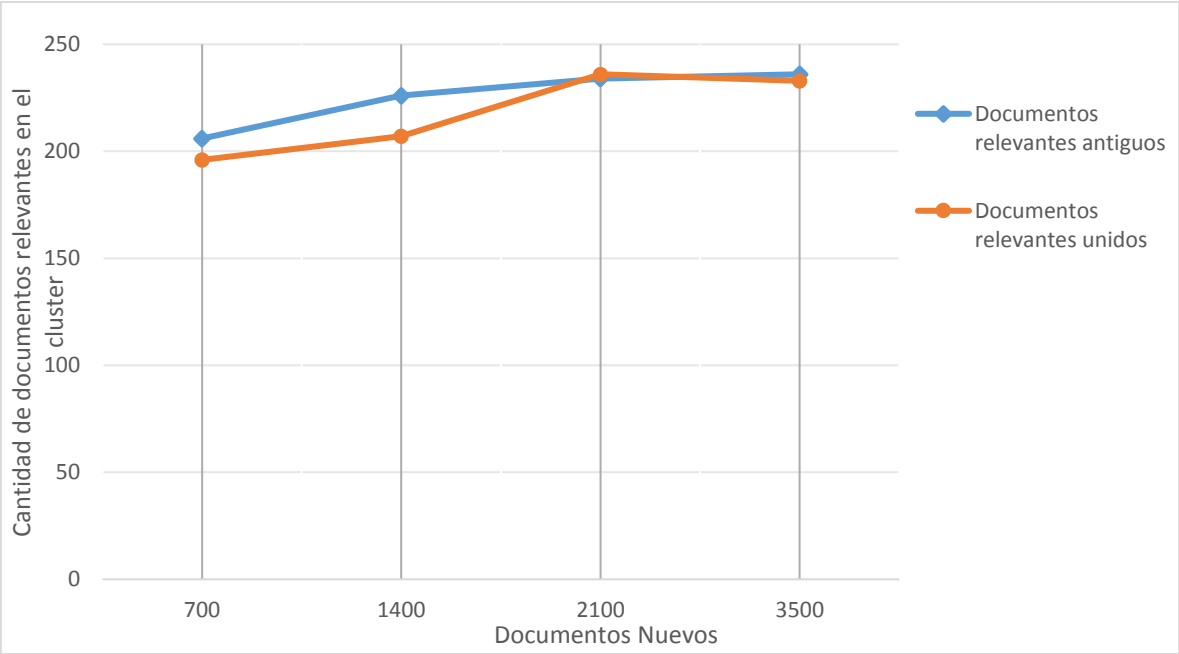

Figura Nº23: Gráfico del cuarto experimento de Single link de la primera etapa.

5.5.1.2 Experimentos Complete link

Experimento 1: En el primer experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº21 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº24.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
700	700	231	258
700	1400	245	236
700	2100	216	247
700	3500	229	225

Tabla Nº21: Resultados del primer experimento de Complete link de la primera etapa.

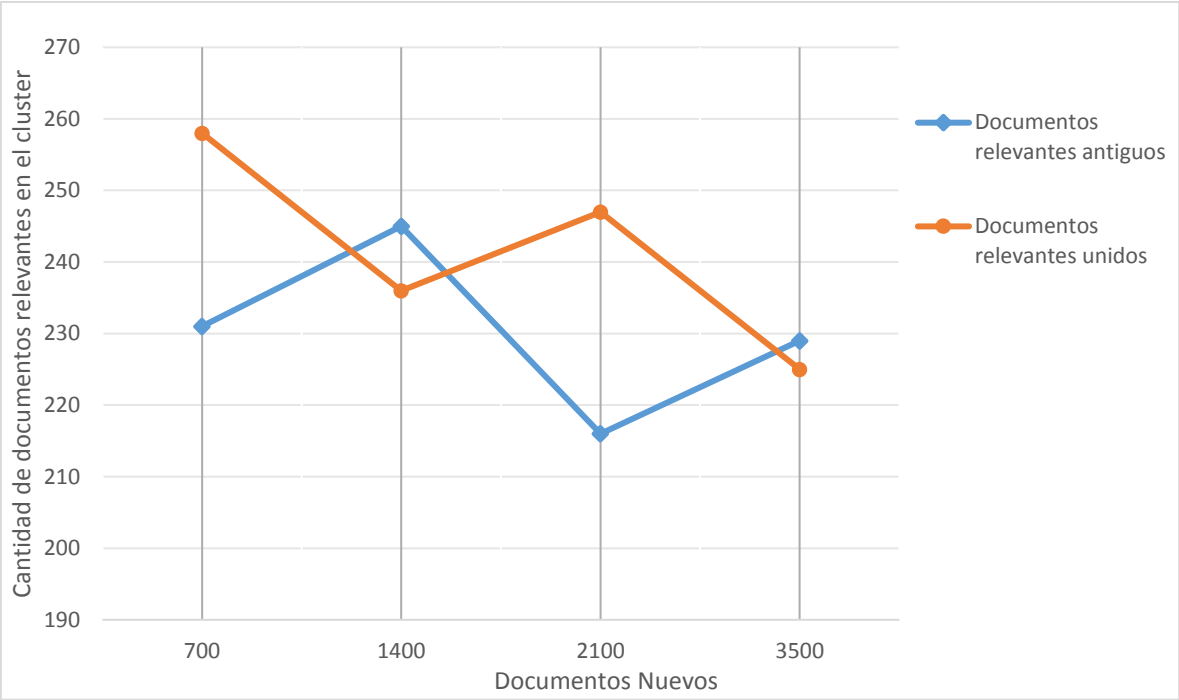

Figura Nº24: Gráfico del primer experimento de Complete link de la primera etapa.

Experimento 2: En el segundo experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº22 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº25.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
1400	700	236	245
1400	1400	238	251
1400	2100	241	251
1400	3500	236	258

Tabla Nº22: Resultados del segundo experimento de Complete link de la primera etapa.

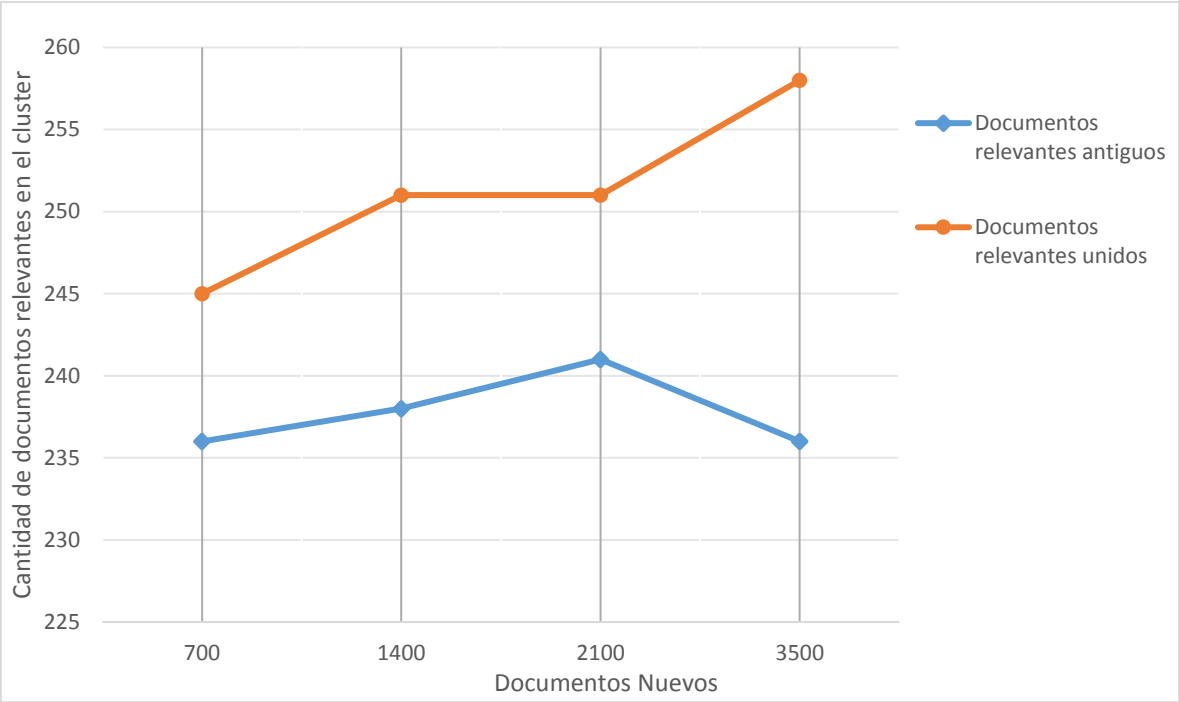

Figura Nº25: Gráfico del segundo experimento de Complete link de la primera etapa.

Experimento 3: En el tercer experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº23 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº26.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
2100	700	250	255
2100	1400	220	221
2100	2100	229	257
2100	3500	240	243

Tabla Nº23: Resultados del tercer experimento de Complete link de la primera etapa.

