

ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE INGENIERÍA EN ELECTRICIDAD Y COMPUTACIÓN

**“TEXT MINING APLICADO A LA CLASIFICACIÓN Y DISTRIBUCIÓN
AUTOMÁTICA DE CORREO ELECTRÓNICO Y DETECCIÓN DE CORREO
SPAM”**

TESIS DE GRADO

Previa a la Obtención del Título de:

INGENIERO EN COMPUTACIÓN

ESPECIALIZACIÓN SISTEMAS TECNOLÓGICOS

Presentada por:

ALTAMIRANO VALAREZO ZOILA VERÓNICA

PINTO ASTUDILLO ALVARO BADIR

SÁNCHEZ GUERRERO JOHANNA DEL CARMEN

GUAYAQUIL – ECUADOR

2007

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este proyecto de graduación, nos corresponde exclusivamente y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de exámenes y títulos profesionales de la ESPOL)


ALTAMIRANO VERÓNICA


PINTO ALVARO


SÁNCHEZ JOHANNA

TRIBUNAL DE GRADUACIÓN



ING. HOLGER CEVALLOS
SUB DECANO DE LA FIEC



ING. FABRICIO ECHEVERRÍA
DIRECTOR DEL TÓPICO



ING. CARMEN VACA
VOCAL PRINCIPAL

ESCUELA SUPERIOR POLITECNICA
FACULTAD DE INGENIERIA
BIBLIOTECA
INV. No. CMPT-ST-64-1

DEDICATORIA

Dedicamos este trabajo a todos aquellos que
creyeron en nosotros y nos apoyaron en todo
momento

RESUMEN

En la actualidad el correo electrónico es un medio de comunicación cada vez más popular. Al ser extremadamente económico y fácil de usar, 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 al problema del SPAM. Los correos basura o SPAM, serán detectados mediante la implementación de un filtro anti-SPAM basado en el algoritmo de Naive Bayes, el filtro obtendrá los datos de una base de datos que contiene palabras SPAM y palabras no SPAM, las cuales han sido obtenidas de mensajes SPAM y mensajes legítimos respectivamente. Posteriormente se realiza 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 implementó como ayuda para el usuario determinar cuales de los contactos indirectos podrían pertenecer a

su lista de contactos, 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 "Contactos 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 para lo cual podríamos considerarlos como un usuario cuyos correos son de no importancia y de esta manera poder clasificar a este usuario.

Finalmente, también se desarrolló un interfaz como ayuda para el administrador del correo, en la que le permitirá conocer con mayor exactitud los usuarios que reciben correo basura o SPAM.

3.2.1 Lenguaje de programación.....	53
3.2.2 Algoritmos relacionados con la solución del problema.....	53

4 CIRCULO VIRTUOSO DE LA MINERÍA DE DATOS	
4.1 ¿Qué es el círculo virtuoso de la minería de datos? - Fases.....	58
4.1.1 Definición del problema.....	59
4.1.2 Exploración de los datos.....	60
4.1.3 Preparación de los datos.....	62
4.1.4 Modelación.....	63
4.1.5 Evaluación del modelo.....	68
4.1.6 Despliegue de los modelos y resultados.....	69

5 ANÁLISIS DE FACTIBILIDAD DE COSTO	
5.1 Descripción de los servicios del sistema.....	70
5.2 Costos del sistema.....	71
5.2.1 Costos de desarrollo.....	71
5.2.2 Costos de implementación.....	71
5.2 Análisis de viabilidad.....	72
5.2.1 Análisis costo - beneficio.....	72

CONCLUSIONES Y RECOMENDACIONES	76
---------------------------------------	----

ANEXOS	
Anexo A.....	Manual del usuario
Anexo B.....	Manual técnico.

ÍNDICE DE GRÁFICOS

Pág.

