Text Mining Aplicado a la Clasificación y Distribución Automática de Correo Electrónico y Detección de Correo SPAM

Pedro Fabricio Echeverría Briones [1]
Zoila Verónica Altamirano Valarezo[2]
Alvaro Badir Pinto Astudillo[3]
Johanna del Carmen Sánchez Guerrero [4]
[1] Director de Tópico, MSIG, Profesor de ESPOL; e-mail: pechever@uniplex.com.ec
[2] Ingeniera en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: valtamir@.fiec.espol.edu.ec
[3] Ingeniera en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: apinto@fiec.espol.edu.ec
[4] Ingeniero en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: jdcsanch@fiec.espol.edu.ec

Resumen

El Text Mining es la más reciente área de investigación del procesamiento de textos. Ella se define como el proceso de descubrimiento de patrones nuevos conocimientos en una colección de textos, es decir, El Text Mining es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos [2]. *En el desarrollo de nuestro proyecto aplicamos Text Mining para la creación de un Anti-SPAM, el mismo que esta diseñado para uso del Algoritmo de Naive Bayes, y los algoritmos de K-mean y reglas de asociación. Teniendo como objetivo principal la detección de correos basuras o SPAM que reciba un usuario o que posea una dirección de correo electrónico.*

Palabras Claves: Text Mining, SPAM, Algoritmo de Naive Bayes.

Abstract

The Text Mining is the most recent area in investigation of the prosecution of texts. She is defined as the process of discovery of patterns and new knowledge in a collection of texts, that is to say, The Text Mining is the process in charge of the discovery of knowledge that they didn't exist explicitly in any text of the collection, but that they arise of relating the content of several of it. In our development project we apply the Text Mining for the creation of an Anti-SPAM, the same one that this designed for use of the Algorithm of Naive Bayes, and the algorithms of they K-mean and association rules. Having as main objective to detect the mail garbage or SPAM that a user or possesses an electronic mail address receives.

1. Introducción

En la actualidad el correo electrónico es un medio de comunicación eficiente y cada vez más popular. Al ser extremadamente económico y fácil de enviar, es también un medio para el comercio electrónico. Desafortunadamente, esto ha causado que vendedores de todo tipo bombardeen los buzones de correo con mensajes no solicitados y no deseados, conocidos como correo basura o SPAM. La consecuencia de esto es la pérdida de tiempo de los lectores de correo electrónico, el coste de recibir este tipo de mensajes (a veces ofensivos) y el riesgo de llenar el espacio de almacenamiento del servidor (o el buzón). Debido a esto, se propuso la siguiente solución a dicho planteamiento.

Dichos correos basura o SPAM, serán detectados mediante la implementación de un filtro anti-SPAM basado en el algoritmo de Naive Bayes, en el cual se obtendrá los datos desde una base de datos conteniendo palabras SPAM y palabras no SPAM, las cuales habrán sido obtenidas de mensajes SPAM y mensajes legítimos respectivamente. Posteriormente realizamos el análisis del contenido del correo haciendo uso de la minería de texto, para luego determinar si dicho contenido es o no un SPAM.

Además del filtro anteriormente mencionado se desarrollo un interfaz como ayuda para el administrador del correo, en la que le permitirá conocer los usuarios que reciben correo basura o SPAM.

2. Definiciones, Métodos y Materiales

2.1. Definiciones

2.1.1 ¿QUE ES EL SPAM?

El SPAM es el correo electrónico no solicitado, normalmente con contenido publicitario, que se envía de forma masiva.

2.1.2 CARACTERÍSTICAS

Algunas de las características más comunes que presentan este tipo de mensajes de correo electrónico son:

- La dirección que aparece como remitente del mensaje no resulta conocida para el usuario, y es habitual que esté falseada.
- El mensaje no suele tener dirección para reenviar.
- Presentan un asunto llamativo.
- El contenido es publicitario: anuncios de sitios web, fórmulas para ganar dinero fácilmente, productos milagro, ofertas inmobiliarias, o simplemente listados de productos en venta en promoción.

Aunque el método de distribución más habitual de este tipo de publicidad es el correo electrónico, existen diversas variantes, cada cual con su propio nombre asociado en función de su canal de distribución [6]:

- ***SPAM:*** Enviado a través del correo electrónico.
- ***Spim:*** Específico para aplicaciones de tipo Mensajería Instantánea (MSN Messenger, Yahoo Messenger, etc).
- ***Spit:*** SPAM sobre telefonía IP. La telefonía IP consiste en la utilización de Internet como medio de transmisión para realizar llamadas telefónicas.
- ***SPAM SMS:*** SPAM destinado a enviarse a dispositivos móviles mediante SMS (Short Message Service).