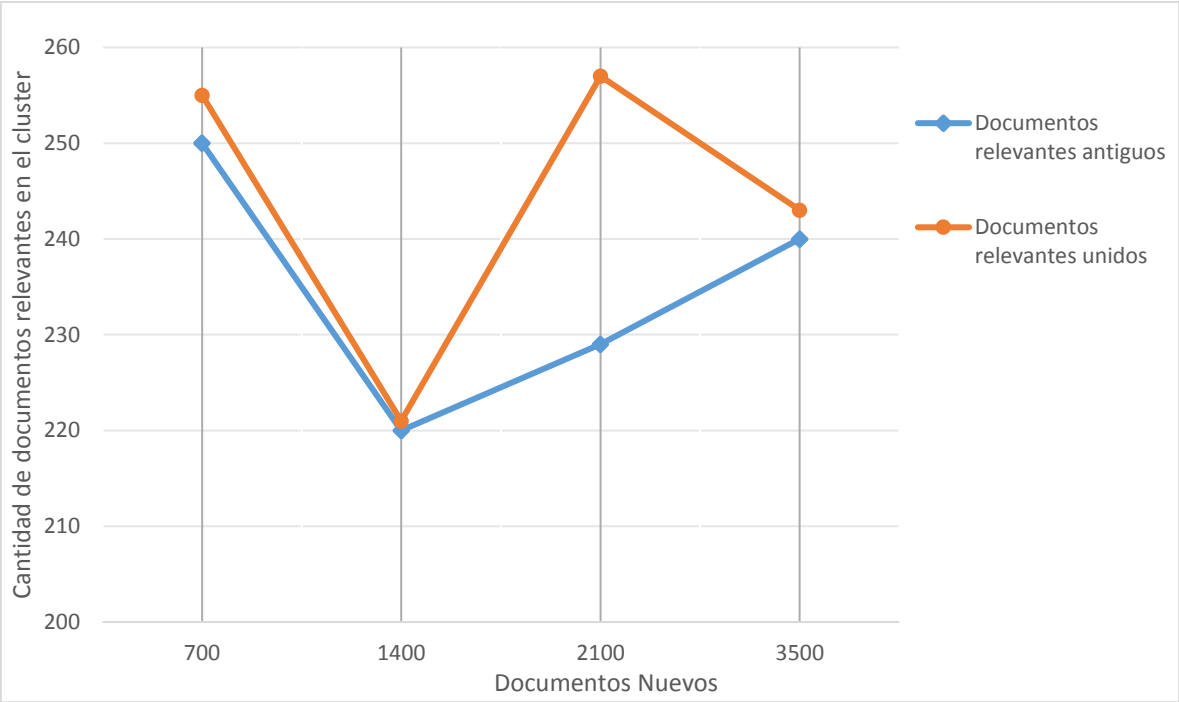

Figura Nº26: Gráfico del tercer experimento de Complete link de la primera etapa.

Experimento 4: En el cuarto experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº24 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº27.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
3500	700	228	235
3500	1400	237	245
3500	2100	232	277
3500	3500	232	223

Tabla Nº24: Resultados del cuarto experimento de Complete link de la primera etapa.

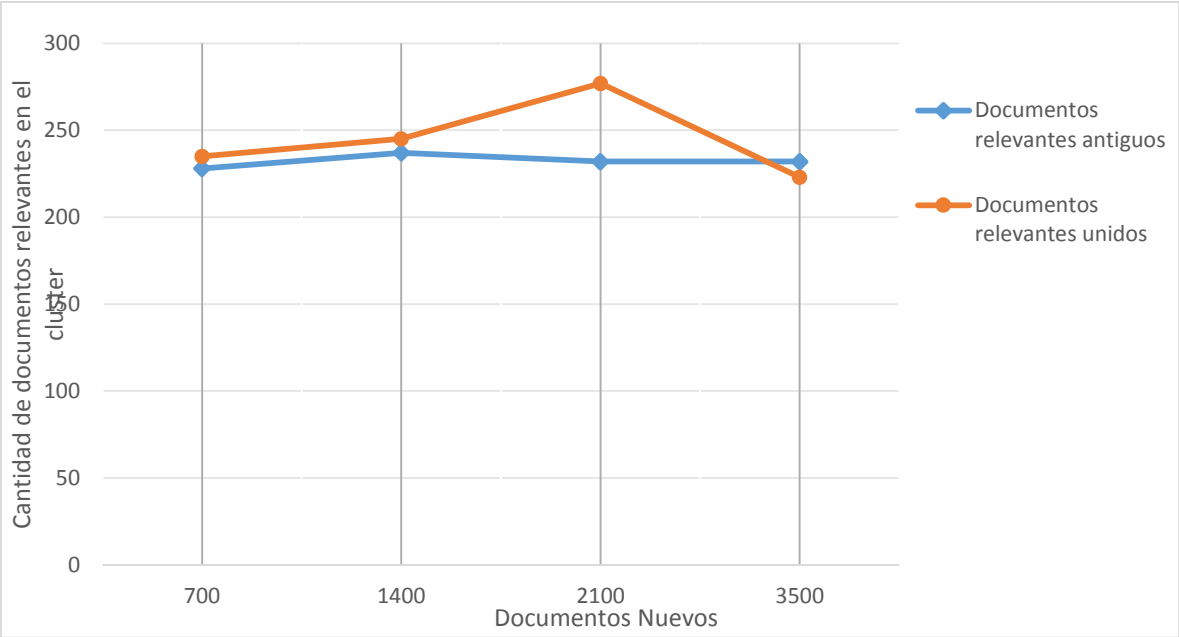

Figura Nº27: Gráfico del cuarto experimento de Complete link de la primera etapa.

5.5.1.3 Experimentos Average link

Experimento 1: En el primer experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº25 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº28.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
700	700	2140	2372
700	1400	1860	2182
700	2100	2059	2478
700	3500	2121	2360

Tabla Nº25: Resultados del primer experimento de Average link de la primera etapa.

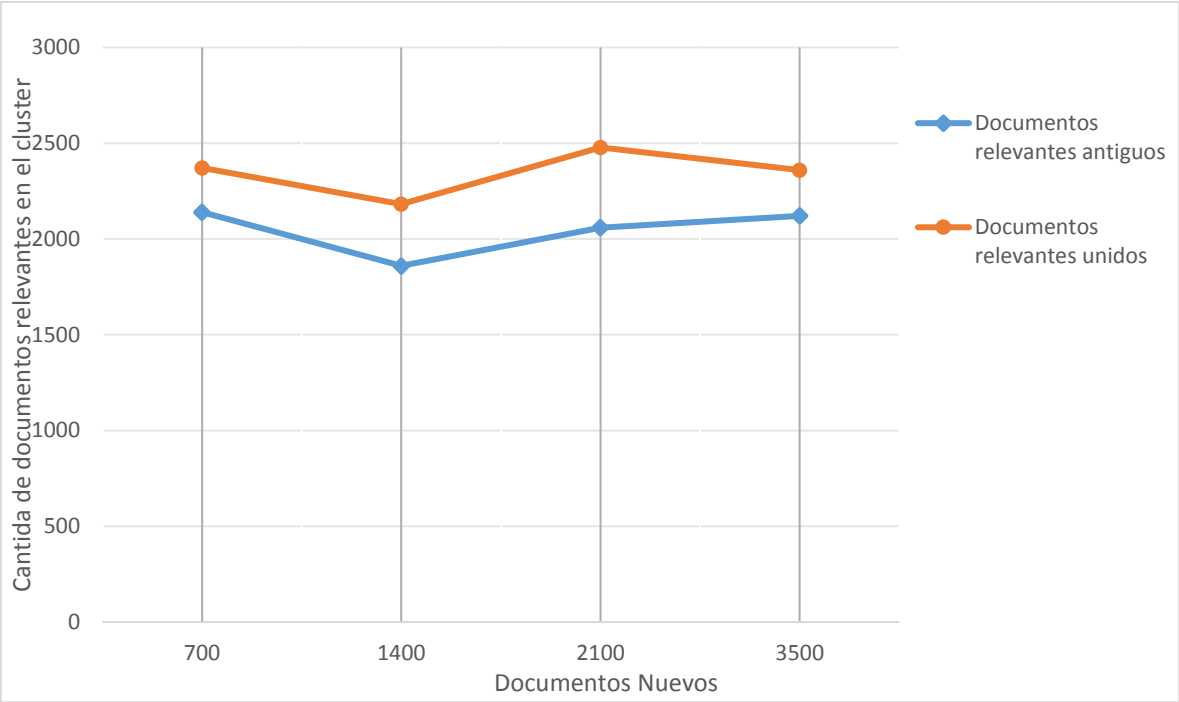

Figura Nº28: Gráfico del primer experimento de Average link de la primera etapa.

Experimento 2: En el segundo experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº26 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº29.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
1400	700	2309	2873
1400	1400	2241	2788
1400	2100	2042	2875
1400	3500	2403	3052

Tabla Nº26: Resultados del segundo experimento de Average link de la primera etapa.

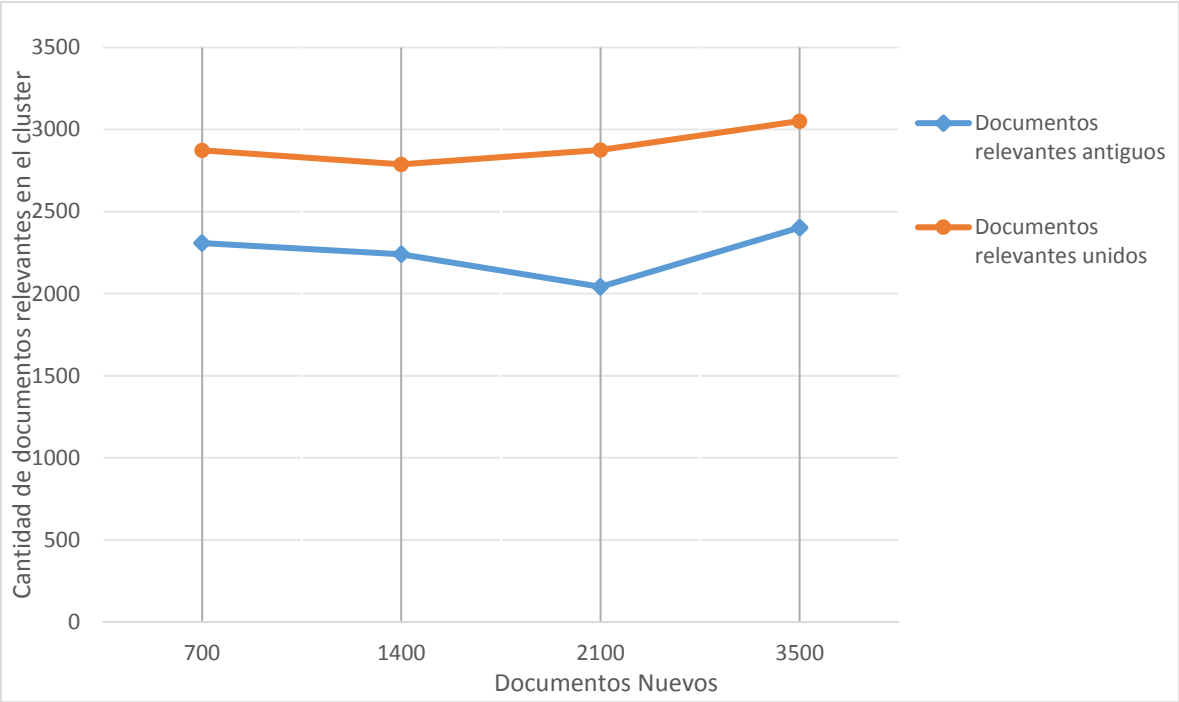

Figura Nº29: Gráfico del segundo experimento de Average link de la primera etapa.