Figura 1.1	Proceso de minería de texto.....	3
Figura 1.2	SPAM pornográfico.....	9
Figura 1.3	SPAM de tecnologías informáticas	10
Figura 1.4	SPAM de salud y medicina	10
Figura 1.5	SPAM de finanzas personales.....	11
Figura 1.6	SPAM de educación / entrenamiento.....	11
Figura 3.1	Diagrama de contexto de casos de uso.....	21
Figura 3.2	Escenario 1.1 Ingreso exitoso de “usuario” y “contraseña”	35
Figura 3.3	Escenario 1.2 Ingreso no exitoso por “usuario” incorrecto Escenario 1.3 Ingreso no exitoso por contraseña	35
Figura 3.4	incorrecta.....	36
Figura 3.5	Escenario 1.4 Ingreso no exitoso por no haber ingresado los datos que son exigidos.....	36
Figura 3.6	Escenario 1.5 Ingreso no exitoso por fallas técnicas correspondientes al ingreso de “usuario” o “contraseña”	37
Figura 3.7	Escenario 2.1 Clasificación exitosa por espacio en el buzón	37
Figura 3.8	Escenario 3.1 Clasificación exitosa por uso de correo.....	38
Figura 3.9	Escenario 4.1 Ingreso exitoso de “usuario” y “contraseña”	38
Figura 3.10	Escenario 4.2 Ingreso no exitoso por “usuario” incorrecto	39
Figura 3.11	Escenario 4.3 Ingreso no exitoso por “contraseña” incorrecta	39
Figura 3.12	Escenario 4.4 Ingreso no exitoso por no haber ingresado los datos que son exigidos	40
Figura 3.13	Escenario 4.5 Ingreso no exitoso por fallas técnicas correspondientes al ingreso del “usuario” o “contraseña”	40
Figura 3.14	Escenario 5.1 Organización exitosa de los correos electrónicos por asunto.....	41
Figura 3.15	Escenario 5.2 Organización no exitosa de los correos electrónicos por asunto.....	41
Figura 3.16	Escenario 5.3 Organización exitosa de los correos electrónicos por remitente o destinatario.....	42
Figura 3.17	Escenario 5.4 Organización no exitosa de los correos electrónicos por remitente o destinatario.....	42
Figura 3.18	Escenario 5.5 Organización exitosa de los correos electrónicos por fecha de envío o recepción.....	43
Figura 3.19	Escenario 5.6 Organización no exitosa de los correos electrónicos por fecha de envío o recepción.....	43
Figura 3.20	Escenario 5.7 Organización exitosa por prioridad.....	44

Figura 3.21	Escenario 6.1 Detección exitosa de correo SPAM....	44
Figura 3.22	Escenario 6.2 Detección no exitosa de correo SPAM.....	45
Figura 3.23	Escenario 7.1 Determinación exitosa de correo SPAM....	45
Figura 3.24	Escenario 8.1 Determinación exitosa de contactos indirectos.....	46
Figura 3.25	Diagrama entidad - relación.....	47
Figura 3.26	Diagrama de los componentes del sistema de los módulos de administración y de usuario.....	48
Figura 3.27	Diseño de flujo de información.....	50
Figura 3.28	Pantalla principal.....	51
Figura 3.29	Pantalla de entrada a cuenta.....	51
Figura 3.30	Pantalla de bandeja de entrada.....	52
Figura 3.31	Descripción teórica del filtrado bayesiano.....	55
Figura 3.32	Descripción gráfica del filtrado bayesiano.....	55
Figura 3.33	Gráfico representativo de los contactos favoritos.....	57
Figura 4.1	Diagrama de proceso de la minería de datos.....	59
Figura 4.2	Ejemplo de la descripción de los datos de la tabla correo.....	61
Figura 4.3	Ejemplo de la descripción de los datos de la tabla contactos.....	62
Figura 4.4	Diseño del filtro Anti – SPAM.....	64
Figura 5.1	Porcentaje de SPAM en el servicio de correo electrónico.....	73

ÍNDICE DE TABLAS

	Pág.
Tabla 1 Estado del arte de la minería de texto.....	3
Tabla 2 Software utilizado para el desarrollo del proyecto_servidor.....	16
Tabla 3 Software utilizado para el desarrollo del proyecto_cliente.....	17
Tabla 4 Costos para el desarrollo del sistema.....	71
Tabla 5 Costos de implantación del sistema.....	71
Tabla 6 Listas de precios de servidores.....	74

1.- FUNDAMENTOS TEÓRICOS

El propósito de este capítulo es que el lector tenga una definición clara de los términos o conceptos que hemos utilizados para el desarrollo de este proyecto.

