Categorización automática de correos electrónicos - Preprocesamiento

Camilo Enrique Rodríguez Torres
Universidad Nacional de Colombia, Facultad de Ingeniería, Departamento de Ingeniería de Sistemas
camilo255423@hotmail.com

***Abstract*— El siguiente documento describe la fase de procesamiento para la categorización automática de correos electrónicos.**

***Index Terms*— Clasificación, agrupamiento, clasificadores de textos, extracción de características, correos electrónicos**

I. Introducción

En los últimos años el volumen de correo electrónico que reciben los usuarios a diario ha crecido vertiginosamente[11]. 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 [17], [9] 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.

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 [3] 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 ([7], [15], [13]) , máquinas con vectores de soporte utilizadas en [1] por Brutlag y Meek , k-vecino más cercano KNN [14] por Payne , hasta técnicas relativamente recientes como sistemas inmunes artificiales utilizados para la detección de correo spam en [16] y en [12]. Los trabajos en categorización de correos bajo aprendizaje no supervisado, como se mencionó anteriormente, se han centrando especialmente en organización jerárquica de folderes o carpetas. Cabe destacar [10] realizado por Manco et al. utilizando clustering particional y jearquico para realizar el agrupamiento. Las características son extraídas a partir de las diferentes partes del mensaje de correo(Remitente destinatarios, hora/fecha, Asunto, Archivos adjuntos y contenido del mensaje); El trabajo de Giacolleto y Aberer en [4] donde se propone 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 correos electrónicos. En este trabajo se utiliza el método de k-medios bisectivo para realizar el proceso de agrupamiento; El trabajo de Khuossainov [8] utilizando clustering jerárquico aglomerativo para reconocer tareas y procesos del usario(por ejemplo compras en línea) a partir del contenido de los mensajes de correo del usuario.

Este documento presenta la primera fase para el desarrollo del sistema categorizador de correos electrónicos. Este primera fase involucra el preprocesamiento de los correos para crear una representación adecuada de los mensajes y poder realizar el proceso de categorización. El documento esta estructurado de la siguiente manera: La sección 3 explica la arquitectura del sistema propuesto y describe el diseño para la fase de procesamiento, el cual incluye, representación de los mensajes en formato xml, identifiación de hilos y procesamiento del cuerpo de los mensajes. El capítulo 4 la selección y extracción de características de los mensajes y el capítulo 5 describe brevemente las pruebas relizadas hasta el momento con este sistema.

II. Categorización automática 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[5]:

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 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, 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.

III. Arquitectura del sistema propuesto

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

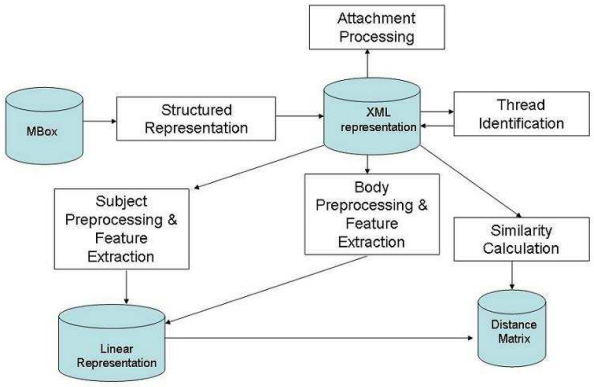

Fig. 1. Arquitectura del Sistema

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.

A. Campos importantes en los mensajes de correo

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.

B. Representación de los mensajes en Xml

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.

C. Identificación de Hilos

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[9]:

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 [9].

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 determinar 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.

D. 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[6]:

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

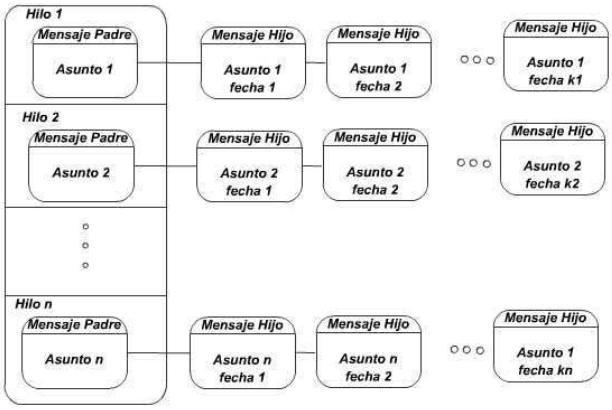

Fig. 2. Representación vectorial de los hilos

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[2].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” [6] para realizar el procesamiento anteriormente descrito.

IV. Selección y Extracción de características del cuerpo del mensaje

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.

Thread Id 1
Id: mails\1
Subject: Global Contracts/Facilities new responsibilities
From: stacey.richardson@enron.com
Date: 07:30:29 AM

Thread Id 2
Id: mails\1000
Subject: Chisholm LOI
From: mark.greenberg@enron.com
Date: 02:34:44 PM
Id: mails\1001
Subject: Chisholm LOI
From: mark.greenberg@enron.com
Date: 03:09:59 PM
Id: mails\1009
Subject: RE: Chisholm LOI
From: lorie.hernandez@enron.com
Date: 07:03:39 AM

Thread Id 3
Id: mails\1017
Subject: Master Netting Agreements - Canadian Security Registrations
From: chris.gaffney@enron.com
Date: 12:34:04 PM

Fig. 3. Ejemplo de identificación de hilos sobre el conjunto de datos de enron

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.