Experimento 3: En el tercer experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº27 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº30.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
2100	700	2398	2767
2100	1400	2484	2731
2100	2100	2435	3173
2100	3500	2613	3177

Tabla Nº27: Resultados del tercer experimento de Average link de la primera etapa.

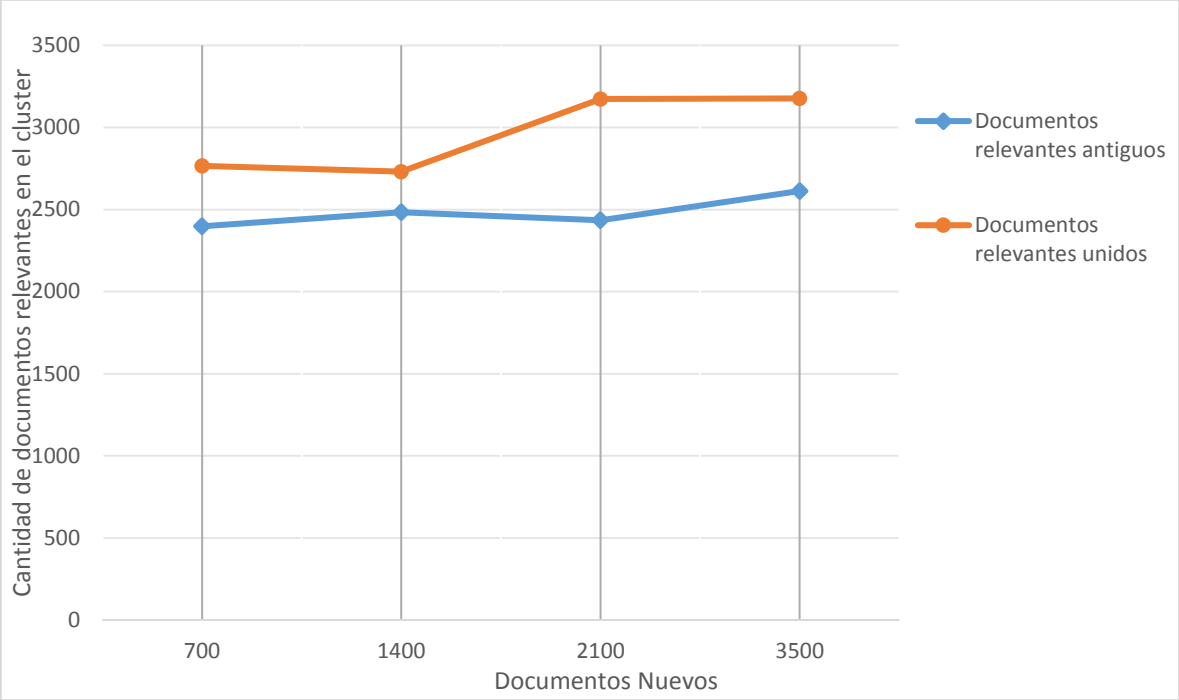

Figura Nº30: Gráfico del tercer experimento de Average link de la primera etapa.

Experimento 4: En el cuarto experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº28 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº31.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
3500	700	2455	3035
3500	1400	2697	3050
3500	2100	2752	2791
3500	3500	2575	3389

Tabla Nº28: Resultados del cuarto experimento de Average link de la primera etapa.

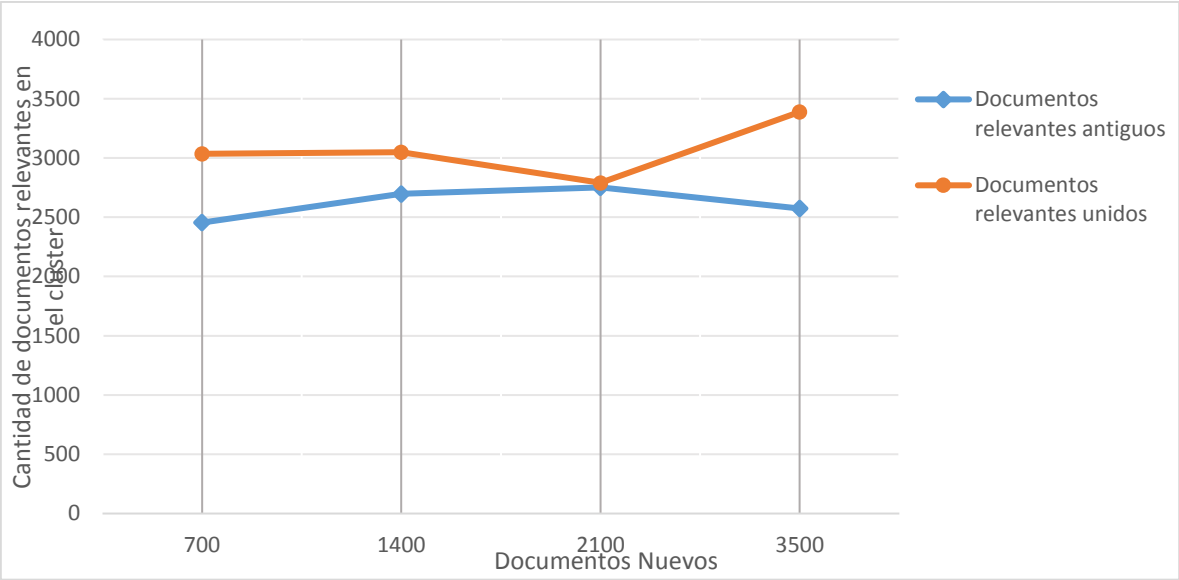

Figura Nº31: Gráfico del cuarto experimento de Average link de la primera etapa.

5.5.2 Experimentos de clustering segunda etapa

5.5.2.1 Experimentos de Single Link

Experimento 1: En el primer experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº29 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº32.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
700	700	218	242
700	1400	223	231
700	2100	196	224
700	3500	240	246

Tabla Nº29: Resultados del primer experimento de Single link de la segunda etapa.

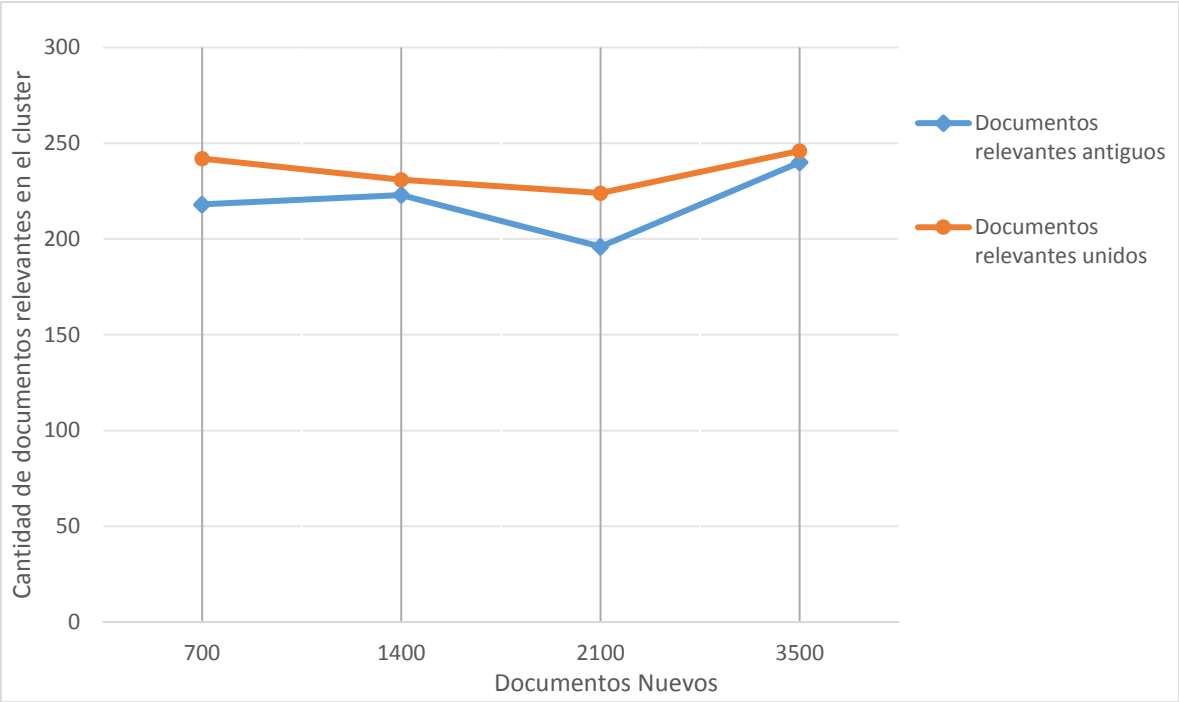

Figura Nº32: Gráfico del primer experimento de Single link de la segunda etapa.