1.1 MINERÍA DE TEXTO (TEXT MINING)

En la actualidad el 80% de información que posee una empresa ó institución sobre sus clientes está contenida en textos es decir, almacenados en forma textual no estructurada entre los que tenemos: informes, correos electrónicos etc. Los avances recientes en el descubrimiento de nuevo conocimiento y su gestión incluyen la aplicación de técnicas de minería de datos para encontrar conocimiento significativo a partir de datos textuales sin estructura.

“La esencia de los métodos de la minería de datos aplicados a los datos numéricos, puede también ser aplicada a datos de texto. Sin duda, este campo de estudio es muy vasto, por lo que técnicas como la categorización del documento, el procesamiento de lenguaje natural, la extracción y recuperación de la información o el aprendizaje automático, entre otras, apoyan a la **minería de texto** (TEXT MINING)”[1].

La extracción de conocimiento a partir de datos de texto tiene como principal objetivo descubrir patrones válidos, interesantes y comprensibles; para lograr este objetivo se hace uso de las diferentes técnicas de minería de datos que en el desarrollo de este capítulo han sido definidas; entre ellas se encuentran las que se utilizó para el desarrollo de este proyecto.

“En ocasiones se confunde a la **minería de texto** (TEXT MINING) con la *recuperación de la información* (Information Retrieval o IR). Ésta consiste en la recuperación automática de documentos relevantes mediante indexaciones de

textos, clasificación, categorización, etc. En cambio, el TEXT MINING se refiere a examinar una colección de documentos y descubrir información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo.” [1]

En conclusión la **minería de texto** (TEXT MINING) es la encargada de detectar o encontrar patrones no triviales, escondido, anónimo y potencialmente útil e incluso información sobre el conocimiento almacenado en las bases de datos textuales (conjunto de datos textuales) no estructuradas.

1.2 PROCESO DE LA MINERÍA DE TEXTO (TEXT MINING)

La **minería de texto** es la encargada de descubrir el conocimiento que no existía formalmente en ningún texto de la colección (grandes colecciones de documentos no estructurados). “Para procesar el texto se debe de realizar dos etapas las cuales son: el de pre-procesamiento y la del descubrimiento.

- Una etapa de pre-procesamiento: En esta primera etapa, los textos se transforman a algún tipo de representación estructurada o semi-estructurada que facilite su posterior análisis.
- Una etapa de descubrimiento: En la segunda etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones o nuevos conocimientos.

Dependiendo del tipo de métodos usados en la etapa de pre - procesamiento es el tipo de representación del contenido de los textos construido; y dependiendo de esta representación, es el tipo de patrones descubiertos.

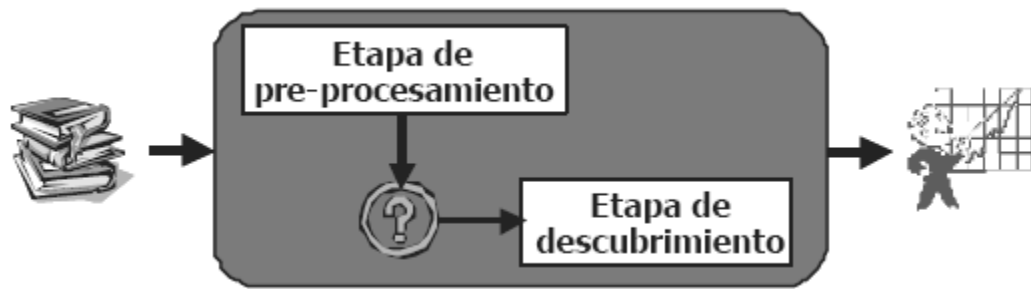


Figura 1.1 Proceso de minería de texto. [2]