2.1.3. TIPOS DE SPAM

En todo el mundo, el SPAM tiende a anunciar cierta gama de bienes y servicios, sin que importe el lenguaje y la geografía. El SPAM pertenece a las siguientes categorías:

- Pornografía
- Tecnologías informáticas
- Finanzas personales
- Educación / entrenamiento
- Otros.

***Pornografía*:** Esta categoría incluye ofertas de productos diseñados para aumentar o mejorar la potencia sexual, enlaces a sitios pornográficos o anuncios de pornografía. Ejemplo:

> Tema: Una ayuda muy barata para su erección :-)
>
> ¡Buenos días!
> ¡Le ofrecemos la ***Viagra*** más barata del mundo! . La puede comprar en: {ENLACE}
> Sinceramente,
> Liza Stokes

Figura 1. SPAM Pornográfico

***Tecnologías informáticas*:** Esta categoría incluye ofertas de hardware y software a bajo precio, como también servicios para los usuarios de sitios Web: hospedaje (hosting), registro de dominios, optimización de sitios Web y así por el estilo. Ejemplo

Tema: Grandes ahorros en software OEM. Todas las mejores marcas a tu disposición.

¿Buscas software de alta calidad a bajo precio? Nosotros tenemos lo que necesitas
Windows XP Professional 2002 .... $50
Adobe Photoshop 7.0 .... $60
Microsoft Office XP Professional 2002 .... $60
Corel Draw Graphics Suite 11 .... $60
y mucho más...

Figura 2. SPAM de Tecnologías informáticas

Finanzas personales**:** El SPAM de esta categoría ofrece seguros, servicios de reducción de deudas, préstamos con bajos intereses. Ejemplo:

Tema: Los prestamistas compiten entre sí...tú ganas.

Reduce tus pagos de hipoteca. ¡Los intereses están creciendo! Dale a tu familia la libertad financiera que se merece.
Refinancia hoy y ahorra.
*Rápido y fácil
*CONFIDENCIAL
*Cientos de prestamistas
*100% gratis
*Obtén las mejores tasas
¡Solicítalo hoy! {ENLACE}
Cualquier crédito será aceptado. Para borrar tu nombre de nuestra base de datos pulsa aquí
Gracias.

Llama 1-800-279-7310
o escríbenos a: 1700 E. Elliot Rd. STE3-C4 Tempe, AZ. 85283

Figura 3. SPAM de Finanzas Personales

Educación**:** Esta categoría incluye ofertas de seminarios, entrenamientos en línea. Ejemplo:

Tema: Obtén un diploma de licenciatura o doctorado desde tu casa.

Llama a {número de teléfono} para averiguar sobre nuestros programas de graduación.
Si estás buscando un grado de licenciatura, doctorado o MBA Podemos darte credenciales completamente verificables para que hagas carrera. BACK ON TRACK!
Sin exámenes ni tesis. Llama: {Teléfono} Call: {Phone Num.}. Nos disculpamos si no querías recibir este mensaje. Para darte de baja de nuestra lista, por favor llama {Teléfono}

Figura 4. SPAM de Educación

2.2. Métodos

El filtro bayesiano es técnica utilizada en la detección de SPAM, estos filtros se basan en estadísticas de palabras que aparecen en los correos no deseados.

Generalmente trabajan con grupos de dos tipos de correos, los legítimos y los SPAM, para cada palabra en estos correos, los filtros calculan las probabilidades de SPAM basadas en la proporción de ocurrencias de los mismos, por ejemplo, la palabra "garantizado" aparece un 99% de las veces en correos SPAM que en correos legítimos y la palabra deducir aparece con un porcentaje menor pues no es usada en los correos no deseados.

2.2.1 Método de Naive Bayes

Utilizamos este método para realizar el filtro Anti-SPAM, como ya hemos mencionados la necesidad de crear este filtro se debe a la gran cantidad de correos basuras o SPAM que recibimos en nuestros correos.

Método a utilizar

Se escogerá una base de datos o muestra de entrenamiento conteniendo palabras SPAM y palabras no SPAM cada grupo de palabra fueron seleccionadas de correo basura y mensajes legítimos respectivamente, Todos los correos de los usuarios son extraídos desde el servidor de correo para posteriormente realizar el respectivo análisis, para esto hacemos uso de las fórmulas que a continuación se expondremos y de acuerdo al resultado obtenido de estas, se podrá definir si el contenido de los correos analizados es un correo SPAM o no.