Experimento 2: En el segundo experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº30 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº33.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
1400	700	227	237
1400	1400	224	214
1400	2100	239	233
1400	3500	234	236

Tabla Nº30: Resultados del segundo experimento de Single link de la segunda etapa.

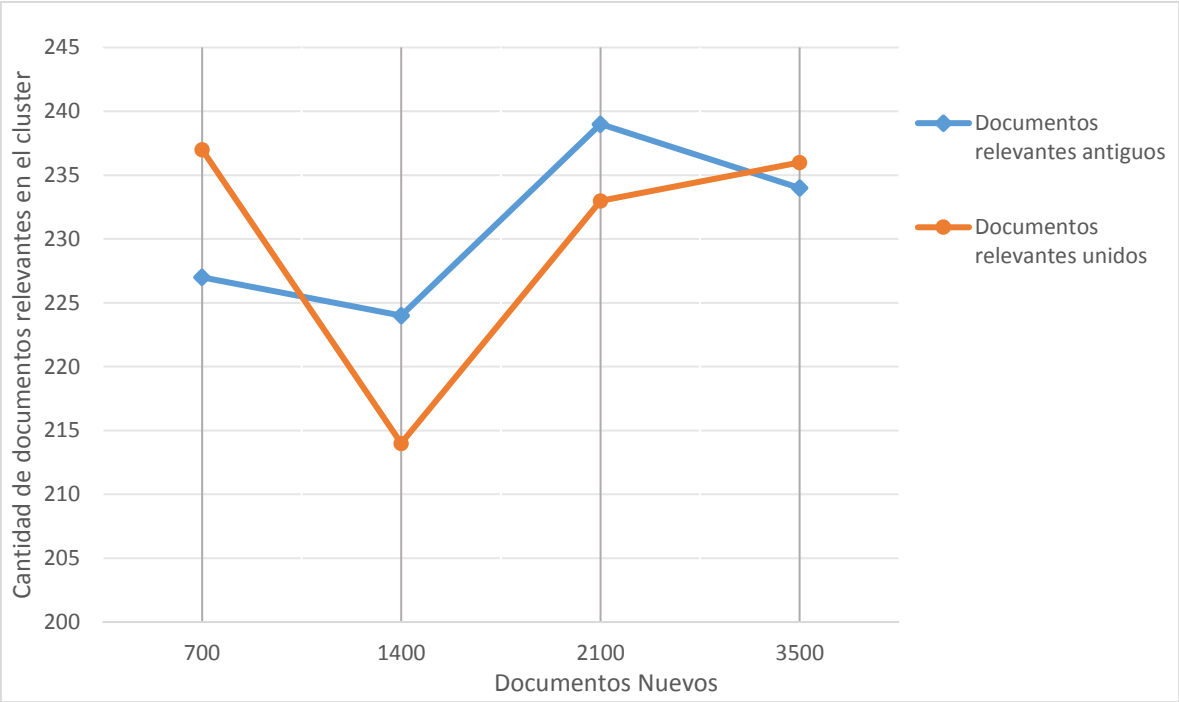

Figura Nº33: Gráfico del segundo experimento de Single link de la segunda etapa.

Experimento 3: En el tercer experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº31 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº34.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
2100	700	242	246
2100	1400	237	243
2100	2100	204	205
2100	3500	220	206

Tabla Nº31: Resultados del tercer experimento de Single link de la segunda etapa.

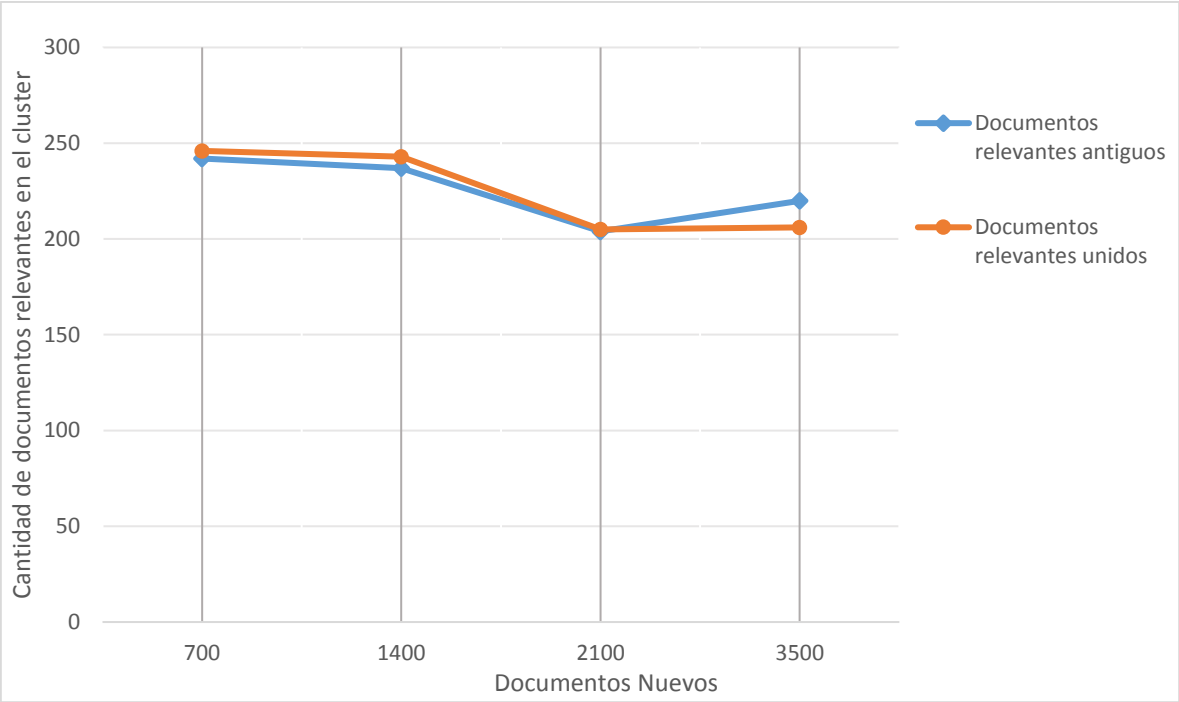

Figura Nº34: Gráfico del tercer experimento de Single link de la segunda etapa.

Experimento 4: En el cuarto experimento se utilizó el algoritmo jerárquico Single link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº32 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº35.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
3500	700	208	226
3500	1400	231	228
3500	2100	229	229
3500	3500	237	232

Tabla Nº32: Gráfico del cuarto experimento de Single link de la segunda etapa.

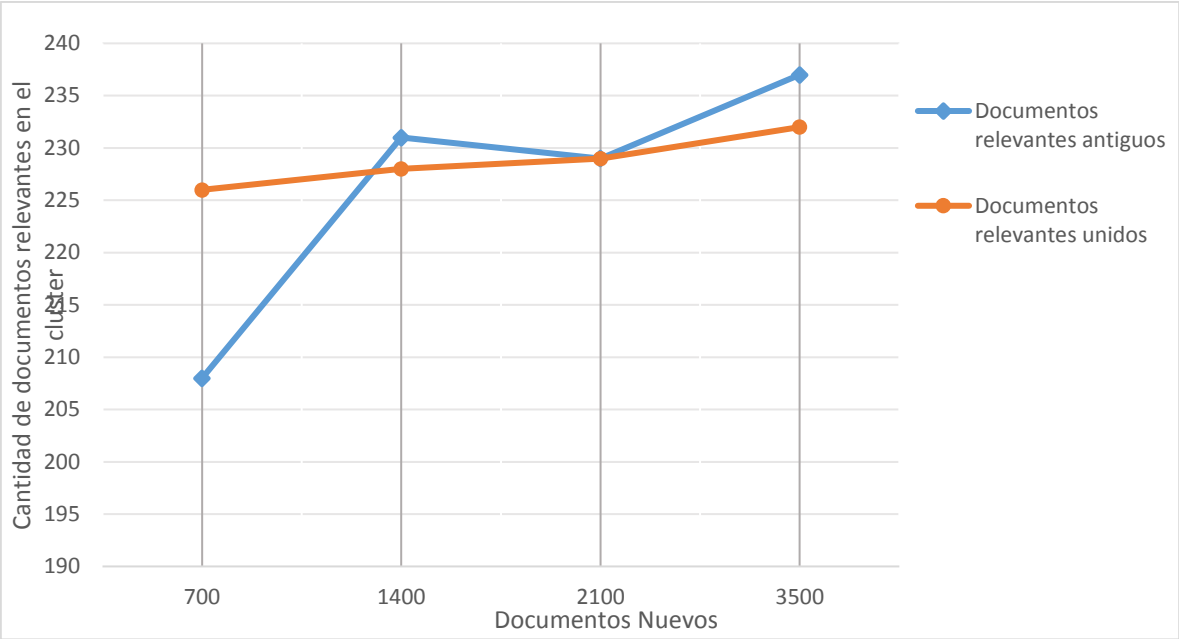

Figura Nº35: Gráfico del cuarto experimento de Single link de la segunda etapa.

5.5.2.2 Experimentos de Complete Link

Experimento 1: En el primer experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº33 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº36.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
700	700	222	256
700	1400	238	242
700	2100	217	243
700	3500	230	250

Tabla Nº33: Resultados del primer experimento de Complete link de la segunda etapa.