La tabla 1 muestra los tres tipos de estrategias empleadas en los actuales sistemas de minería de texto. Como se observa, todos estos métodos limitan a un nivel temático o de entidad de sus resultados, haciendo imposible descubrir cosas más detalladas como: consensos, tendencias y desviaciones. ” [2]

Etapa de pre-procesamiento	Tipo de representación	Tipo de descubrimientos
Categorización	Vector de temas	Nivel temático
Full-text	Secuencia de palabras	Patrones de lenguaje
Extracción de información	Tabla de datos	Relaciones entre entidades

Tabla 1 Estado del arte de la minería de texto. [2]

A pesar de los tipos de estrategias que son empleadas en los sistemas actuales existe otros son llamados *agrupamientos de grafos conceptuales*, no es recomendable utilizarlos ya que no existen tantas fuentes sobre esta técnica.

1.3 TECNICAS DE LA MINERÍA DE DATOS

Las técnicas de la minería de datos, nos permiten resolver el problema de la extracción de conocimiento; lo cual se lo realiza, partiendo de grandes cantidades de información teniendo como objetivo descubrir patrones, información válida, novedosa, interesante y sobre todo comprensible.

“Entre las técnicas tenemos:

- 1.3.1 **Técnicas algebraicas y estadísticas:** Se basan, generalmente, en expresar modelos y patrones mediante fórmulas algebraicas.
- 1.3.2 **Técnicas bayesianas:** Se basan en estimar la probabilidad de pertenencia (a una clase o grupo), mediante la estimación de las probabilidades condicionales inversas o a priori, utilizando para ello el teorema de Bayes.
- 1.3.3 **Técnicas basadas en conteos de frecuencias y tablas de contingencia:** Estas técnicas se basan en contar la frecuencia en la que dos o más sucesos se presenten conjuntamente.

Las ventajas de estas técnicas son: una buena efectividad, un considerable ahorro en términos de mano de obra del experto, y cierta independencia del dominio” [16].

Para el desarrollo del proyecto, utilizamos una de las técnicas bayesianas como el algoritmo del NAIVE BAYES.

1.4 GENERACIÓN DE UN CLASIFICADOR AUTOMÁTICO

Como antes se mencionó las técnicas bayesianas fueron las que se escogieron para el desarrollo de este proyecto, las cuales son usadas en la generación de un clasificador automático.

“Un clasificador automático se genera mediante la aplicación de un conjunto de ejemplos de clasificación, constituido por documentos ya clasificados. Generalmente existe una fase de entrenamiento durante la cual el clasificador ajusta sus parámetros de operación sobre la base de los errores cometidos

hasta lograr un nivel de efectividad aceptable. Los documentos presentados en esta fase se consideran correctamente clasificados ” [3]

La finalidad del clasificador de correo electrónico es que utilice el conjunto de mensajes que están clasificados (bandeja de entrada y correo SPAM) para así instruirse del ejemplo para efectuar las tareas respectivas. El propósito del sistema es que parta sin conocimiento y que a futuro aprenda a descartar mediante la intervención del usuario. Ya que la clasificación es un proceso que sufre alteraciones no solo de forma, ni de fondo sino también por los cambios de interés que tiene los usuarios; es necesario que sea ajustado después de algún tiempo de uso.

1.4.1 CLASIFICADOR PROBABILÍSTICO NAIVE BAYES

Los CLASIFICADORES BAYESIANOS son clasificadores estadísticos, que pueden predecir tanto las probabilidades del número de miembros de clase, como la probabilidad de que una muestra dada pertenezca a una clase particular.

“Los clasificadores probabilísticos están basados en la idea de que existe una distribución de probabilidad para definir la pertenencia de un documento con respecto a un conjunto de clases.

El clasificador supone que la aparición de una palabra en un documento es independiente de la aparición de las demás palabras. Esto es lo que se conoce como la aproximación "bag of words", utilizada en técnicas de recuperación de la información para simplificar el tratamiento matemático del problema. ” [3]

La “CLASIFICACIÓN BAYESIANA” se basa en el TEOREMA DE BAYES, y los CLASIFICADORES BAYESIANOS han demostrado una alta exactitud y velocidad cuando se han aplicado a grandes bases de datos.