- Teorema de Bayes
    - P(B|A) = (P(A|B) * P(B)) / P(A)
- Teorema de Bayes para filtro de correo
    - Evento "SPAM" = mensaje es SPAM
    - Evento "palabras" = mensaje contiene palabras malas
- P(SPAM|palabras)=P(palabras|SPAM) * P(SPAM | P(palabras)
- *Probabilidad de que un mensaje es SPAM, dado que contiene palabras malas.*

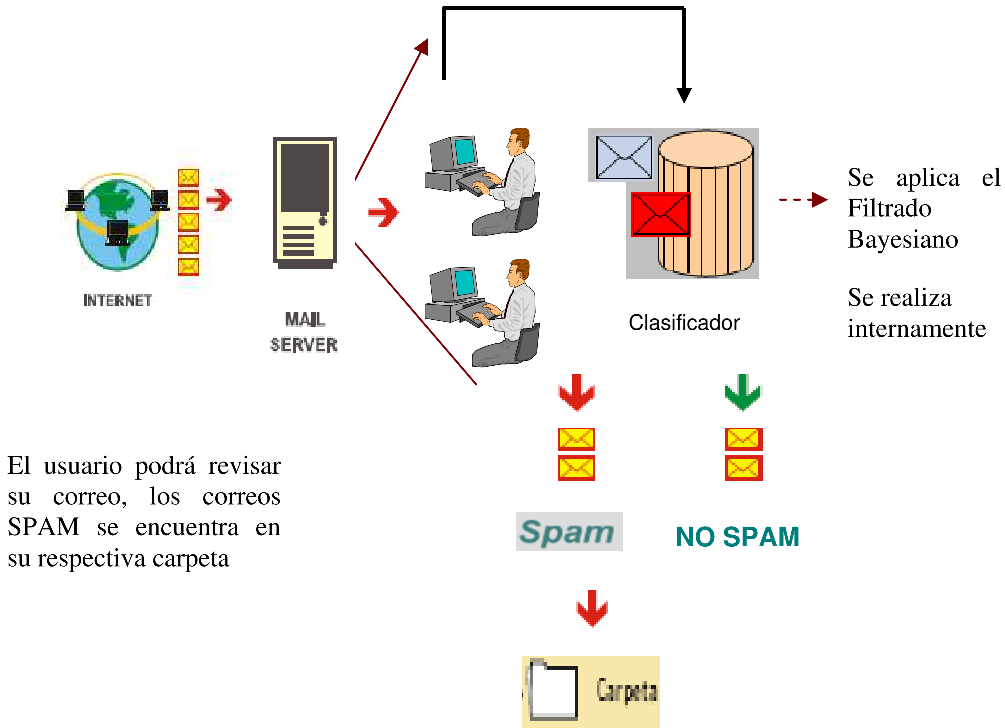

Figura 5. Descripción gráfica del filtrado bayesiano

2.2.2. Método de Regla de Asociación

Utilizamos este método para realizar los Contactos Indirectos, a continuación se explica con detalle lo planteado.

Origen

Este proceso surge de la necesidad de los diferentes usuarios de conocer los tipos de contactos que no pertenecen a su lista de favoritos, pero de una u otra manera están relacionados con sus contactos mas afines ya sea porque se encuentran en sus listas de contactos o estén relacionados por los correos

Esto es con el fin de determinar si estos contactos al que llamamos "C*ontactos indirectos*" pueden pertenecer a nuestra lista de contactos personales o a un determinado grupo el cual podríamos tener personalizados de acuerdo a los contactos de nuestra conveniencia.

Determinar si estos contactos indirectos están repercutiendo en nuestro buzón de correos, enviando tipos de correos denominados SPAM par lo cual podríamos considerarlos como un usuario cuyos correos son de no importancia y de esta manera poder clasificar a este usuario.

Método a utilizar

Se realiza una representación del algoritmo de asociación para verificar estos tipos de usuarios.

Creamos una matriz **nxn** donde n esta representado por todos contactos favoritos agregados por los diferentes usuarios. Esta matriz de contactos estarán con datos de 1's y 0's de acuerdo a la relación que se tenga entre ellos, 1 para si estos contactos tienen relación directa de acuerdo si esta agregado a su lista de contactos favoritos y cero en el caso de que no tengan relación alguna.

	c1	c2	c3	c4	c5	...	cn
c1	0	1	0	0	1	...	1
c2	1	0	1	0	1	...	0
c3	0	0	0	0	1	...	0
c4	0	0	1	0	1	...	1
c5	1	0	0	1	0	...	0
:	:	:	:	:	:	...	1
cn	1	1	0	0	0	...	0

Figura 6. Gráfico representativo de los contactos favoritos

Esta relación se la realiza de acuerdo a los contactos favoritos de todos los usuarios con respecto a los míos. Esta relación tiene una cobertura, el cual esta dado para un contacto indirecto determinado como relación directa de uno o mas de mis contactos favoritos. Esta cobertura me servirá para determinar la confianza (en %) que podría tener para determinar si este usuario o no puede ser parte de mis contactos de acuerdos a las relaciones encontradas.

Estos parámetros serán personalizados, lo que indicaría que el usuario puede solo querer ver los detalles de aquellos contactos indirectos cuya cobertura sea específica y con grado de confianza determinado.

