# Contents

# 1 Introducción y definición del problema

En los últimos años el volumen de correo electrónico que reciben los usuarios a diario ha crecido vertiginosamente[30]. Prueba de ello es que compañías prestadoras de servicio gratuito de correo tales como yahoo, google y hotmail han aumentado la capacidad de almacenamiento por usuario. Esta gran cantidad de información hace cada vez más difícil la tarea de administración del correo por parte del usuario [40, 25] y en consecuencia, la comunidad científica ha puesto su atención en el desarrollo de mejores sistemas automáticos para categorización de correo. En los últimos años se han desarrollado diversas técnicas para la categorización de correo de manera automática con buenos resultados. Sin embargo la mayoría de estas técnicas se basan en aprendizaje supervisado y los trabajos que han utilizado aprendizaje no supervisado se han enfocado en una estructura jerárquica de fólderes. En el primer caso, en los enfoques bajo aprendizaje supervisado, el usuario establece previamente los fólderes en los que desea almacenar los mensajes y el sistema se entrena para que aprenda a identificar cada categoría. Un caso especial de categorización automática de correo al que se ha prestado especial atención es el filtrado de correo spam (correo no solicitado generado de manera automática) y en el cual se han logrado grandes avances en las técnicas utilizadas. En el segundo caso, bajo aprendizaje no supervisado, si bien se realiza una categorización completamente automática, el enfoque más común utilizado hasta el momento se basa en una organización de los mensajes en fólderes presentando limitaciones sobre todo cuando involucra grandes volúmenes de información. Por esta razón nuevos enfoques utilizados para la organización de información utilizados con éxito en problemas de categorización automática de textos (organización de documentos, búsqueda de información en Internet) tales como redes semánticas y mapas conceptuales pueden ser utilizados para crear sistemas de categorización automática de correo de tal manera que le permitan al usuario visualizar, analizar y extraer información de manera más fácil que en los sistemas tradicionales.

# 2 Estado del Arte Categorización Automática de Correo Electrónico

A continuación se hace una revisión del estado del arte de los sistemas automáticos para categorización de correo utilizados hasta el momento así como de las nuevas técnicas que están bajo estudio . la sección se presenta de la siguiente manera: En la sección 2.1 se hace una descripción general del problema de categorización de textos y los enfoques utilizados para desarrollar este tipo de sistemas. En la 2.2 se hace una revisión de los trabajaos realizados

para clasificación de correo y en la 2.3 se hace una revisión de los trabajos en categorización de correo bajo aprendizaje no supervisado. En la sección 2.4 se revisa el panorama de las técnicas utilizadas para visualizar los resultados de los sistemas categorizadores de correo.

## 2.1 categorización automática de de correos electrónicos

El problema de categorización automática de correo electrónico consiste en organizar y presentar de manera apropiada la información contenida en el buzón de tal manera que le permita al usuario visualizar y extraer la información contenida en estos. La categorización automática de e-mails es un problema particular de la categorización automática de textos, si bien ambos tienen muchos puntos en común también tienen muchos en los que difieren, y necesitan de soluciones específicas. Uno de los puntos clave es que la organización de correo eletrónico es subjetiva y su estructura (folderes, carpetas o grupos) cambia constantemente. El usuario puede que algunas veces quiera ver su correo por remitente, otros por tema u otros por fecha. Prácticamente el problema de categorización de correos se puede subdividir en 4 subproblemas: Representación del mensaje (extracción de características), selección de las características más relevantes, el proceso de categorización propiamente dicho, y por último la presentación de manera adecuada de la categorización al usuario. Para la extracción de características la información de los correos puede pertenecer a diferentes partes de la estructura del mensaje[14]:

1. Información personal ( To Remitente, CC, destinatarios)
2. Información de hilos (un mensaje puede ser la respuesta a otro mensaje y este a su vez de otro mensaje).
3. Fecha de envío
4. Subject o asunto
5. Cuerpo del mensaje

Cada uno de estas partes del mensaje es un problema de extracción de características a tratar. Por ejemplo, la fecha es un dato no categórico, y los destinatarios podrían ser tanto categóricos (las direcciones de origen) como numéricos(número de destinatarios).El cuerpo del mensaje en si, ya es un problema de clasificación de textos que se puede abordar utilizando extracción de palabras, de frases o utilizando de redes semánticas.

La categorización implica, a parte de determinar la mejor técnica de clasificación o de agrupamiento, encontrar la manera más adecuada de medir la similaridad entre mensajes y la manera de que el sistema etiquete de forma automática cada grupo de mensajes encontrado. Una vez obtenida la categorización se debe determinar la mejor forma de presentar los resultados al usuario.

La categorización automática de correo se puede realizar utilizando dos enfoques: El primero es bajo aprendizaje supervisado(clasificación) y el segundo,

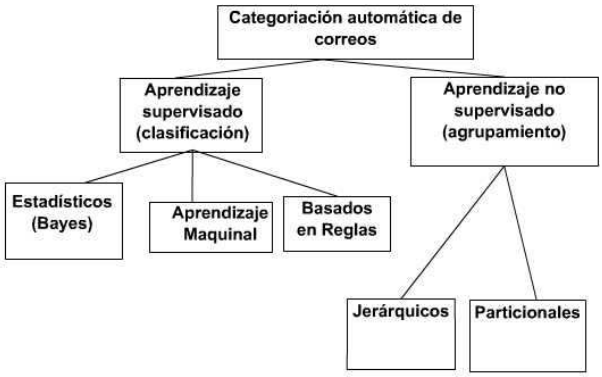


Figure 1: Principales Técnicas aplicadas para la categorización de correo electrónico

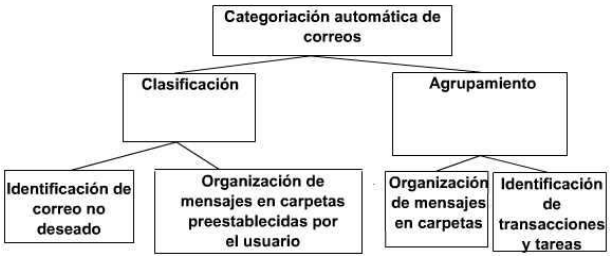


Figure 2: Principales Aplicaciones de la Categorización de correo electrónico

bajo aprendizaje no supervisado(agrupamiento). En el primer caso el usuario es quien determina las categorías en las cuales se va a colocar los correos electrónicos, por lo general, son carpetas que el usuario crea y que el sistema utiliza para ubicar los e-mails dentro de ellos. En el segundo caso, el sistema mismo es el que genera las categorías en las que se van ubicar los correos. No existen carpetas o categorías preestablecidas por el usuario. Trabajos en categorización automática de correo realizados utilizando aprendizaje supervisado y no supervisado serán revisados en la sección 3 y 4 respectivamente.

### 2.1.1 Selección y Extracción de características

Como se mencionó anteriormente, existe diferente tipo de información que puede ser extraida y seleccionada para realizar el proceso de categorización[14]. Por ejemplo en [29] las características son extraídas a partir de las diferentes partes del mensaje: Remitente: destinatarios, hora/fecha, Asunto, Archivos adjuntos y contenido del mensaje. Esta información conforma el vector de características x = (y w). Los componentes del vector **y** es información estructurada del mensaje (tal como dominio de la dirección del remitente, dia de la semana, tipo de extensión del archivo adjunto y tamaño del mensaje).La información no

estructurada es la que se encuentra en el asunto y en el cuerpo del mensaje, en la sección 4 se da una descrpción más detallada de este trabajo. En el trabajo realizado por Svetlana [22] se incluye como alternativa características temporales de los mensajes para mejorar el sistema clasificador . Sin embargo , gran parte del proceso categorizador recae principalmente en las características extraídas del cuerpo del cuerpo y encabezado del mensaje) .