1.5 FILTROS BAYESIANOS

Los filtros bayesianos se basan en estadísticas de palabras que aparecen en los correos no deseados, lo que hace el filtro es leer todos los correos que nos llegan y una vez analizado su contenido, determinar según las palabras que contengan en la base de datos si son correos “SPAM” o no, si son correos “SPAM” son enviados a su respectiva carpeta del mismo nombre de lo contrario son enviados a la bandeja de entrada. Para que un filtro bayesiano funcione, lo que tenemos que hacer es enseñarle, marcando poco a poco todo el correo basura que recibimos. Para que comience a ser efectivo el filtro debemos de proporcionarles mensajes SPAM y no SPAM (mensajes buenos), de este modo sabrán diferenciar su contenido.

1.5.1 FILTRADO BAYESIANO

Utilizamos este método para la realización del SPAM, en nuestra base de datos tenemos dos tablas llamadas *palabrasSPAM* y *palabrasnoSPAM* en las cuales en una de ellas se almacena las palabras más comunes catalogadas como SPAM como son: VIAGRA, HIPOTECA, PESO etc. y en la otra las palabras que no son catalogadas como SPAM, después se las compara con el contenido del mensaje entrante, es decir a medida que se va realizando las comparaciones entre el contenido del correo electrónico con el contenido de las tablas de la base de datos antes mencionadas, se van obteniendo las

probabilidades para luego determinar si dicho correo es SPAM o no, pero a lo largo de este trabajo el lector podrá encontrar con más detalle como funciona el filtro bayesiano desarrollado en este proyecto.

1.6 ¿QUE ES EL SPAM?

En pocas palabras el SPAM quiere decir *envío indiscriminado de mensajes de correo electrónico no solicitados*, estos generalmente se tratan de publicidad de ofertas de productos, bienes y servicios, de páginas web, etc.

“Actualmente, se calcula que entre el 60 y el 80% de los correos que se envían son SPAM. El SPAM es perjudicial para todos, hasta para la empresa que los envía.” [4]

Pero el SPAM puede también servir como medio de propagación de un peligro mayor, como son los llamados *virus informáticos*.

1.6.1 CARACTERÍSTICAS

Los usuarios deben conocer las características comunes que tienen los correos basura denominados también como correos “SPAM” tales como asuntos llamativos y contenido publicitario entre otras.

“Las principales características son las que se presentan a continuación:

- La dirección que aparece como remitente del mensaje no resulta conocida para el usuario.
- El mensaje no suele tener dirección para reenviar.
- Presentan un asunto llamativo.
- El contenido es publicitario

- Aunque el método de distribución más habitual de este tipo de “**malware**¹” [6] (software malicioso), es el correo electrónico, existen diversas variantes, las cuales son: SPAM, SPIM, SPIT [7] y los SPAM SMS.” [5]

Los enviados a través del correo electrónico son los SPAM; los que se usan en aplicaciones de tipo mensajería instantánea son los llamados SPIM; también tenemos los de publicidad a través de telefonía sobre Internet llamados SPIT y por último el SPAM destinado a enviarse a dispositivos móviles mediante SMS (Short Message Service) se los denomina SPAM SMS.

1.6.2 TIPOS DE SPAM

La llegada de Internet a nuestras vidas ha traído una serie de beneficios pero este enorme crecimiento de la red también ha acarreado una serie de problemas para quienes se comunican a través de este medio y para quienes brindan servicios de acceso a Internet.