2.3. Materiales

Nuestro sistema será implementado con las herramientas actuales y adaptables a los requerimientos exigidos por el mismo, permitiendo con esto tener una gran eficiencia, un fácil uso y a su vez tenga una mayor escalabilidad. Las herramientas a utilizar son mostradas en las siguientes tablas y se las detallan en los siguientes puntos a tratar:

Software	Descripción
Windows 2003 Server	Sistema Operativo
Exchange Server 2003	Contenedor de Correos

Tabla 1. Software recomendado para el desarrollo del Servidor de Correo

Software	Descripción
Windows XP Professional	Sistema Operativo
Visual C# Net	Ambiente de Desarrollo
SQL Server 2005	Motor de Base de Datos
Crystal Reports	Generador de reportes

Tabla 2. Software recomendado para el desarrollo del Cliente

2.3.1. Plataforma

Se eligió trabajar con la plataforma *Microsoft Windows XP Professional* ya que es uno de los entornos de escritorio más usados a nivel Personal, Empresarial y Corporativos.

De acuerdo a nuestro enfoque el cual está dado a los niveles antes mencionados, el sistema será más adaptable y comprensible para dichos usuarios ya que es una herramienta de uso diario para ellos.

Otra plataforma a utilizar es *.NET*, dada que es desarrollado por el mismo Microsoft lo cual permitirá un mejor acoplamiento con las demás herramientas a utilizar haciendo así nuestro sistema más estable y compatible con los cambios tecnológicos que se puedan dar en un futuro.

2.3.2. Herramientas de Desarrollo

VISUAL C# NET: Es un lenguaje de propósito general orientado a objetos creado por Microsoft para su nueva plataforma, además ofrece una interfaz común para trabajar de manera cómoda y visual con cualquiera de los lenguajes de la plataforma .NET

SQL SERVER 2005: Además de las ventajas de reducir los tiempos de inactividad de las aplicaciones, mayor escalabilidad y rendimiento y estrictos controles de seguridad, el SQL Server 2005 comprenderá también mejoras significativas de gestión de la información empresarial en los siguientes aspectos:

- Disponibilidad
- Escalabilidad
- Seguridad
- Facilidad de gestión
- Interoperabilidad

3. Resultados obtenidos

- Como resultado se obtuvo una herramienta para podernos “Defender” de los mensajes basura de algunas maneras, aunque no hay métodos infalibles. Por un lado, con unas ciertas normas que debemos seguir ante los mensajes SPAM que recibamos. Por otro, filtrando los mensajes que nos llegan para intentar recibir la menor cantidad posible de este correo no solicitado.

- En la parte de que proporcionamos a los usuarios conocer los tipos contactos que no pertenecen a su lista de favorito, Se les permite determinar si estos contactos al que llamamos “C*ontactos indirectos*” pueden pertenecer a su lista de contactos personales o a un determinado grupo el cual podríamos tener personalizados de acuerdo a los contactos de nuestra conveniencia, o si estos contactos indirectos están repercutiendo en nuestro buzón de correos, enviando tipos de correos denominados SPAM par lo cual podríamos considerarlos como un usuario cuyos correos son de no importancia y de esta manera poder clasificar a este usuario.

- Finalmente en la interfaz que proporcionamos al Administrador de correo, en la que podrá encontrar una clasificación de usuario por el espacio de disco que ocupan sus correos y por el número de correos SPAM que reciben los mismos, se obtuvo como resultado diferentes grupos de usuarios entre los que ocupan mayor espacio de disco y así mismo lo que mayor correo basura reciben en sus correo.

4. Conclusiones

- Se optimizó la capacidad de clasificar si el contenido de los correos recibidos son o no SPAM, permitiendo así que el espacio de disco no se ocupado por correos basuras.

- Se permitió que el Administrador de correo pueda tener un mejor conocimiento de cuales son los usuarios que mayores correos SPAM, recibían o que ocupaban demasiado espacio de disco.

- También proporcionamos al usuario conocer los tipos de contactos que no pertenecen a su lista de favoritos, pero de una u otra manera está relacionada con sus contactos mas afines ya sea porque se encuentran en sus listas de contactos o estén relacionados por los correos.

7. Referencias

[1] Tesis de Ingeniería en Informática, http://www.fi.uba.ar/laboratorios/lsi/servente-tesisingenieriainformatica.pdf

[2] Text Mining, http://www.lsi.us.es/~ferrer/mdoc/MineriaTexto_md01.pdf

[3] Adriaans, P (1996). Data mining. Addison-Wesley

[4] Pyle. D. (1999). Data preparation for data mining. Morgan Kaufman

[5] Berry, M. (1997) Data mining techniques. Wiley

[6] Filtro Anti-Spam http://www.vsantivirus.com/mm-ia-antispam.htm