La técnica más utilizada en la actualidad para clasificación de correo y en general para clasificación de documentos es la denominada bolsa de palabras ("bag of words"). Esta técnica toma aquellas palabras (t) que ocurren con mayor frecuencia dentro del documento (Di):

$$F(t_k, D_i) = \frac{ocurrencias(t_k, D_i)}{N}$$

donde N es el número total de términos en el documento i.

una alternativa es tomar las frecuencias logarítmicas

$$F(t_k, D_i) = log(1 + F(t_k, D_i))$$

Esto debido a que la frecuencia de los palabras es muy irregular (más de la mitad de los términos aparecen solo una vez dentro del documento).

Sin embargo, esta metodología tiene una desventaja. Algunas veces una misma palabra puede presentar una frecuencia elevada no solo en un documento , sino en muchos, por lo tanto no sería una característica representativa de dicho documento. Existen otras técnicas que intentan evitar este problema y que han demostrado tener mejores resultados que la frecuencia de términos. Por ejemplo la del nivel de entropía, la cual está dada por:

$$w_{ki} = F_{log}(t_k, D_i) * (1 + e(t_k, D))$$

donde

$$e(t_k D) = \frac{1}{log|D|} \sum_{J=1} \frac{oc(t_k, D_j)}{oc(t_{k,}, D)} * log \frac{oc(t_k, D_j)}{oc(t_{k,}, D)}$$

y e(tk,D) es la entropía del K-ésimo término(palabra), D la colección de documentos(mensajes de correo en este caso).

Se han desarrollado otras técnicas en donde la extracción de características no se basa en las palabras de los documentos sino, por ejemplo, en frases o en la relación semántica de las palabras. Sin embargo hasta el momento ninguna de dichas técnicas ha podido superar el rendimiento de la técnica de la bolsa de palabras[35] o en algunos casos se ha mejorado el rendimiento a cambio de un incremento en costo computacional. Se han realizado pocos trabajos sobre este tema orientado a clasificación de correo, cabe destacar el trabajo realizado por [11], donde se estudia el uso de frases para el incremento del rendimiento de clasificadores de correos.

Las redes semánticas también se han utilizado con éxito en la extracción de características para la clasificación de textos, aunque no se encontró referencia en trabajos de clasificación de correos, especificamente para reducir la dimensionalidad del conjunto de datos. En el trabajo realizado por Hang y Wertem

[39]utilizan una red semántica para el inglés denominada WordNet para realizar un reducción del número de palabras a utilizar en el proceso de clasificación de textos. Esta red busca palabras con el mismo significado, o que están relacionadas entre si por un mismo concepto para reemplzarlas por un significado más global. Los resultados muestran un incremento en el desempeño del sistema clasificador cuando usan la técnica de bolsa palabras acompañadas de las redes semánticas que la bolsa de palabras por si sola. Un extensivo análisis sobre los algoritmos para la selección de características se puede encontrar en trabajo de Forman[12] .

## 2.2 categorización automática de correos electrónicos bajo aprendizaje supervisado (Clasifcación )

La clasificación automática de correo es un procedimiento que se realiza bajo aprendizaje supervisado. En este tipo de sistemas, el usuario es quien determina las categorías en las cuales se va a colocar los correos electrónicos, por lo general, folderes que el usuario crea y que el sistema utiliza para ubicar los correos, es decir, el usuario es el que determina la estructura de organización del buzón de correos. En el proceso de desarrollo de sistemas automáticos de clasificación intervienen dos etapas: La etapa de entrenamiento, y la etapa de prueba. La primera etapa consiste en presentarle al sistema muestras de ejemplo de cada clase a clasificar. En la segunda etapa, el sistema utiliza el conocimiento adquirido para realizar la clasificación de nuevas muestras. Este tipo de categorización implica que se conozcan de antemano las categorías o clases en las cuales se van a clasificar las muestras (los mensajes de correo).

Dentro de los trabajos realizado en categorización de correo bajo aprendizaje supervisado se han utilizado gran variedad de técnicas de aprendizaje de máquina, que van desde sistemas basados en reglas como el presentado por Cohen en [8] donde se utiliza un sistema de aprendizaje de reglas(RIPPER) , métodos estádisticos (los cuales han sido tradicionalmente utilizados en clasificación de correo) tales como el Bayesiano ([17], [34], [32]) , máquinas con vectores de soporte utilizadas en [6] por Brutlag y Meek , k-vecino más cercano KNN [33] por Payne , hasta técnicas relativamente recientes como sistemas inmunes artificiales utilizados para la detección de correo spam en [36] y en [31].

Los dos tipos de aplicaciones en las que se utiliza este tipo de categorización son por un lado el filtrado de correo no deseado, principalmente sistemas antispam, y por otro la clasificación automática de correos en carpetas.

Hasta el momento existen varios enfoques para crear sistemas anti-spam: Las listas de correos basadas en DNS(DNS-based Blackhole) , los métodos heurísticos y los métodos estadísticos. De estos últimos, el método más utilizado y efectivo es es el filtrado bayesiano, principalmete el algoritmo propuesto por Graham [16] ha sido el punto de partida para el desarrollo de nuevos algoritmos [17], [34], los cuales logran un 99% de aciertos en la identificación de spam. De los trabajos realizados para resolver este tipo de problemas vale la pena resaltar el de Pantel y Lin[32], donde utiliza un clasificador Bayesiano y una lista de stopwords creada dinámicamente, es decir, en lugar de crear una lista de palabras

mas comunes, que es lo que por lo general se hace, la lista se crea directamente a partir de el contenido de los mensajes incluyendo las palabras menos frecuentes en estos. Los resultados experimentales son comparados con el algoritmo RIPPER propuesto por Cohen en [8] , logrando el clasificador Bayesiano un 94% de precisión contra un 86% del clasificador RIPPER.

También se han utilizado técnicas menos tradicionales diferentes al clasificador bayesiano, tales como algoritmos genéticos y redes inmunes. En [20] se utiliza programación genética para crear un sistema anti-spam y es también es comparado con el clasificador Bayesiano, encontrando que ambos sistemas presentan un rendimiento similar aunque advierten que es necesesario realizar mas estudios.

En [36],los autores proponen un sistema inmune artificial para la clasificación de E-mails en correo deseado y no deseado. Las células B del sistema contienen información de las características de los correos no deseados. Los antígenos son los correos a clasificar. El correo no deseado se coloca en un lugar distinto al correo deseado (tal y como lo hace hotmail). Si el usuario borra este correo significa que el sistema hizo bien la clasificación, si no lo borra significa que lo hizo mal. Este proceso de realimentación con el usuario permite que el sistema se mantenga en constante aprendizaje. El artículo compara el sistema inmune propuesto con el clasificador bayesiano, Los resultados experimentales demuestran que el sistema inmune artificial propuesto presenta rendimiento similar al del clasificador bayesiano. Otro trabajo sobre el tema utilizando sistemas inmunes es el realizado por Terri [31] , que se concentra en el correo spam. Los antígenos son los correos a identificar y los anticuerpos contienen información de características de mensajes spam. Igualmente el sistema se adapta en un proceso de realimentación con el usuario cuando este indica que un correo fue identificado erroneamente como spam. . Este trabajo concluye que este sistema no tiene un rendimiento igual que el algorítmo de Graham . Desafortunadamente no incluyen los resultados experimentales de este algoritmo que permitan hacer un análisis comparativo.