“Hoy en día, SPAM es una palabra familiar, ya que el 70-80% de todo el tráfico de correo es SPAM. En la mayor parte de los casos el SPAM es publicidad y la experiencia nos muestra que los spammers escogen ciertos bienes y servicios para promoverlos. ” [8]

El SPAM se ha convertido en una vía para promocionar productos o servicios ilegales o rechazables, como cadenas de dinero, acceso a pornografía, difusión

¹ Malware: Proviene de una agrupación de las palabras malicious software. Este programa o archivo, que es dañino para el ordenador, está diseñado para insertar virus, gusanos, troyanos o spyware intentando conseguir algún objetivo, como podría ser el de recoger información sobre el usuario o sobre el ordenador en sí.

de pornografía infantil y otros. También se basan en el engaño a los clientes y en falsas promociones para conseguir direcciones de usuarios.

1.6.2.1 SPAMs MAS COMUNES EN EL CORREO ELECTRÓNICO

Existen diferentes categorías de SPAM en los correos electrónico entre las más comunes tenemos las siguientes: pornografía, salud y medicina, tecnologías informáticas, finanzas personales, educación/entrenamiento, lotería y juegos de apuestas, comercial, políticos, religiosos y de hostigamientos.

Pornografía: “Incluye ofertas de productos diseñados para aumentar o mejorar la potencia sexual, anuncios o enlaces a sitios pornográficos” [8]. 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 1.2 SPAM pornográfico [8]

En la figura 1.2 se muestra un correo basura con contenido sexual, en el cual le ofrecen el viagra más barato del mundo captando la atención del usuario, en este caso los del sexo masculino.

Tecnologías informáticas: “Incluye ofertas de hardware y software a bajo precio, como también servicios para los usuarios de sitios web: hospedaje (hosting), registro de dominios” [8]. 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 1.3 SPAM de tecnologías informáticas [8]

En la figura 1.3 podemos observar el contenido de un correo basura tecnológico con una lista software de alta calidad y a muy bajo costo.

Salud y medicina: “Incluye anuncios de sistemas para adelgazar, para el cuidado de la piel, la cura de la calvicie, suplementos de dietas” [8]. Ejemplo:

Tema: Pierda hasta el 19% de su peso. Un nuevo sistema para adelgazar está aquí.

Hola, tengo una oferta especial para ti...
¿QUIERES PERDER PESO?

El sistema más poderoso de pérdida de peso ahora está disponible sin prescripción médica. Completamente natural: Adipren720 ¡100% de garantía, o te devolvemos el dinero! Más detalles sobre Adipren720: {ENLACE}

- Pierde hasta el 19% del peso de su cuerpo. - Pierde el 20-35% de grasa abdominal.
- Reducción del 40-70% de grasa subcutánea.
- Aumenta tu metabolismo en un 76,9% sin hacer ejercicios.
- Quema las grasas calorizadas.
- Aumenta la confianza en ti mismo y la autoestima.

Figura 1.4 SPAM de salud y medicina [8]

En la figura 1.4 encontramos el contenido de un correo basura de salud, en el cual se ofrece al usuario que tienen sobrepeso o que tienen algunas libritas de más; conseguir una apariencia mejor con un nuevo producto.

Finanzas personales: “Ofrece seguros, servicios de reducción de deudas, préstamos con bajos intereses” [8]. 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. Gracias. Llama 1-800-279-7310 o escríbenos a: 1700 E. Elliot Rd. STE3-C4 Tempe, AZ.

Figura 1.5 SPAM de finanzas personales [8]

En la figura 1.5 se observa el contenido de un correo basura de finanzas personales en la que le ofrecen a los usuarios refinanciar la deuda que posee, con la facilidad del que el crédito sea aprobado inmediatamente.

Educación: “Incluye ofertas de seminarios, entrenamientos y diplomaturas en línea” [8]. 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.}.

Figura 1.6 SPAM de educación / entrenamiento [8]

En la figura 1.6 encontramos un correo basura de educación cuyo contenido esta dirigido para los usuarios que desean obtener un grado de licenciatura o doctorado, pero que no cuenta con mucho tiempo libre.

Político: “Incluye cartas calumniantes o amenazas políticas de extremistas y posibles terroristas. A pesar de que estos son sólo mensajes molestos para los usuarios, las fuerzas de seguridad necesitan hacer un seguimiento de estos envíos, ya que pueden dar pistas sobre amenazas reales o ser un medio de comunicación real entre los terroristas”. [8]

Debemos recalcar que algunas categorías de SPAM son más peligrosas que otras; considerando que algunos de los mensajes contienen proposiciones de negocios no solicitadas, mientras que otros podrían contener algún tipo de virus.