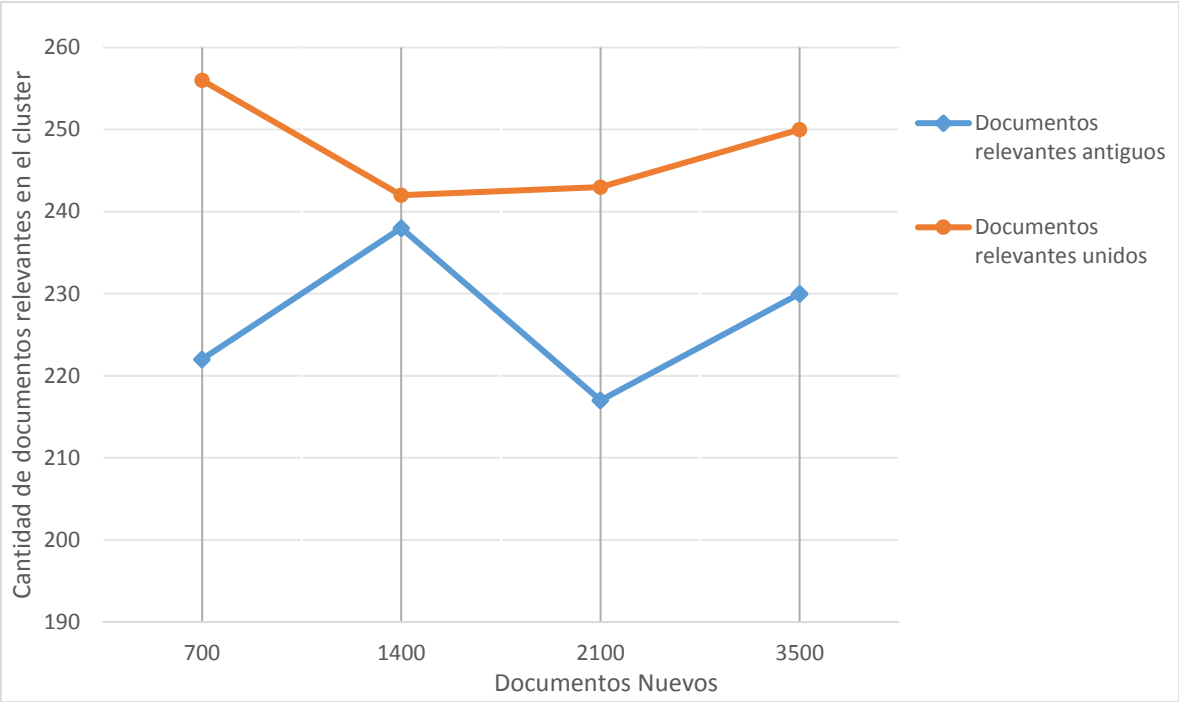

Figura Nº36: Gráfico del primer experimento de Complete link de la segunda etapa.

Experimento 2: En el segundo experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº34 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº37.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
1400	700	208	216
1400	1400	221	228
1400	2100	235	202
1400	3500	248	262

Tabla Nº34: Resultados del segundo experimento de Complete link de la segunda etapa.

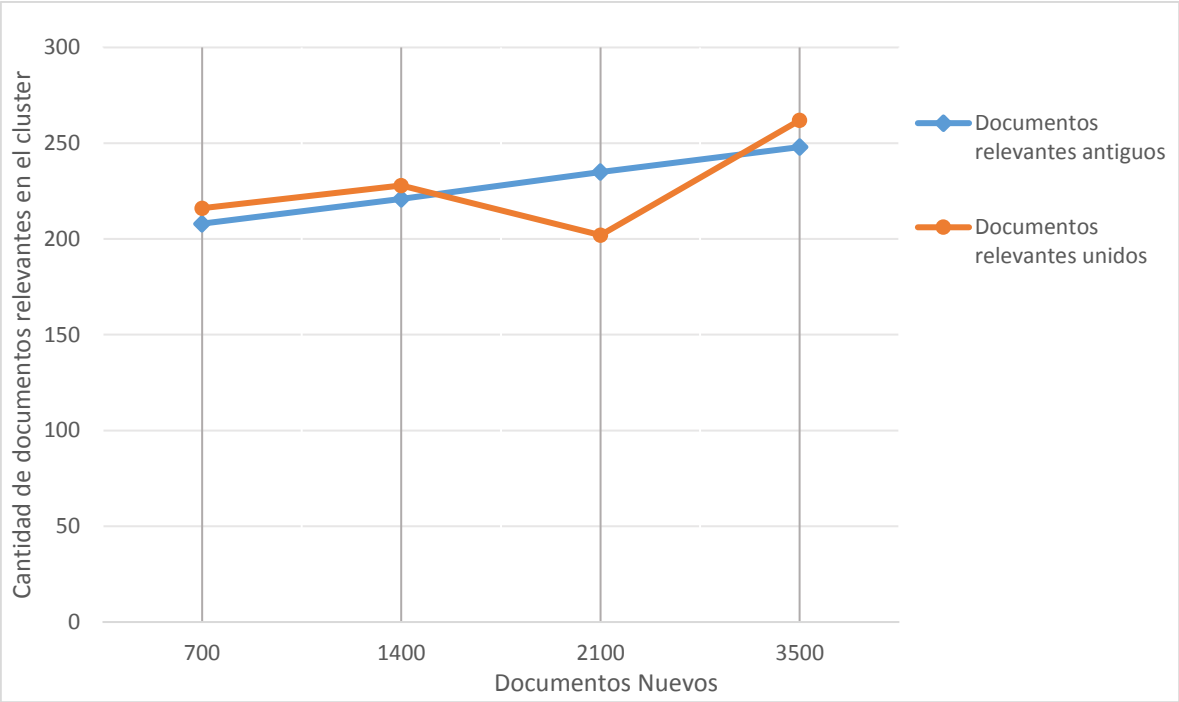

Figura Nº37: Gráfico del segundo experimento de Complete link de la segunda etapa.

Experimento 3: En el tercer experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº35 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº38.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
2100	700	240	248
2100	1400	265	236
2100	2100	229	223
2100	3500	244	215

Tabla Nº35: Resultados del tercer experimento de Complete link de la segunda etapa.

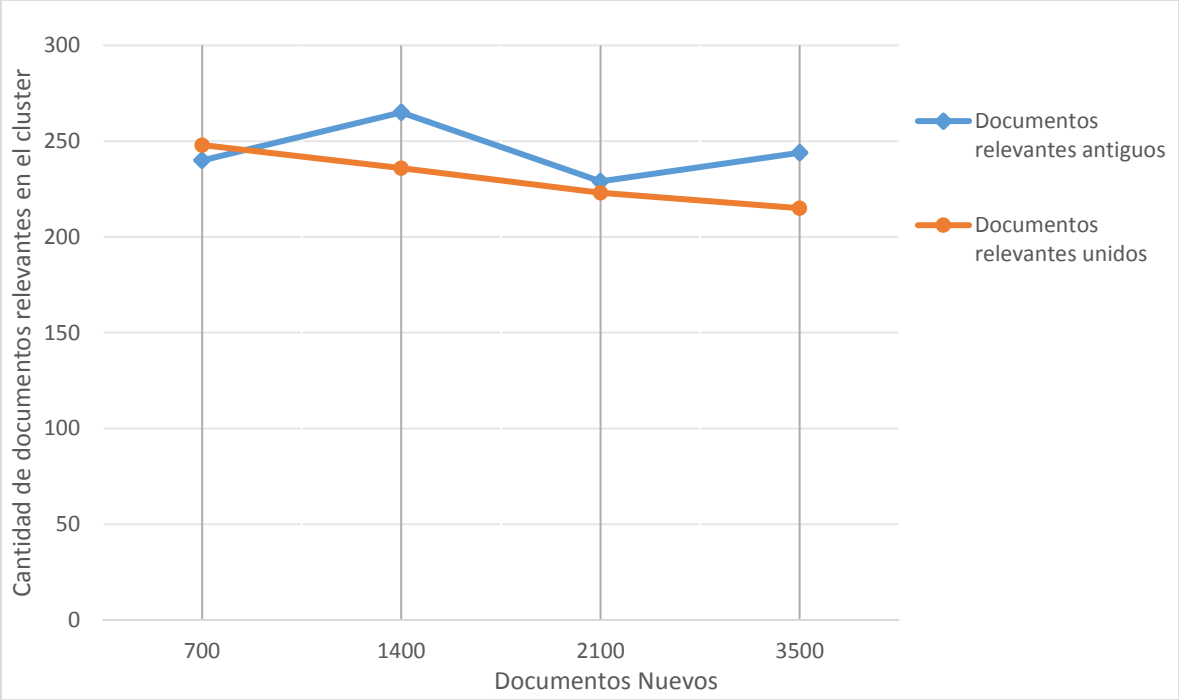

Figura Nº38: Gráfico del tercer experimento de Complete link de la segunda etapa.

Experimento 4: En el cuarto experimento se utilizó el algoritmo jerárquico Complete link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº36 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº39.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
3500	700	233	242
3500	1400	264	230
3500	2100	263	254
3500	3500	227	231

Tabla Nº36: Resultados del cuarto experimento de Complete link de la segunda etapa.