En la clasificación de correo en general, no se ha logrado un rendimiento tan alto como los logrados en el filtrado de correo spam[10]. Dentro de este tipo de trabajos cabe destacar:

Cohen [8] : En este trabajo se utiliza un sistema de aprendizaje de reglas(RIPPER) y se compara con el tradicional TF-IDF-Bayesiano demostrando que ambos algoritmos se desempeñan bien con un bajo número de muestras.

Brutlag y Meek [6]: Los autores utilizan el algoritmo de SMV(máquina con vectores de soporte) logando un desempeño del 70 % al 90% de rendimiento(accuracy).

Payne y Edawrds: [33] donde utilizan un sistema de reglas por inducción denominando CN2(Clark & Nibbet 1989) y una variación del algoritmo del k-vecino más cercano KNN , denominado IBPL1.

Si se desea revisar otros trabajos realizados en clasificación de correo, consultar el trabajo de [9]

## 2.3 categorización automática de correos electrónicos bajo aprendizaje no supervisado (Agrupamiento)

El agrupamiento (clustering), como su nombre lo indica es la división de la entrada de datos en grupos que tengan cierta similitud. Mediante aprendizaje maquinal este es un mecanismo que se realiza bajo aprendizaje no supervisado. Existe un gran número de algoritmos y enfoques de para realizar agrupamiento ( acerca de los algoritmos de agrupamiento revisar el trabajo de Berkhin[5]). En text mining las técnicas de agrupamiento se utilizan en diversas etápas que van desde la reducción de la dimensionalidad del vector de características, asociación de palabras, hasta el agrupamiento de documentos propiamente dicho y la recuperación de información[37].

Para agrupamiento de correos electrónicos las técnicas más utilizadas han sido los algoritmos particionales y los algoritmos jerárquicos. La clasificación de correo electrónico ha sido más ampliamnete explorada que el agrupamiento, sin embargo en los últimos años se han venido realizando investigaciones en esta área. A continuación se describe brevemente algunos de los más representativos:

- En [29] realizado por manco et al. se utiliza clustering particional y jearquico para realizar agrupamiento automático de correos electrónicos. Las características son extraídas a partir de las diferentes partes del mensaje: Remitente: destinatarios, hora/fecha, Asunto, Archivos adjuntos y contenido del mensaje. Esta información conforma el vector de características x = (y w). Los componentes del vector **y** es información estructurada del mensaje (tal como dominio de la dirección del remitente, dia de la semana, tipo de extensión del archivo adjunto y tamaño del mensaje).La información no estructurada es la que se encuentra en el asunto y en el cuerpo del mensaje.Habiendo extraido los vectores de características que representan a los mensajes, se define la similaridad entre el vector $x_i$ y el vector $x_j$ de la siguiente forma:

$$s(x_i, x_j) = \alpha s_1(y_i^n, y_j^n) + \eta s_2(y_i^c, y_j^c) + \gamma s_3(w_i, w_j)$$

donde $s_1(y_i^n, y_j^n)$ es la similaridad entre la información estructurada de tipo numérico , $s_2(y_i^c, y_j^c)$ es la similaridad entre la información estructurada de tipo categórico , $s_3(w_i, w_j)$ es la similaridad entre la información no estructurada y $\alpha, \eta, \gamma$ son coeficientes entre 0 y 1 para ajustar la ifluencia de cada una de las partes.

La medida de distancia utilizada para encontrar $s_1$ es la distancia ecludiana, para $s_2$la función de direchlet y para $s_3$la distancia coseno.

El etiquetado de cada cluser se realiza de la siguiente forma: Del vector de datos numéricos $y^n$ se utiliza la media, del vector de datos categóricos $y^c$ se utiliza la moda, y para el vector w, se eligen aquellos términos más frecuentes en el cluster que no se

| Categorical | Numeric |
|---|---|
| Sender domain (e.g. yahoo.com) | Message length |
| Most frequent recipient radix domain (e.g., gov, com, edu) | Nr. of recipients |
| Weekday | Nr. of messages received from the same sender |
| Time period (e.g., early morning, afternoon, evening) | |
| Attachment file extension (e.g., jpg, ps, xls) | |

Figure 3: Información estructurada extraídas de los mensajes de correo (trabajo de Manco et al.)

encuenten en ningún otro cluster. Finalmente, se realiza la agrupación el método jerárquico aglomerativo para encontrar los centroides iniciales y luego se utiliza k-medios.

Para analizar los resultados experimentales se comparó el agrupamiento del sistema con el realizado previamente por un usuario sobre un conjunto de mensajes. El uso de agrupación jerarquica para encontrar los centroides iniciales representó una ganancia en el rendimiento(accuracy) del sistema.El uso de información estructurada numérica no produjo mejores en el desempeño del sistema, en cambio el uso de información estructurada categórica representó un incremento en el desempño del sistema en algunos casos.

- Giacolleto y Aberer proponen en [13] un sistema que integra la estructura de carpetas predefinidas por el usuario con una estructura generada completamente por el sistema extraída a partir de los mensajes de correo electrónico. En este trabajo se utiliza el método de k-medios bisectivo para realizar el proceso de agrupamiento. El número de clusters que realiza el algoritmo depende del número mínimo de carpetas solicitadas por el usuario. La extracción de características se realiza utlizando frecuencia de términos(TF-IDF, term frecuency - inverse document frecuency). Como es posible que al final queden mensajes no relacionados entre si en algunas carpetas el sistema realiza un paso adicional haciendo una revisión de cada carpeta y crea nuevas carpetas en caso de que sea necesario. Luego de tener dos tipos de estructuras de carpetas, una generada por el computador y otra por el usuario, el sistema se encarga de integrarlas en una sola. El etiquetado de las carpetas, es de decir de los grupos(clusters), es un problema basntante dificil de resolver no solo en este caso, si no en los problemas de agrupamiento en en general. En este trabajo cada grupo se etiqueta de la siguiente forma:

- Email nombre - asunto, cuando la mayoría de correos de de cluster se refiere al nombre de una persona sobre determinado asunto. Se eligen los términos más frecuentes que hacen referencia al asunto.
- Asunto: Cuando la característica más observada es el asunto(subject) se

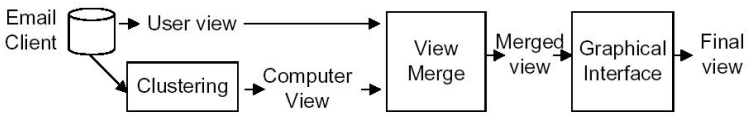


Figure 4: Solución propuesta por Giacolleto y Aberer para el problema de categorización automática de correo electrónico

asinga esta etiqueta a la carpeta.

- Email recibido de - nombre: Cuando la mayoría de correos vienen de determinada persona.
- Discusión - nombre - asunto: Correos que discuten un asunto con una persona. El la característica nombre aparece tanto en el de(from) del mensaje, como en la lista de destinatarios.