2.- ANALISIS DEL SISTEMA

A lo largo de este capítulo se encontrará el análisis de los requerimientos y las herramientas necesarias para el desarrollo de este proyecto.

2.1 DEFINICIÓN DE REQUERIMIENTOS DEL SISTEMA

El sistema debe cumplir con los siguientes requerimientos:

- Analizar correos electrónicos en el módulo de administración.
 - Determinar el comportamiento de los usuarios dentro de la empresa.
 - Organizar los correos de acuerdo al comportamiento del usuario.
- Analizar correos electrónicos en el módulo del usuario.
 - Organizar correos de acuerdo a las preferencias.
 - Organizar correos de acuerdo a sus atributos.
- Detectar correos SPAM.
 - Eliminar correos considerados como SPAM.

La implementación de cada uno de estos requerimientos del sistema brindará al usuario mayor facilidad en el uso de la información que contenga en su cuenta electrónica, obteniendo como resultado una menor pérdida de tiempo.

2.2 DEFINICIÓN DE LOS ALCANCES DEL SISTEMA

Contemplando los requerimientos del sistema, se han definido los siguientes alcances:

Los correos que residen en el servidor por parte de todos los usuarios podrán ser organizados y clasificados de las siguientes maneras:

- Organizados por asunto.
- Organizados por remitente o destinatario.

- Organizados por fecha de recepción o envío.
- Organizados por prioridad.
- Clasificados por espacio en buzón.
- Clasificados por utilización de correo.

Los usuarios podrán clasificar sus correos mediante:

- Grupos personalizados por los usuarios.
- Asunto y remitente
- Clasificados por prioridad.
- Clasificados por fecha.

La detección de SPAM se va a realizar mediante el análisis del contenido de los correos entrantes.

2.3 ESPECIFICACIÓN DEL SISTEMA

El sistema consta de dos roles:

2.3.1 Módulo de administración

2.3.2 Módulo del usuario

La creación de estos dos módulos surgen de la necesidad de que tanto el administrador y los usuarios de las cuentas electrónicas, puedan organizar y clasificar la información de dichas cuentas en una manera más eficaz.

2.3.1 Módulo de administración: El sistema deberá:

- 1.- Autenticar mediante un administrador de correo del MAIL SERVER.
- 2.- Una vez autenticado se deberá recibir todos los correos almacenados en el MAIL SERVER.

3.- Ya listado los correos, se procederá a la organización y clasificación de éstos según el criterio del administrador. Los que serán:

- Clasificados por espacio en buzón
- Clasificados por utilización de correo

Estas clasificaciones, permitirán al administrador conocer cuales son los usuarios que ocupan mayor espacio en su correo, así también cuales son los que reciben correos SPAM, y otro tipo de información.

2.3.2 Módulo del usuario: El sistema deberá:

1. Verificar la cuenta del usuario registrada en el MAIL SERVER la cual contiene la información del mismo y a su vez autenticar la cuenta ya existente.

2.- Una vez autenticado se deberá receptor todos los correos almacenados para dicha cuenta.

3.- Una vez listado los correos, se procederá a su organización. Los que serán:

➤ Organizar correos de acuerdo a las preferencias.

- **Por grupos personalizados por los usuarios:** El usuario podrá crear diferentes grupos para el almacenamiento de sus correos entrantes.

➤ Organizar correos de acuerdo a sus atributos.

- **Por el asunto y remitente:** El usuario podrá clasificar sus correos mediante el remitente o realizando una clasificación filtrando el campo asunto por medio de un criterio establecido por el usuario.

- **Clasificados por prioridad:** Será organizada por medio de prioridades y se clasificará mediante búsqueda personalizada para datos establecidos por el sistema de correo.
 - **Clasificados por fecha:** Será organizado descendentemente o ascendentemente y se clasificará mediante una búsqueda personalizada de acuerdo a un período deseado.
- Eliminar correos considerados como SPAM.