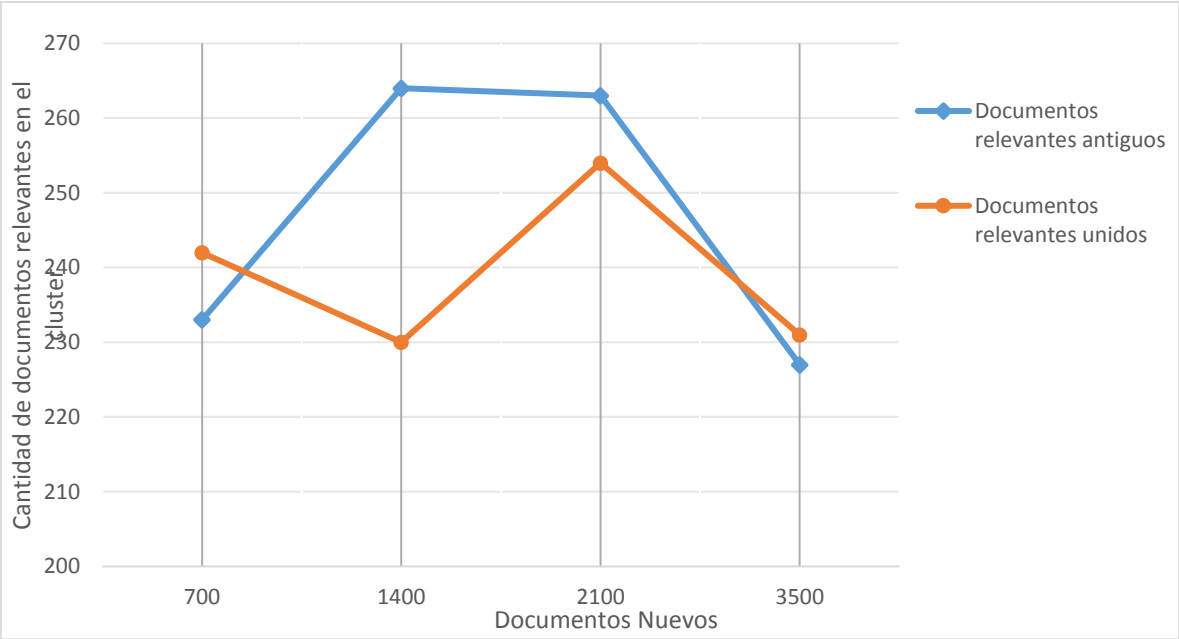

Figura Nº39: Gráfico del cuarto experimento de Complete link de la segunda etapa.

5.5.2.3 Experimentos de Average Link

Experimento 1: En el primer experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 700 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº37 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº40.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
700	700	1957	2412
700	1400	1936	2449
700	2100	2007	2489
700	3500	2072	2446

Tabla Nº37: Resultados del primer experimento de Average link de la segunda etapa.

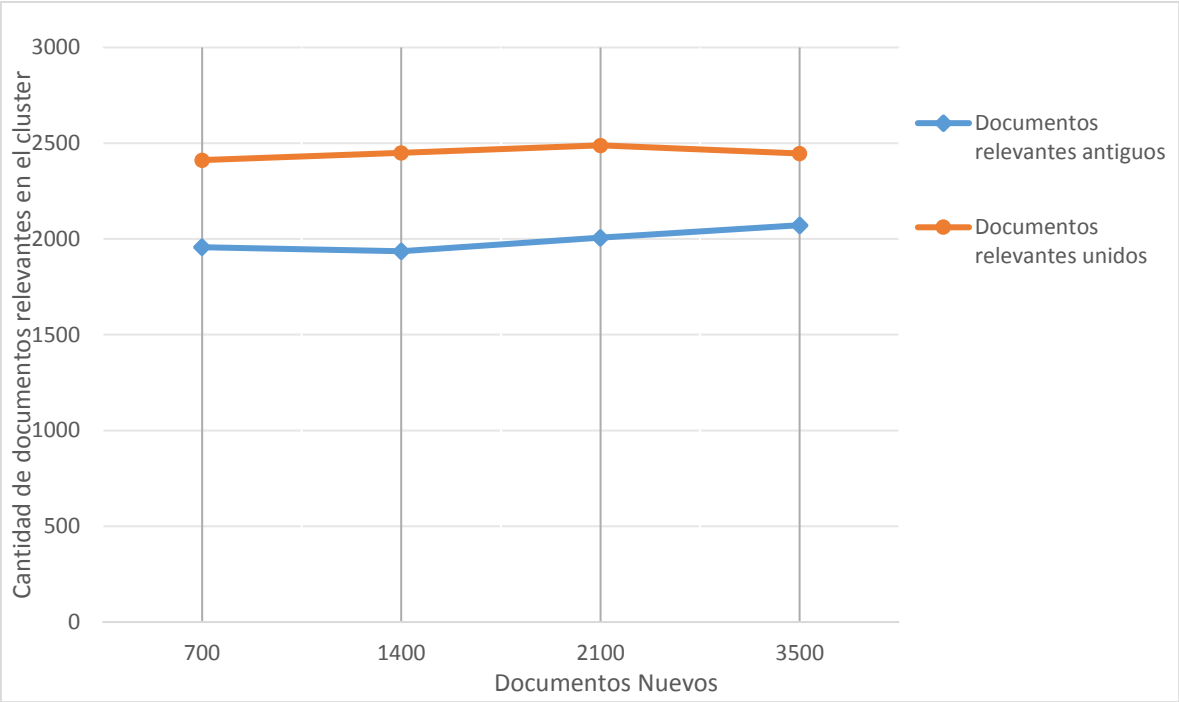

Figura Nº40: Gráfico del primer experimento de Average link de la segunda etapa.

Experimento 2: En el segundo experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 1400 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº38 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº41.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
1400	700	2464	2512
1400	1400	2209	2591
1400	2100	2462	2898
1400	3500	2311	2837

Tabla Nº38: Resultados del segundo experimento de Average link de la segunda etapa.

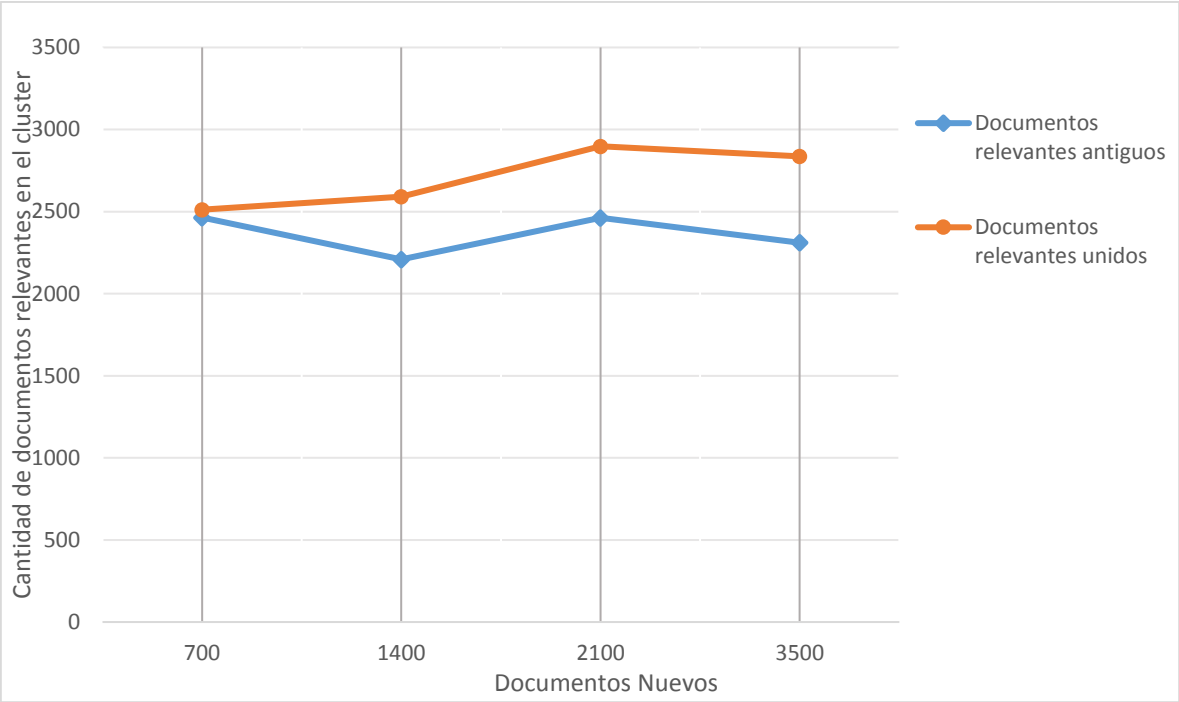

Figura Nº41: Gráfico del segundo experimento de Average link de la segunda etapa.

Experimento 3: En el tercer experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 2100 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº39 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº42.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
2100	700	2679	2864
2100	1400	2848	2755
2100	2100	2528	2919
2100	3500	2405	2906

Tabla Nº39: Resultados del tercer experimento de Average link de la segunda etapa.