Los resultados experimentales de este trabajo ponen de manifiesto dos problemas : El sistema produce carpetas o folderes con correos sin relación, según los autores, debido a la inclusión de información personal de los correos, lo cual implicaría una mejora en la extracción de características de este tipo; El segundo problema se refiere a la falta de claridad de las estructuras de folderes creadas por el sistema, particularmente la de carpetas compartidas por mensajes.Para resolver este problema, los autores plantean dos alternativas: O bien rediseñar la interfaz para lograr un mejor entendimiento por parte del usuario, o bien rediseñar el algoritmo para evitar que se generen estructuras de folderes demasiado complejas.

[21]

Una caso especial de categorización bajo aprendizaje no supervisado que se ha tratado es la clasificación de correos relacionados con transacciones por internet(por ejemplo e-bussines) En este tipo de trabajo se intenta primero determinar la tarea con las que están relacionados los correos para luego generar un modelo que permita describir el proceso de dicha tarea y de ese modo categorizar cada mensaje de correo en la transacción y actividad adecuada. En el trabajo propuesto por kushmerick [23] este proceso se realiza de la siguiente forma:

1. Identifiación de tareas: Particionar los mensajes de acuerdo al tipo de transacción o actividad asociado.
2. Identificación de transiciones: Dentro esa transacción, a que estado de esta pertenece.(envío, confirmación, etc)
3. Inducción automática de un modelo: A partir de una serie de actividades y mensajes identificados generar un modélo que describa el proceso.

4. Clasificación automática de los mensajes: Identificar a que transacción pertence un correo entrante.

El primer paso se realiza agrupando los mensajes con el mismo id de transacción. Luego, para la identificación de transiciones se utiliza clustering jerárquico aglomerativo, utilizando como medida de similaridad una combinación entre LCSS(máxima longitud entre subsequencia comúnes entre dos mensajes) y TF-IDF (frecuencia inversa de términos). A partir de la partición por actividad y por transición del grupo de mensajes, se crea un autómata de estados finitos que describe cada uno de los procesos. En la última etapa, la clasificación de un mensaje entrante de acuerdo al modelo generado, se utiliza aprendizaje supervisado (concretamente, máquinas con vectores de soporte), donde los mensajes con los que se generó el modelo sirven como datos de entrenamiento.

## 2.4 Extracción de Conocimiento, Interfaces y visualización de resultados

Los usuario utilizan los sistemas de correo para realizar principalmente cinco tipos de actividades[38]:

- Flujo de mensajes: Una primera actividad realizada por los usuarios normalmente es determinar que mensajes han llegado recientemente al buzón de correo. Esta acción implica detener la activadad que el usuario esté realizando en el momento para revisar el mensaje entrante. Actualmente los sistemas dan alguna clase de señal y en algunos casos algun tipo información del mensaje para que el usuario tenga algun criterio y determine si lo revisa inmediatamente o no. Un sistema que apoye el primer tipo de actividad debe ser capaz de identificar rápidamente la información contenida en el mensaje y darle una idea inicial al usuario para que este decida si suspende o no la actividad que está realizando.

- Administrar los correos almacenados: Luego de almacenados las personas necesitan poner en diferntes carpetas tantos los mensajes leídos como los no leídos para su posterior recuperacíon. Para realizar este proceso los usuarios utilizan dos métodos: Revisar los mensajes de correo por orden de llegada y por prioridad del correo. Por lo tanto un sistema inteligente debe ser capaz de identificar cuales son los mensajes prioritarios para el usuario y cuales no.

- Administración de tareas: Los correos electrónicos en muchas ocasiones son un soporte para las tareas que tiene que realizar un usuario, recordando sus actividades [28].

- Almacenamiento: Implica el proceso de almacenamiento propiamente dicho realizado principalmente mediante estructura de carpetas jerárquicas. En promedio, una persona utiliza 40 carpetas con una profundidad de tres niveles. Sin embargo, este enfoque tiene la desventaja de que un mensaje solo se puede ubicarse físicamente en una sola carpeta y sin embargo

pertenecer a dos o mas carpetas. Una solución desarrollada por algunos sistemas es realizar consultas para realizar la búsqueda. Al final la visualización de los correos se produce por una serie de consultas definidas previamente por el usuario, o generadas automáticamente por el sistema. Ejemplos de este tipo de sistemas se pueden encontrar en [1, 14, 3].

- Recuperación de información: La quinta actividad depende mucho de la actividad de la almacenamiento, puesto que la fácil ubicación de información depende de una buena organización de los correos. El sistema propuesto utiliza una interfaz que muestra los mensajes al usario agrupados por hilos de conversación.

El segundo problema a tratar cuando se desarrollan sistemas de correos es el de la visualización de información. La visualización de información es un tema aplicado a diferentes áreas, por ejemplo biológia y química( árboles evolución, árboles filogenéticos, mapas moleculares, mapas genéticos) o sistemas de información (estructuras de datos, diágramas de estado transición, redes semánticas, y administración de documentos, solo por mencionar algunos)[18].

El problema que se intenta resolver con la visualización de información es encontrar la relación entre los datos que se le presentan al usuario, donde cada nodo representa un dato o un conjunto de datos y las aristas las relaciones entre ellos.

Existen tres problemas asociados a la visualización de información [18]:

- El tamaño del grafo: El número de datos(nodos) presentados al usuario es un punto a resolver para el algoritmo de visualización. Un grafo demasiado denso va a dificultar la comprensión de la información para el usuario.

- Predictibilidad: Este término hace referencia a que al correr el algoritmo de visualzación varias veces sobre el mismo grafo debe producir la misma representación de los datos.

- Complejidad del algoritmo de visualización: Los algorimtos de visualización permiten al usuario navegar por la información en tiempo real, por lo tanto la complejidad de estos es un punto crítico a tener en cuenta.

En recuperación de información y categorización de documentos se han utilizado diferentes tipos de estructuras para visualizar los resultados: Por ejemplo, el navegador de MeHSB(sistema para la administración de documentos médicos) utiliza una visualización jerárquica para presentar las etiquetas de categorías. Una mejora a este tipo de sistema es la inclusión de metadatos para permitirle al usuario visualizar subcategorías semánticamente asociadas a cada categoría.

Otro tipo de técnica utilizada para visualizar información, como ya se mencionó, es la de agrupamiento. El agrupamiento organiza los documentos basados en su grado de similaridad y los centroides de cada grupo determina el tema que identifica al grupo. Ese tipo de sistemas presenta un resumen temático que identifica cada grupo. Esto permite al usuario explorar los grupos de de su interes

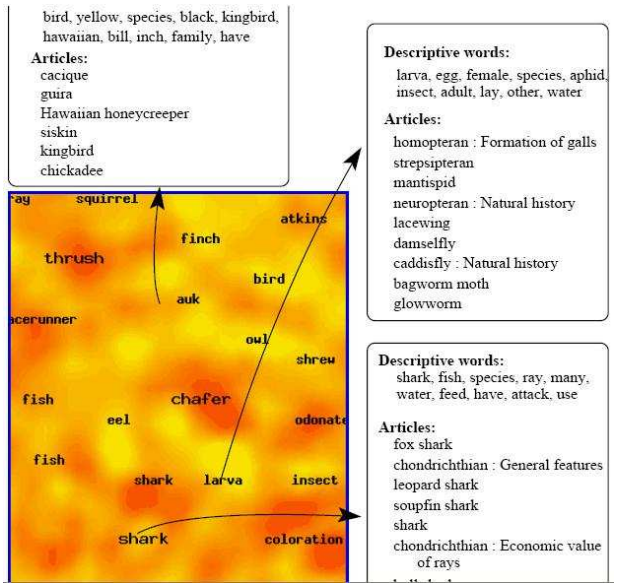


Figure 5: Interfaz utilizando Mapas autoorganizados

y al examinar cada uno de ellos volver a utilizar agrupamiento. Una técnica de agrupamiento utilizada para visualizar información son los mapas autoorganizados, donde cada region de documentos está caracterizada por palabras o frases y la cercanía entre regiones indican algun tipo de relación semántica entre estas. Para una revisión mas detallada de estos trabajos consultar[4].