V. RESULTADOS PRELIMINARES

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.

Se realizaron 3 pruebas para cada algoritmo: Una primera utilizando solo el asunto del mensaje, una segunda utilizando fecha y asunto y una última utilizando fecha, asunto y cuerpo.

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

VI. DISCUSIÓN

En el agrupamiento jerárquico, es claro que el uso de la información del cuerpo del mensaje no ayuda en el agrupamiento de los mensajes. Esto se debe a que en muchos casos el texto contenido en el cuerpo es casi nula y no aporta

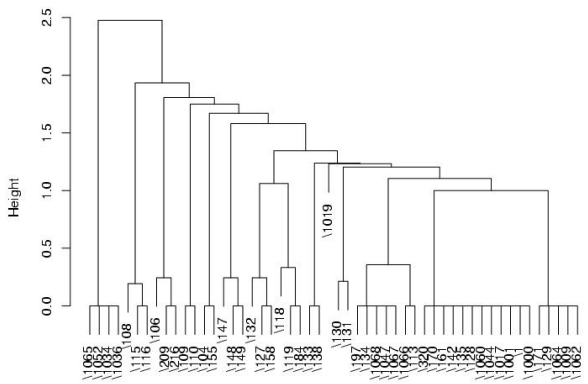

Fig. 4. Agrupamiento Jerárquico utilizando asunto

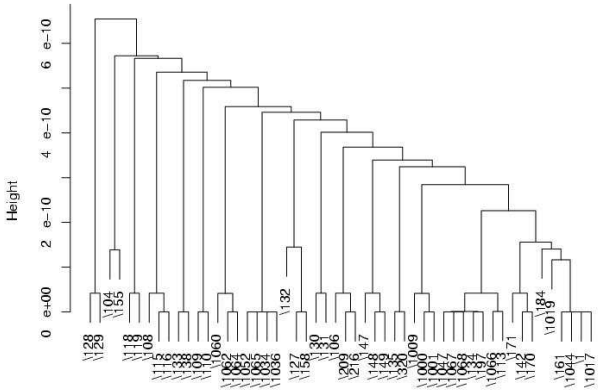

Fig. 5. Agrupamiento Jerárquico utilizando fecha y asunto

ningún información discriminante para el agrupamiento, y por el contrario puede introducir ruido.

La inclusión de la fecha incrementa el desempeño del agrupamiento, identificando grupos de mensajes mas pequeños y mas similares entre si.

VII. TRABAJO FUTURO

Teniendo ya una arquitectura para el procesamiento de los mensajes de correo, el trabajo futuro incluye definir por unlado una representación lineal de los mensajes que integre toda esta información extraida, hilos, asunto, cuerpo del mensaje y por otro las medidas de similaridad y la matriz de distancia entre mensajes para poder abordar la fase de categorización como tal.

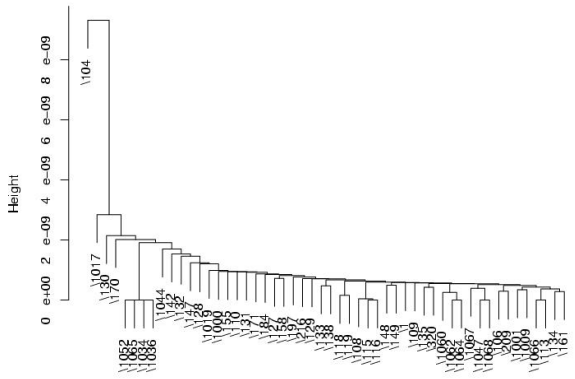

Fig. 6. Agrupamiento Jerárquico utilizando fecha, asunto, cuerpo

Thread Id: 1: mails\1,
Thread Id: 2: mails\1000,mails\1001,mails\1009,
Thread Id: 3: mails\1017,
Thread Id: 4: mails\1019,
Thread Id: 5: mails\1034,mails\1036,mails\1052,mails\1065,
Thread Id: 6: mails\104,mails\155,
Thread Id: 7: mails\1044,
Thread Id: 8: mails\1047,
Thread Id: 9: mails\106,mails\209,mails\216,
Thread Id: 10: mails\1060,mails\1062,mails\1064,
Thread Id: 11: mails\1066,mails\1067,mails\1068,
Thread Id: 12: mails\108,mails\115,mails\116,
Thread Id: 13: mails\109,mails\110,
Thread Id: 14: mails\113,mails\134,
Thread Id: 15: mails\118,mails\119,
Thread Id: 16: mails\127,mails\132,mails\158,
Thread Id: 17: mails\128,mails\129,
Thread Id: 18: mails\130,mails\131,
Thread Id: 19: mails\133,mails\138,
Thread Id: 20: mails\135,mails\320,
Thread Id: 21: mails\142,mails\170,mails\171,
Thread Id: 22: mails\147,mails\148,mails\149,
Thread Id: 23: mails\161,
Thread Id: 24: mails\184,
Thread Id: 25: mails\197,

Fig. 7. Identificación de hilos para 50 mensajes del conjunto de datos de enron

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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