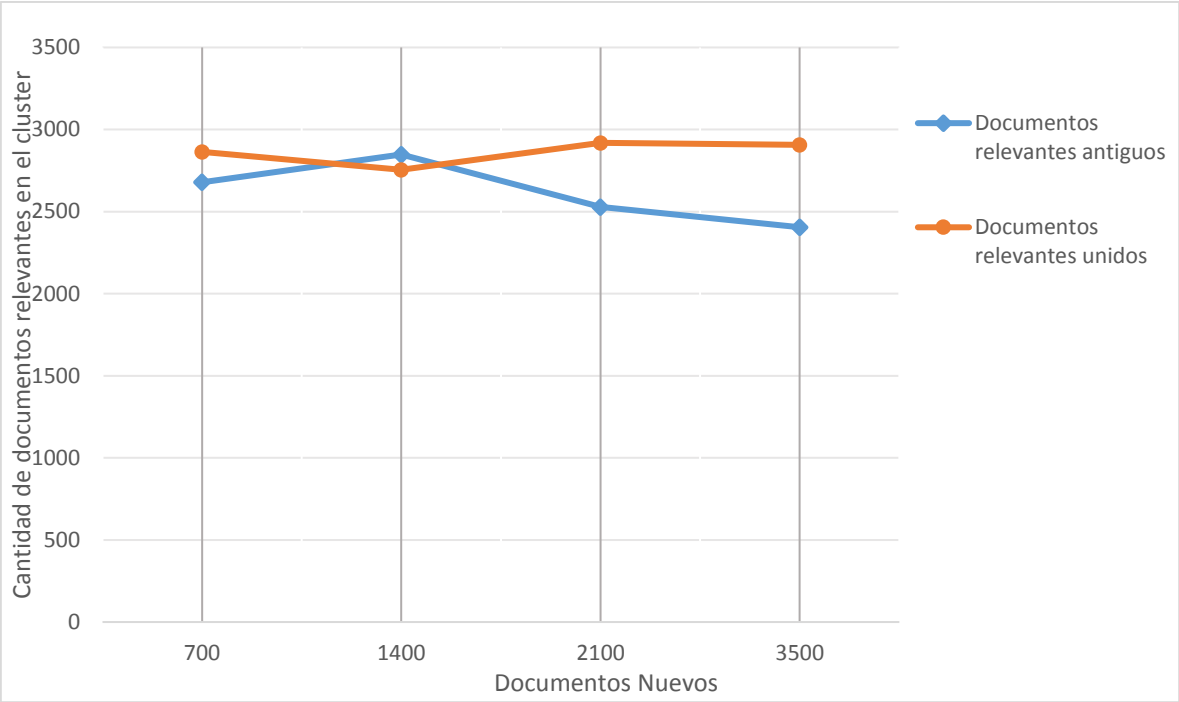

Figura Nº42: Gráfico del tercer experimento de Average link de la segunda etapa.

Experimento 4: En el cuarto experimento se utilizó el algoritmo jerárquico Average link, donde se utilizan dos conjuntos de documentos, el primer conjunto de documentos contiene a los documentos antiguos, mientras que el segundo conjunto de documentos contiene a los documentos unidos. El conjunto de documentos antiguos estaba compuesto por 3500 documentos. Mientras que el conjunto de documentos nuevos varía en 700, 1400, 2100 y 3500 documentos. Además, el número de términos es de 1400. En la Tabla Nº40 se muestran los resultados obtenidos con las distintas cantidades de documentos, los cuales también se pueden observar gráficamente en la Figura Nº43.

Documentos		Documentos relevantes visitados	
Antiguos	Nuevos	Doc. Antiguos	Doc. Unidos
3500	700	2966	3144
3500	1400	2943	3161
3500	2100	2279	3570
3500	3500	2671	3265

Tabla Nº40: Resultados del cuarto experimento de Average link de la segunda etapa.

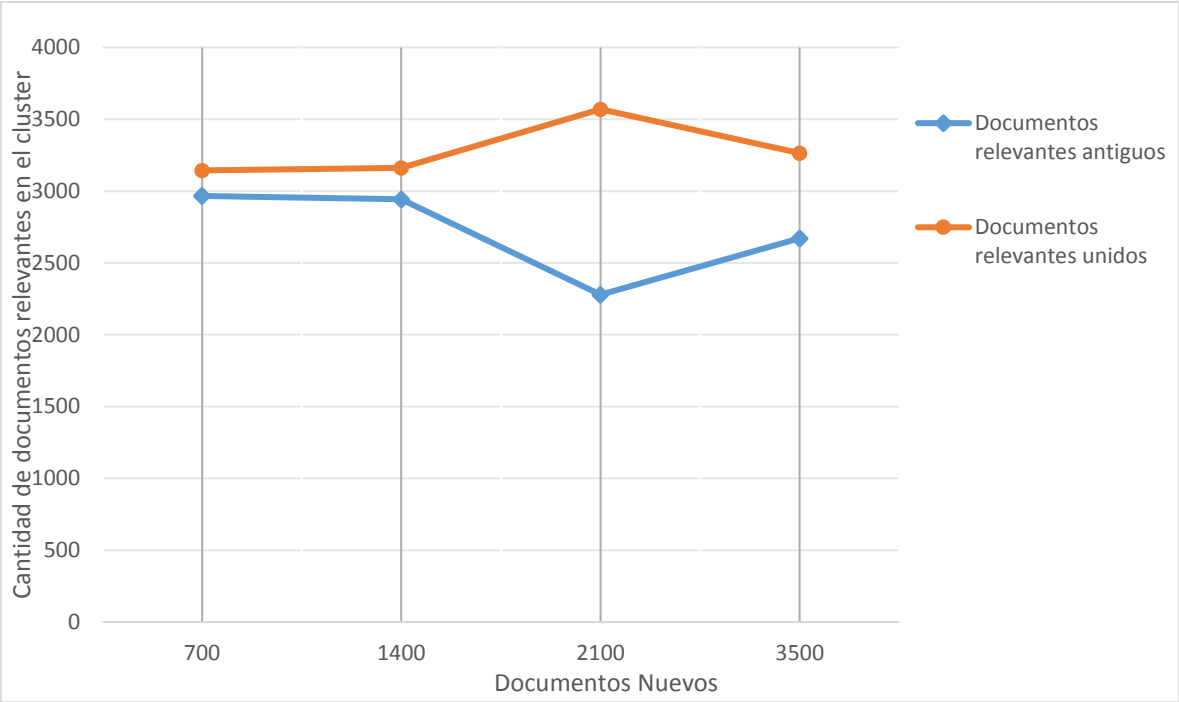

Figura Nº43: Gráfico del cuarto experimento de Average link de la segunda etapa.

5.6 Conclusiones de los experimentos

Luego de la realización de los experimentos de precisión, es importante destacar que el enfoque principal de esta investigación es mejorar la precisión, lo cual se cumple en todos los experimentos realizados tanto en la primera como en la segunda etapa de los experimentos. Estos experimentos se realizaron con diferentes cantidades de documentos nuevos y antiguos. Como se dijo anteriormente, la precisión aumenta en todos los experimentos, por lo que se puede concluir que resulta conveniente rehacer el cluster en todos los casos. Sin embargo, el aumento más notorio de precisión se ve cuando la cantidad de documentos antiguos es menor a la cantidad de documentos nuevos. De esto se puede deducir que, al tener una cantidad de documentos nuevos ampliamente mayor a la cantidad de documentos antiguos, aumenta la probabilidad de que estos documentos nuevos sean relevantes, esta situación se ve reflejada en el experimento 4 de precisión, de ambas etapas experimentales, donde la diferencia entre documentos nuevos y antiguos fue mayor, que es el caso de 700 documentos antiguos y 3500 documentos nuevos. Por el contrario, el menor aumento en la precisión se produce cuando los documentos antiguos son 3500 y los

nuevos 700, situación ocurrida en el experimento 1 de precisión de las dos etapas experimentales. Se debe destacar que al aplicar el algoritmo Monte Carlo al algoritmo probabilístico se logra mejorar la precisión en todas las pruebas realizadas, aunque, se debe tener en cuenta que la aplicación de este algoritmo tiene como consecuencia un mayor costo en tiempo ejecución. Sin embargo, el objetivo principal de esta investigación es mejorar la precisión, por lo tanto, será conveniente aplicar el algoritmo Monte Carlo.

Teniendo en cuenta los resultados de los experimentos de precisión, donde se comprueba que es conveniente rehacer el cluster, se realiza los experimentos para determinar que método para rehacer el cluster resulta más conveniente. El parámetro a considerar para determinar que algoritmo tiene mejor desempeño, es la cantidad de documentos relevantes visitados al recorrer los clústeres generados con cada uno de los algoritmos. De los métodos utilizados, el que mejor resultados entrega es el algoritmo Average Link. La eficacia del algoritmo Average Link es notoriamente mayor al de Single Link y Complete Link, tal como queda demostrado en los experimentos de la sección 5.5.1.3 y 5.5.2.3. Al igual que en los experimentos de clustering el mejor resultado se obtiene cuando la diferencia entre documentos antiguos y nuevos es mayor, siendo el caso de 700 documentos antiguos y 3500 nuevos el que obtiene mejores resultados. Por el contrario, pese a tener resultados favorables, cuando se tienen una cantidad de documentos antiguos notoriamente mayor a los nuevos, como es el caso de 3500 antiguos y 700 nuevos se encuentran los resultados más acotados.

En cuanto a los algoritmos Single Link y Complete Link los resultados demostraron una tendencia similar en relación a cuando mejora, o disminuye la cantidad de documentos relevantes visitados al recorrer el cluster, sin embargo, se observó una diferencia entre los documentos relevantes visitados al recorrer el cluster, la cual es más notoria al momento de utilizar Complete Link, tal como se puede observar en los experimentos 2 y 3 de Complete Link de la primera etapa y en el experimento 1 de la segunda etapa. Si bien Single Link mostraba un comportamiento similar, este solo puede ser apreciado en el experimento 1 de Single Link de la segunda etapa.

CAPITULO 6: CONCLUSIONES GENERALES

6.1 Contribuciones

Esta investigación tiene como objetivo la obtención de datos empíricos de algoritmos en línea para el clustering de documentos. Para cumplir este objetivo se fijó una serie de objetivos específicos.

- Analizar los conceptos involucrados en recuperación de la información y de los algoritmos de clustering. Investigar sus principales características y comprender su problemática.

  En los capítulos 2 y 3 se analizan los distintos elementos que componen la teoría de recuperación de la información y de sistemas de recuperación de la información. El capítulo 2 se encuentra centrado en la recuperación de la información. Se comienza explicando como la representación de documentos y consultas, las operaciones de consultas, así como también el matching entre consultas y documentos. Luego, se definen los sistemas de recuperación de información y como estos son evaluados. También se escogen la medida que será utilizada al momento de realizar los experimentos. Posteriormente en el capítulo 3, son definidos los conceptos envueltos en el clustering de documentos.