Para la visualización de los mensajes correos electrónicos los sistemas utilizan por lo general arboles de carpetas, debido, como se mencionó anteriormente, a que estos organizan los correos de manera jerárquica. Algunas interfaces interesantes que caben mencionar [38] agrupando por hilos de conversación, ya que una de las estrategias de visualización de los usuarios depende del número de mensajes que se puedan ver.

El uso de mapas autoorganizados para visualizar información ha sido ampliamente utilizado en organización de documentos (ejemplo de este tipo de trabajos se puede encontrar en [24, 39, 27, 26]) y también se ha empezado a explorar para visualizar mensajes de correo[2].

# 3 Sistema Categorizador propuesto

## 3.1 Representación de los mensajes

La primera parte del preprocesamiento incluye representar la información contenida en los mensajes de correo de una manera más estructurada. El formato escogido con este fin fue Xml. A partir de esta nueva representación se realizan

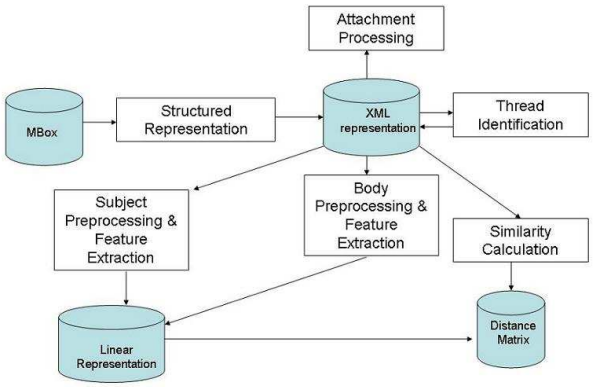


Figure 6: Arquitectura del Sistema

dos tareas de preprocesamiento indispensables: La identificación de hilos y el procesamiento del cuerpo de los mensajes.

Con los arhivos procesados de esta manera se procede a realizar la selección de características, que para el caso del cuerpo del mensaje se utlizará inicialmente la técnica de recuencia de términos. Las técnicas para la extracción del resto de características no están aun definidas y hacen parte del trabajo actual, por lo tanto no se presenta en este documento.

A partir de la representación Xml de los mensajes y de la extracción de características se construye una representación lineal de los mensajes(inicialmente se utilizará vectores de características) y una matriz de distancia entre estos para de esta manera proceder a hacer la categorizacíon.

A continuación se describe brevemente cada una de las fases de procesamiento.

### 3.1.1 Identificación de campos importantes

En la sección anterior se decribieron la estrucutra general de un correo electrónico. A continuación se presentan de manera más detallada los campos utilizados para la realización de este trabajo. La mayor parte de la información del mensaje está en el encabezado, el cual se puede dividir en cuatro grupos de campos: fecha de envío, los datos de origen, la dirección de destino y los campos de información . El primer campo mencionado está conformado por el tipo de dato "Date", seguido la especificación de la hora. El segundo campo es una serire de datos que contienen información de la persona que envía el mensaje. Contiene el campo “from”, y puede contener la máquina de origen(“ sender”) y los campos reply-to. Si el campo “from” contiene mas de una dirección de correo,el campo campo “sender debe obligatoriamente aparecer. El tercer grupo de campos contienen the information del destinatario: to, cc, bcc fields. ("To:"

lista de direcciones), ("Cc:" lista de direcciones), ("Bcc:" lista de direcciones). The bcc (Blind Carbon Copy) es un campo para procesar con cuidado puesto que puede aparecer o no dependiendo del origen y el destinatario. Los campos de información son opcionales, (subject, comments y keywords).

Adicionalmente existen campos que presentan el historail reciente de los mensajes, pero el problema de estos es que no siempre están implementados de la misma manera.

### 3.1.2 Representación Estructurada (XMTP)

Lo primero es transformar los mensajes del formato original(Mbox, por ejemplo) en el formato XML para obtener una representación más estructurada del mensaje. Por otro lado para hacer uso del protocolo XMPT, el cual es una versión xml del protocolo SMPT. Una de las ventajas de este protocolo es que los mensajes se pueden transformar facilmente en código html para presentarselo al usuario.

### 3.1.3 Identificación de hilos

### 3.1.4 Preprocesamiento del asunto y del cuerpo del mensaje (Normalización, Stop words, Stemming, Lemmatization )

Un hilo es una conversación entre dos o mas personas mediante el intercambio de mensajes de correo electrónico.La identificación de hilos no es una tarea sencilla ya que no todos los servidores de correo incluyen información estructurada en el mensaje de respuesta de la misma manera. Algunos sistemas de correo copian el ID del mensaje o algún otro campo de identificación del mensaje padre. Otros, copian el asunto(subject) padre al asunto hijo precedido por RE:.

La diversidad de clientes de correo ha impedido seguir un estandar para la representación de hilos en mensajes de correos. Por otro lado, los estandares no ayudan en la tarea de recuperar la estructura de los hilos ya que es frecuente la eliminación de los campos In-Reply-To por los programas de almacenamiento.

Existen tres tipos de información para la identificación de un hilo[25]:

Asunto (Subject) : Ayuda a la identificación de si un mensaje pertenece a un hilo de mensaje pero no indica a que parte del hilo pertenece.

Información delimitada por símbolos: Consiste en la parte del cuerpo que contiene texto del mensaje padre y que por lo general está delimitada por el caracter ">".

Información no delimitada: Es la parte la información del cuerpo que contiene el texto del mensaje hijo es decir es la información del cuerpo del mensaje en si.

El uso de información marcada y del asunto de mensajes es la mejor manera de identificar hilos de mensajes. Para una revisión detallada del proceso de identificación de hilos examinar [25].

Para la realización de este trabajo se utiliza la identificación de hilos examinando el asunto(subject) del mensaje padre. Para evitar la desventaja mencionada anteriormente en este tipo de técnica, es decir que no se puede deter-

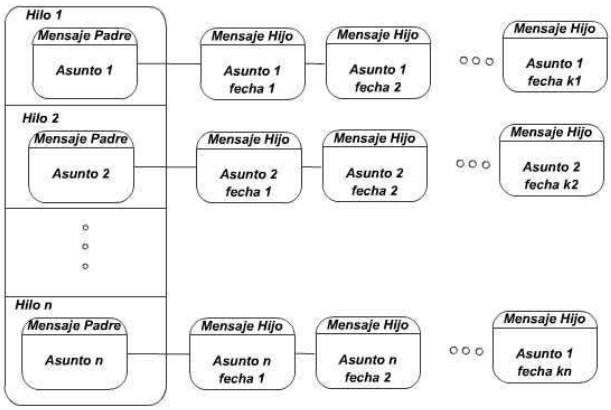


Figure 7: Representación vectorial de los hilos

minar la posición del mensaje dentro del hilo, se utiliza la fecha de envío del mensaje.

Es decir, un hilo es un vector de hilos con el mismo asunto(subject), cuyo primer elemento es el mensaje padre y los elementos que le siguen son los mensajes hijo ordenados cronológicamente.

## 3.2 Procesamiento del Asunto y el Cuerpo del mensaje

Eñ procesamiento del cuerpo y del asunto del mensaje involucra cuatro tareas que se describen brevemente a continuación[15]:

1. Normalización: Consiste en eliminar puntuación, eliminar acentos, y convertir todo el texto en minúsculas.

2. Eliminación de StopWords: En esta etapa las palabras mas comunes, como artículos, preposiciones , disyunciones y conjunciones son eliminadas del texto.

3. Stemming: Cada uno de las palabras del texto es reducida a su forma raíz. Por ejemplo se elminan la conjugación de los verbos, y se deja la raiz de este. Para este propósito existen muchos algoritmos, entre ellos el de Martin Potter[7].En este trabajo se utiliza el algoritmo de "snowball" tambien creado por este autor.

4. Lematización: Las palabras con el mismo significado son cambiadas por una única forma, por ejemplo un sinónimo.

En este trabajo se utilizan las bibliotecas "Lucene" [15] para realizar el procesamiento anteriormente descrito.

### 3.2.1 Preprocesamiento de Archivos Adjuntos.

## 3.3 Arquitectura de Preprocesamiento

## 3.4 Extracción de características

### 3.4.1 Frecuencia de palabras

La técnica más utilizada en la actualidad para para clasificación de textos es la denominada bolsa de palabras ("bag of words"). Esta técnica toma aquellas palabras (t) que ocurren con mayor frecuencia dentro del documento (Di):

$$F(t_k, D_i) = \frac{ocurrencias(t_k, D_i)}{N}$$

donde N es el número total de términos en el documento i.

una alternativa es tomar las frecuencias logarítmicas

$$F(t_k, D_i) = log(1 + F(t_k, D_i))$$

Esto debido a que la frecuencia de los palabras es muy irregular (más de la mitad de los términos aparecen solo una vez dentro del documento).Sin embargo, esta metodología tiene una desventaja. Algunas veces una misma palabra puede presentar una frecuencia elevada no solo en un documento , sino en muchos, por lo tanto no sería una característica representativa de dicho documento. Una técnica que intenta solucionar este problema es

$w_{k,i} = F(t_k, D_i) log(\frac{|D|}{|D_i \in D | t_k \in D_i|})$

denominada Term Frecuency Inverse Document (tfidf), donde el termino $F(t_k, D_i)$ denota frecuencia del término tk en el documento i. D es el número total de documentos y $|D_i \in D | t_k \in D_i|$el número de documentos que contienen al término tk.En otras palabras, un término es relevante si tiene alta ocurrencia dentro de un documento y baja en los otros documentos.

Existen otras técnicas que intentan evitar este problema y que han demostrado tener mejores resultados que la frecuencia de términos. Por ejemplo la del nivel de entropía, la cual está dada por:

$$w_{ki} = F_{log}(t_k, D_i) * (1 + e(t_k, D))$$

donde

$$e(t_k D) = \frac{1}{log|D|} \sum_{J=1} \frac{oc(t_k, D_j)}{oc(t_{k,}, D)} * log \frac{oc(t_k, D_j)}{oc(t_{k,}, D)}$$

y e(tk,D) es la entropía del K-ésimo término(palabra), D la colección de documentos(mensajes de correo en este caso).

Para la extracción de características mediante la técnica de frecuencia de palabras se utilza las bibliotecas SOMLIB.

### 3.4.2 Medidas de similaridad

Como se mencionó anteriormente, un mensaje de correo electrónico está compuesto por tipo de información diversa. Por está razón no es fácil encontrar una única medida de similaridad que funcione adecuadamente. Se utilizaron las medidas de similaridad mas comunes para estos tipos de datos ( para frecuencia de palabras y para datos de tipo numérico,por ejemplo la fecha): Distancia euclidiana y distancia seno.

La distancia euclidiana entre P=(p1,p2,p3,..,pn) y Q=(p1,p2,p3,...,qn) se define como:
$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$
la distancia coseno entre P y Q se define como:
$\frac{A \cdot B}{\|A\|\|B\|}$

## 3.5 Categorización

### 3.5.1 Modelo propuesto

### 3.5.2 Agrupamiento de los mensajes de correo

Para la categorización el correo se exploraron las siguientes técnicas bajo aprendizaje no supervizado; Agrupamiento jerarquico, agrupamiento particional y mapas autoorganizados.

**Agrupamiento aglomerativo enlace simple [19]**

1. Se inicia con todos los patrones desagrupados, estando en el nivel L(0)=0 y la sequencia de agrupamiento m=0;

2. Se encuentra el par de grupos mas similar (r) y (s) de acuerdo con:$d[(r),(s)] = min\, d[(i),(j)]$,es decir la distancia mínima entre todos los subgrupos de r y los subgrupos de s.

3. Se incrementa la sequencia m: m=m+1. Se unen los grupos (r) y (s) en un nuevo grupo para formar el siguiente agrupamiento m.$L(m) = d[(r),(s)]$

4. Se actualiza la matriz de proximidad, D, eliminando los correspondientes grupos (r) y (s) y se adicionan las correspondientes filas y columnas con el nuevo grupo creado. La proximidad entre el nuevo grupo (r,s) y el antiguo grupo(k) se define como:$d[(k),(r,s)] = min\, d[(k),(r)],d[(k),(s)]$

5. Si todos los patrones se encuentran en algun grupo, se detiene el algoritmo , si no se vuelve al paso 2.

**Agrupamiento particional k-medios**

El objetivo del algoritmo de k medios es particionar el conjutno de patrones en k grupos. Para llevar acabo esto, se definen k centroides, uno para cada grupo. Inicialmente, estos centroides se eligen de manera aleatoria. Luego se procede a

encontrar para cada patrón el centroide mas cercano y asginarlo a su respectivo grupo. Uno vez asignados todos los datos a un grupo se calcula nuevamente los centroides como el centro de gravedad de cada grupo y se vuelve a repetir el paso de reasignación de datos a cada centroide. Este procedimiento se repite hasta que los centroides no cambien ya deposición.

1. Se inicializan los k centroides
2. se asigna cada dato al que tiene el centroide más cercano.
3. Cuando todos los datos han sido asignados, se recalculan las posiciones de los k centroides.
4. Se repiten los pasos 2 y 3 hasta que los centroides no se mueven mas.

**Mapas Autoorganizativos**

Un mapa auto organizado es una grilla conformada por neuronas(unidades) cada una de las cuales se representa por un vector de pesos de dimensión d. Cada neurona tiene una relación de vecindad con las otras neuronas lo cual determina la tipología del mapa. Las tipologías más comunes son la rectangular y la hexagonal. El algoritmo es simliar al de los kmedios, es decir se toma un dato de entrada(también de dimensión d) y se encuentra la neurona cuyo vector de pesos sea mas similar al de datos de entrada. La diferencia es que en los mapas autoorganizados no solo se actualizan los valores de la neurona sino también la detodas las neuronas vecinas.

El mapa se entrena de manera iterativa. En cada iteración se elige de manera aleatoria un dato de entrada y se encuentra la neurona que contenga el vector de pesos mas similar a este, denominada BMU(best-matching unit):

$||x - m_c|| = min\{x - m_i\}$donde $||.||$ es la medida de similaridad utilizada.