- Revisar una tesis realizada previamente (Frederick Lara - Delia Moncada, 2016), la cual servirá de base para extender los experimentos.
  Se estudia el proyecto de título realizado por Frederick Lara - Delia Moncada (2016), el cual ayuda a comprender a grandes rasgos los procesos necesarios para la realización de los experimentos.
- Implementar y analizar el comportamiento de, al menos, tres algoritmos de clustering en línea (Single link, Complete link, Average link).

  Para cumplir este objetivo se implementaron los algoritmos de clustering mencionados, y se estudia su funcionamiento.

- Desarrollar un algoritmo que permita realizar los experimentos bajo otras condiciones.

Se realiza la implementación del algoritmo tipo Monte Carlo, con el fin de mejorar la precisión obtenida con el uso del algoritmo probabilístico. En todos los experimentos realizados se logra mejorar la precisión, aunque a un costo de tiempo de ejecución bastante elevado, sin embargo, el objetivo principal es mejorar la precisión por lo que los resultados obtenidos se consideran satisfactorios.

- Obtener resultados experimentales, considerando como parámetros el número de documentos en los clusters, el número de términos relevantes (para los documentos que ya están en el cluster), el número de documentos nuevos entre otros.

  A lo largo del capítulo 5 se realizaron dos tipos de experimentos tanto para el estudio de la precisión y de los algoritmos de clustering, estos fueron divididos en dos etapas experimentales en donde se obtienen los resultados bajo distintos escenarios, los cuales involucran cambios en la cantidad de términos y documentos. La primera etapa consta en la extensión de los experimentos para el algoritmo probabilístico, mientras que la segunda etapa consta de los experimentos realizados con el algoritmo probabilístico de tipo Monte Carlo para simular los juicios de usuario.

- Evaluar y comparar resultados obtenidos con el fin de determinar conveniencia del algoritmo en función de precisión y cantidad de documentos.

  Son comparados los resultados obtenidos, posteriormente son analizados y se obtienen conclusiones en base a ellos.

6.2 Trabajos Futuros

En cuanto a los trabajos futuros que pueden ser considerados a partir de esta investigación se pueden destacar dos ideas, en primer lugar se pueden extender los experimentos realizados para evaluar diferentes escenarios experimentales, para así conseguir mejoras en los valores tanto de precisión como de eficacia de los algoritmos de clustering.

Bibliografía

- Anderberg, M.R. (1973). Cluster Analysis for Applications. New York: Academic Press. Anick, P.G. and Vaithyanathan, S. (1997). Exploiting clustering and phrases for context-based information retrieval. In Proceedings of the 20th Annual ACM SIGIR Conference, pp. 314-323. Philadelphia, PA.

- Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- Barry, C.L. (1994). User defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 45(3).

- Bharat, K. and Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In In SIGIR Conference on Research and Development in Information Retrieval.

- Borlund, P. and Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, 53.

- Broder, A. (2002). A taxonomy of web search. SIGIR Forum, 36(2).

- Chowdhury, G. (2010). Introduction to Modern Information Retrieval, Third Edition. Facet Publishing, 3rd edition.

- Croft, W.B. (1977). Clustering large files of documents using the single-link method. Journal of the American Society for Information Science, 28.

- Croft, W.B. (1978). Organizing and searching large files of document descriptions. Ph.D. Thesis, Churchill College, University of Cambridge.

- Croft, W.B. and Harper, D.J. (1979). Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35.

- Dubin, D.S. (1996). Structure in document browsing spaces. Ph.D. Thesis, School of Information Sciences, University of Pittsburgh.

- Ellis, D., Furner-Hines, J., Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. Perspectives in Information Management, 3(2).

- Good, I.J. (1958). Speculations concerning information retrieval. Research report PC-78, IBM Research Centre, Yorktown Heights, New York.

- Gordon, A.D. (1987). A review of hierarchical classification. Journal of the Royal Statistical Society, Series A, 150(2).

- Griffiths, A., Robinson, L.A., Willett, P. (1984). Hierarchic agglomerative clustering methods for automatic document classification. Journal of Documentation, 40(3).

- Gutiérrez-Soto, C. (2016). Exploring the Reuse of Past Search Results in Information Retrieval.

- Harman, D. (1993). Overview of the first trec conference. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93, pages 36–47, New York, NY, USA. ACM.

- Jardine, N. and Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. Computer Journal, 11(2):177-184.

- Jardine, N. and Sibson, R. (1971). Mathematical Taxonomy. New York: Wiley.

- Jardine, N. and van Rijsbergen, C.J. (1971). The use of hierarchical clustering in information retrieval. Information Storage and Retrieval, 7:217-240.

- Jones, W.P. and Furnas, G.W. (1987). Pictures of relevance: A geometric analysis of similarity measures. Journal of the American Society for Information Science, 38(6):420-442.

- Keen, E. M. (1992). Presenting results of experimental retrieval comparisons. Inf. Process. Manage., 28(4) :491–502.

- Kirriemuir, J.W. and Willett, P. (1995). Identification of duplicate and near-duplicate full-text records in database search-outputs using hierarchic cluster analysis. Program, 29(3):241-256.

- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment.J. ACM, 46(5) :604–632.

- Lewis, D. D. and Jones, K. S. (1996). Natural language processing for information retrieval. Commun. ACM, 39(1):92–101.

- Milligan, G.W., Soon, S.C., Sokol, L.M. (1983). The effect of cluster size, dimensionality, and the number of cluster on recovery of true cluster structure. IEEE Transactions on Patter Recognition and Machine Intelligence, 5(1).

- Moncada Delia y Lara Frederick (2016) Algoritmo de clustering en línea, para recuperación de la información.

- Ouksel, A. (2002). Mining the world wide web: An information search approach by george chang, marcus j. healey (editor), james a. m. mchugh, jason t. l. wang. SIGMOD Rec., 31(2) :69–70.

- Peat, H.J. and Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the American Society for Information Science, 42(5):378-383.

- Preece, S.E. (1973). Clustering as an output option. Proceedings of the American Society for Information Science, 10.

- Reid, J. (2000). A task-oriented non-interactive evaluation methodology for. Information Retrieval Systems, Information Retrieval, 2 :115–129.

- Robertson, S. E., van Rijsbergen, C. J., and Porter, M. F. (1981). Probabilistic models of indexing and searching. In Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval, SIGIR ’80, pages 35–56, Kent, UK, UK. Butterworth & Co.

- Rocchio, J.J. (1966). Document retrieval systems - Optimization and evaluation. PhD Thesis, Report ISR-10 to the National Science Foundation, Harvard Computation Laboratory.

- Rorvig, M. (1999). Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. Journal of the American Society for Information Science, 50(8):639-651.

- Salton, G. (1971). The SMART Retrieval System and ;Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

- Salton, G., (ed.) (1971). The SMART Retrieval System - Experiments in Automatic Document Retrieval. New Jersey, Englewood Cliffs: Prentice Hall Inc.

- Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. Commun. ACM, 18(11) :613–620.

- Salton, G. and Wong, A. (1978). Generation and search of clustered files. ACM Transactions on Database Systems, 3(4):321-346.

- Salton, G. and McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

- Salton, G. and McGill, M. J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.

- Salton, G., Buckley, C.: Readings in information retrieval. In Sparck Jones, K., Willett, P., eds.: Readings in information retrieval. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997) 355-364.

- Schamber, L., Eisenberg, M.B., Nilan, M.S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. Information Processing & Management, 26(6):755-776.

- Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single link cluster method. Computer Journal, 16:30-34.

- Sneath, P.H.A. and Sokal, R.R. (1973). Numerical taxonomy: the principles and practice of numerical classification. San Francisco: W.H. Freeman.

- Sparck Jones, K. (1971). Automatic keyword classification for information retrieval. London: Butterworths.

- Sparck Jones, K. and Willett, P., editors (1997a). Readings in Information Retrieval. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

- Theodoridis, S. and Koutroumbas, K. (1999). Pattern Recognition. San Diego: Academic Press.

- Tombros, A., Villa, R., and Rijsbergen, C. J. V. (2002). The effectiveness of query-specific hierarchic clustering. In in information retrieval. Information Processing and Management.

- Tombros, A., Rijsbergen, C.J.V.: Query-sensitive similarity measures for informa-tion retrieval. Knowledge and Information Systems 6 2004.

- van Rijsbergen, C.J. and Sparck Jones, K. (1973). A test for the separation of relevant and nonrelevant documents in experimental retrieval collections. Journal of Documentation, 29(3).

- Van Rijsbergen, C.J. (1979). Information Retrieval. London: Butterworths, 2nd Edition.

- van Rijsbergen, C.J, Harper D.J., Porter, M.F. (1981). The selection of good search terms. Information Processing & Management, 17.

- Van Rijsbergen, C.J. (1986). A new theoretical framework for information retrieval. In Proceedings of the 9th Annual ACM SIGIR Conference, pp. 194-200. Pisa, Italy.

- Voorhees, E.M. (1985a). The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval. Ph.D. Thesis, Technical Report TR 85-705 of the Department of Computing Science, Cornell University.

- Voorhees, E.M. (1985b). The cluster hypothesis revisited. In Proceedings of the 8th Annual ACM SIGIR Conference, pp. 188-196. Montreal, Canada.

- Ward, J.H. (1963). Hierarchical grouping to minimize an objective function. Journal of the American Statistical Association, 58:236-244.

- Watanabe, S. (1969). Knowing and guessing: a quantitative study of inference and information. New York: Wiley.

- Willett, P. (1983). Similarity coefficients and weighting functions for automatic document classification: an empirical comparison. International Classification, 3:138-142.

- Willett, P. (1988). Recent trends in hierarchic document clustering: A critical review. Information Processing & Management, 24(5).