Una vez encontrada la BMU se adapta el vector de pesos de esta siguiendo:

$m_i(t+1)| = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_i(t)]$

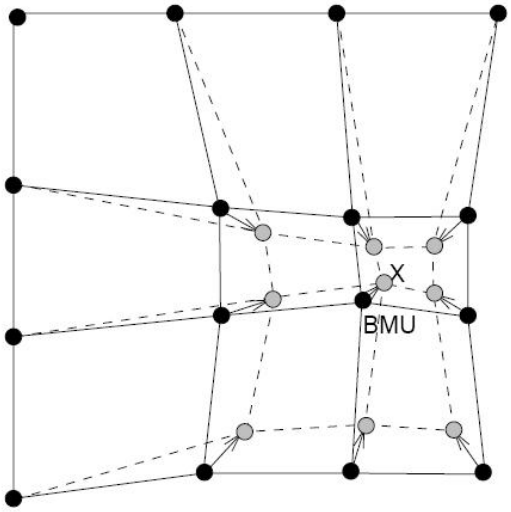


Figure 8: En el SOM se actualiza la BMU y sus vecinas

#### 3.5.3 Etiquetado de cada uno de los grupos

#### 3.5.4 Relación entre mensajes

#### 3.5.5 Clasificación de un nuevo mensaje entrante en una categoría.

### 3.6 Visualización

#### 3.6.1 Esquema de visualización de las categorías.

#### 3.6.2 Esquema de Navegación.

## 4 Experimentos y Resultados

### 4.1 Conjunto de Métricas utilizadas

### 4.2 Conjuntos de Datos

#### 4.2.1 Enron Dataset

### 4.3 Resultados

Para pruebas de la etapa de preprocesamiento se utilizó correos del conjutno de datos de enron. Este procesamiento incluye las etapas previamente descritas en este documento: Representación del mensaje(formato xml), Identificación de hilos, preprocesamiento del Asunto, Preprocesamiento del cuerpo del mensaje.

El número de mensajes utilizados fue de 50 correos electrónicos del buzón de un usuario. Los algoritmos utilizados para realizar el agrupamiento fue Agrupamiento Jerárquico utilizando distancia promedio.

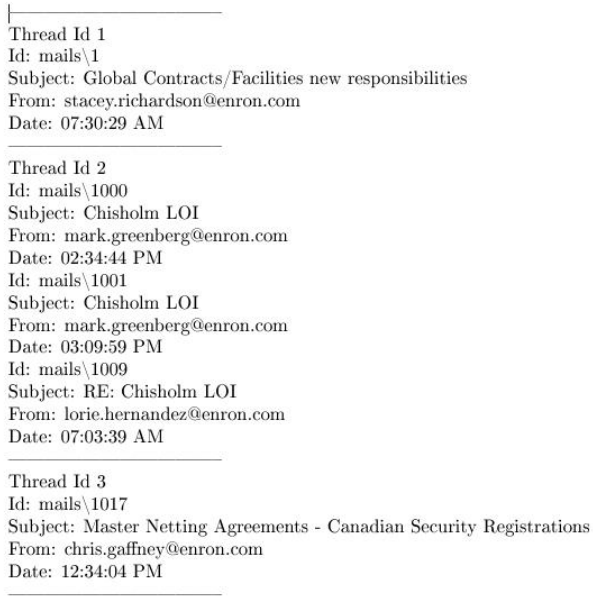


Figure 9: Ejemplo de identificación de hilos sobre el conjunto de datos de enron

Se realizaron 3 pruebas utilizando dos tipos de medidas de similaridad: Distancia euclidiana y distancia coseno. Una primera prueba se utilizó solo el asunto del mensaje, una segunda, fecha y asunto y una última, fecha, asunto y cuerpo.

Los resultados fueron comparados con los hilos de mensajes a los que pertenece cada uno.

## 4.4 Discusión

Los resultados de este trabajo indican que debido a la gran diversidad de información contenida en los mensajes de correo es difícil establecer una medida de similaridad adecuada para realizar el agrupamiento. La distancia coseno funciona bien cuando se incluye información del asunto y del cuerpo del mensaje, mas no con información de la fecha. La distancia euclidiana funciona bien con los campos fecha y asunto, pero no cuando se incluye información del cuerpo del mensaje. Por otro lado aunque los mapas autoorganizados permiten una buena visualización del agrupamiento , se deben incluir modificaciones en el algoritmo o en la representación de los datos para lograr un buen desempeño del sistema categorizador.

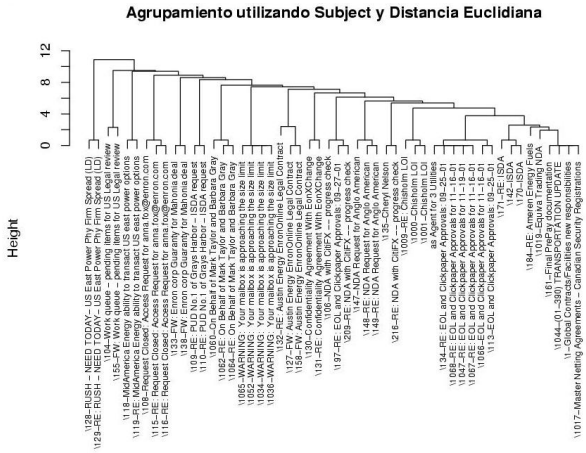


Figure 10: Agrupamiento Jerárquico utilizando asunto y distancia euclidiana

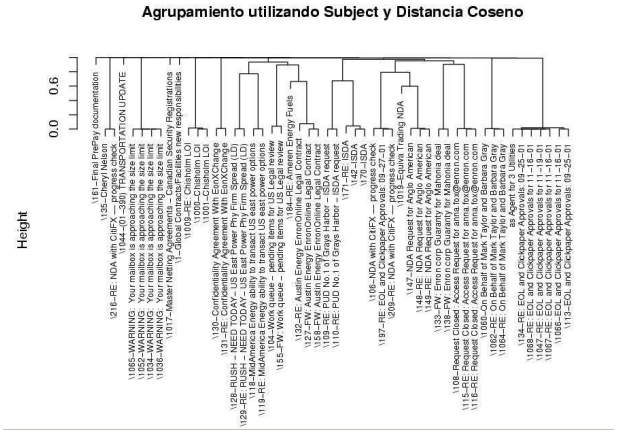


Figure 11: Agrupamiento Jerárquico utilizando asunto y distancia coseno

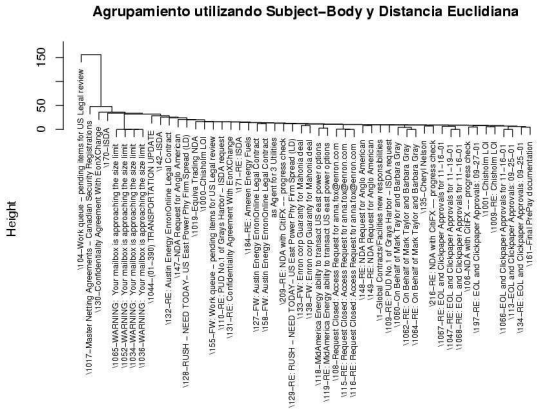


Figure 12: Agrupamiento Jerárquico utilizando asunto-cuerpo y distancia euclidiana

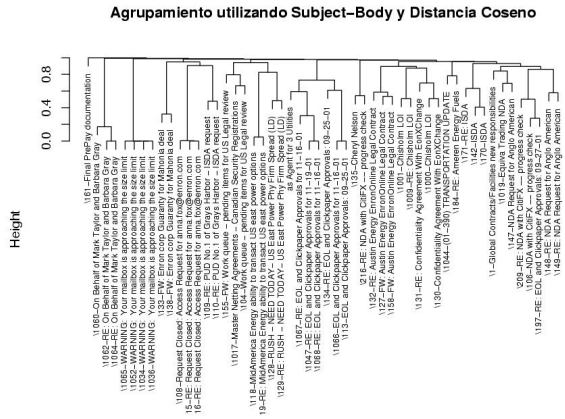


Figure 13: Agrupamiento Jerárquico utilizando asunto-cuerpo y distancia coseno

# 5 Conclusiones

# References


[1] M. Schroeder T. Wobber A. Birrell, S. Perl. The pachyderm e-mail system, 1997.

[2] Schreck Tobias A. Keim Daniel, Mansmann Florian. Mailsom - visual exploration of electronic mail archives using self-organizing maps. In *Second Conference on Email and Anti-Spam (CEAS 2005), Stanford University, Palo Alto, CA, USA*, 2005.

[3] Rana Kashif Ali. Ais and semantic query. Technical report, hp labs, 2004.

[4] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[5] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[6] Jake D. Brutlag and Christopher Meek. Challenges of the email domain for text classification. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 103–110, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[7] S.E. Robertson C.J. van Rijsbergen and M.F. Porter. New models in probabilistic information retrieval. Technical report, British Library Research and Development Report, no. 5587, 1980.

[8] W.W. Cohen. Learning to classify English text with ILP methods. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 124–143. IOS Press, 1996.

[9] Elisabeth Crawford, Judy Kay, and Eric McCreath. Automatic induction of rules for e-mail classification. In the Proceedings of the Sixth Australasian Document Computing Symposium, 2001, 2001.

[10] Elisabeth Crawford, Judy Kay, and Eric McCreath. Iems - the intelligent email sorter. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 83–90, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[11] Patrick Jon Crawford Elisabeth, Koprinska Irena. Phrases and feature selection in e-mail classification. In P. Bruza, A. Moffat, and A. Turpin, editors, *Proc. 9th Australasian Document Computing Symposium*, Melbourne, December 2004. Department of Computer Science and Software Engineering, The University of Melbourne.

[12] George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, March 2003.

[13] Aberer K. Giacoletto E. Automatic expansion of manual email classifications based on text analysis. In *2nd International Conference on Ontologies, Databases, and Applications of Semantics for large scale information systems (ODBASE)*, 2003.

[14] Carvalho Vitor R. Goodman Joshua. Implicit queries for email. In *Second Conference on Email and Anti-Spam (CEAS 2005), Stanford University, Palo Alto, CA, USA*, 2005.

[15] Otis Gospodnetic and Erik Hatcher. *Lucene in Action*. Manning Publications, 2004. GOS o 05:1 1.Ex.

[16] Paul Graham. A plan for spam. In *Reprinted in Paul Graham, Hackers and Painters, Big Ideas from the Computer Age, OReally, 2004*, 2002.

[17] Paul Graham. Better bayesian filtering. In *Proceedings of the 2003 Spam Conference*, Jan 2003.

[18] Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, /2000.

[19] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[20] Hooman Katirai. Filtering junk e-mail: A performance comparison between genetic programming & naïïve bayes. Technical report, University of Waterloo, 1999.

[21] Kushmerick Nicholas Khoussainov Rinat. Email task management: An iterative relational learning approach. In *Second Conference on Email and Anti-Spam (CEAS 2005), Stanford University, Palo Alto, CA, USA*, 2005.

[22] Abu-Hakima Suhayya Kriritchenko Svetlana, Matwin Stan. Email classification with temporal features.

[23] Nicholas Kushmerick and Tessa Lau. Automated email activity management: an unsupervised learning approach. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pages 67–74, New York, NY, USA, 2005. ACM Press.

[24] Krista Lagus, Samuel Kaski, and Teuvo Kohonen. Mining massive document collections by the websom method. *Inf. Sci.*, 163(1-3):135–156, 2004.

[25] David D. Lewis and K. A. Knowles. Threading electronic mail - a preliminary study. *Information Processing and Management*, 33(2):209–217, 1997.

[26] Xia Lin. Visualization for the document space. In *VIS '92: Proceedings of the 3rd conference on Visualization '92*, pages 274–281, Los Alamitos, CA, USA, 1992. IEEE Computer Society Press.

[27] Xia Lin, Dagobert Soergel, and Gary Marchionini. A self-organizing semantic map for information retrieval. In *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 262–269, New York, NY, USA, 1991. ACM Press.

[28] Wendy E. Mackay. More than just a communication system: diversity in the use of electronic mail. In *CSCW '88: Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, pages 344–353, New York, NY, USA, 1988. ACM Press.

[29] Giuseppe Manco, Elio Masciari, Massimo Ruffolo, and Andrea Tagarelli. Towards an adaptive mail classifier. In Italian Association for Artificial Intelligence Workshop Su Apprendimento Automatico: Metodi Ed Applicazioni, 2002.

[30] Kenricj Mock. An experimental framework for email categorization and management. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 392–393, New York, NY, USA, 2001. ACM Press.

[31] White Tony Oda Terri. Developing an immunity to spam. In *GECCO 2003. Genetic and Evolutionary Computation Conference, Chicago, IL, USA, July 12-16, 2003*, volume 2723, pages 231 – 242, 2003.

[32] Patrick Pantel and Dekang Lin. Spamcop: A spam classification & organization program. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.

[33] T. R. Payne and P. Edwards. Interface agents that learn: An investigation of learning issues in a mail agent interface. *Applied Artificial Intelligence*, 11(1):1–32, 1997.

[34] Gary Robinson. A statistical approach to the spam problem. *Linux J.*, 2003(107):3, 2003.

[35] Sam Scott and Stan Matwin. Feature engineering for text classification. In Ivan Bratko and Saso Dzeroski, editors, *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pages 379–388, Bled, SL, 1999. Morgan Kaufmann Publishers, San Francisco, US.

[36] A. Secker, A Freitas, and J. Timmis. Aisec: An artificial immune system for e-mail classification. In R. Sarker, R. Reynolds, H. Abbass, T. Kay-Chen, R. McKay, D Essam, and T. Gedeon, editors, *Proceedings of the Congress on Evolutionary Computation*, pages 131–139, Canberra. Australia, December 2003. IEEE.

[37] Pierre Senellart and Vincent D. Blondel. Automatic discovery of similar words. In Michael W. Berry, editor, *A Comprehensive Survey of Text Mining*. Springer-Verlag, August 2003.

[38] G. Venolia, L. Dabbish, J. Cadiz, and A. Gupta. Supporting email workflow. Microsoft Research Tech Report MSR-TR-2001-88, 2001.

[39] Stefan Wermter and Chihli Hung. Selforganizing classification on the Reuters news corpus. In *Proceedings of COLING-02, the 19th International Conference on Computational Linguistics*, Taipei, TW, 2002.

[40] Steve Whittaker, Victoria Bellotti, and Paul Moody. Revisiting and reinventing email. *HCI Special Issue on Email*, 20, 2005.