Como se puede observar el módulo de usuario contiene varias opciones que le permitirán, manipular más eficazmente la información que se encuentre en el correo electrónico.

2.4 HERRAMIENTAS PARA LA IMPLEMENTACIÓN DEL SISTEMA

El sistema será implementado con 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.

Servidor de Correo

Software	Descripción
WINDOWS 2003 SERVER	Sistema operativo
EXCHANGE SERVER 2003	Contenedor de correos

Tabla 2 Software utilizado para el desarrollo del proyecto _ servidor
Fuente: Autores

Cliente

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 3 Software utilizado para el desarrollo del proyecto _ cliente
Fuente: Autores

2.4.1 PLATAFORMAS

Se eligió trabajar con la plataforma *Microsoft Windows XP Professional*, debido a que es una plataforma que provee los entornos de escritorio más usados a nivel personal, empresarial y corporativo.

2.4.2 HERRAMIENTAS DE DESARROLLO

VISUAL C# NET: Se utilizó este lenguaje ya que es de propósito general orientado a objetos creado por Microsoft para su nueva plataforma, además ofrece al programador una interfaz común para trabajar de manera cómoda y visual con cualquiera de los lenguajes de la plataforma .NET

SQL SERVER 2005: Se utilizó SQL SERVER 2005 porque además de las ventajas que nos brinda al ser una herramienta de fácil uso respecto al ingreso y manipulación de los datos, nos ofrece métodos para acceder a la información, desde el sistema, de una manera rápida como son los Store-Procedure. Como usuario de la herramienta se puede planear trabajos en varios servidores, acceder a archivos en otros equipos, realizar copias de seguridad en

ubicaciones de red, mandar notificaciones mediante e-mail, hacer modificaciones en el registro, entre otros beneficios.

Siendo administrador de la base de datos se tienen otros beneficios tales como poder crear y modificar cuentas de usuario, tener acceso a las claves de registros, acceso total al sistema de archivos del equipo, iniciar o detener los servicios del SQL Server, poder acceder a los registros de equipos clientes (si se van a utilizar controladores ODBC), poder utilizar el monitor del sistema y crear registros de log de cualquier tipo de contador, entre otros beneficios.

3. DISEÑO E IMPLEMENTACIÓN DEL SISTEMA

3.1 DISEÑO

El propósito de este capítulo es que el lector comprenda el diseño que se realizó para el desarrollo del proyecto. Además de definir las estrategias de diseño que se utilizó para la implementación.

3.1.1 DISEÑO DE CLASES

Documentación de actores

Nombre: Administrador del MAIL SERVER.

Descripción: Es el que se encarga de la clasificación de los usuarios de acuerdo al uso del correo y al espacio de disco que utilizan.

Funcionalidad: Interviene como actor primario en los casos de uso.

Nombre: Usuarios.

Descripción: Son los encargados de organizar y clasificar sus propios correos de acuerdo a sus preferencias.

Funcionalidad: Interviene como actor primario en los casos de uso.

Nombre: MAIL SERVER o servidor de correos.

Descripción: Es que proporciona al sistemas los datos de usuarios para realizar el análisis correspondiente.

Funcionalidad: Interviene como actor secundario en los casos de uso.

Lista de casos de uso

Caso de Uso 1: Autenticar el usuario como administrador de MAIL SERVER.

Caso de Uso 2: Clasificar a los usuarios en el módulo de administración por el espacio de buzón.

Caso de Uso 3: Clasificar a los usuarios en el módulo de administración por el uso de correo.

Caso de Uso 4: Autenticar al usuario con su cuenta de correo.

Caso de Uso 5: Organizar los correos en el módulo del usuario.

Caso de Uso 6: Detectar correos SPAM en el módulo del usuario.

Caso de Uso 7: Marcar correo SPAM en el módulo del usuario.

Caso de Uso 8: Determinar que contactos indirectos pueden formar parte de la listas de contactos del usuario.

DIAGRAMA DE CASOS DE USO

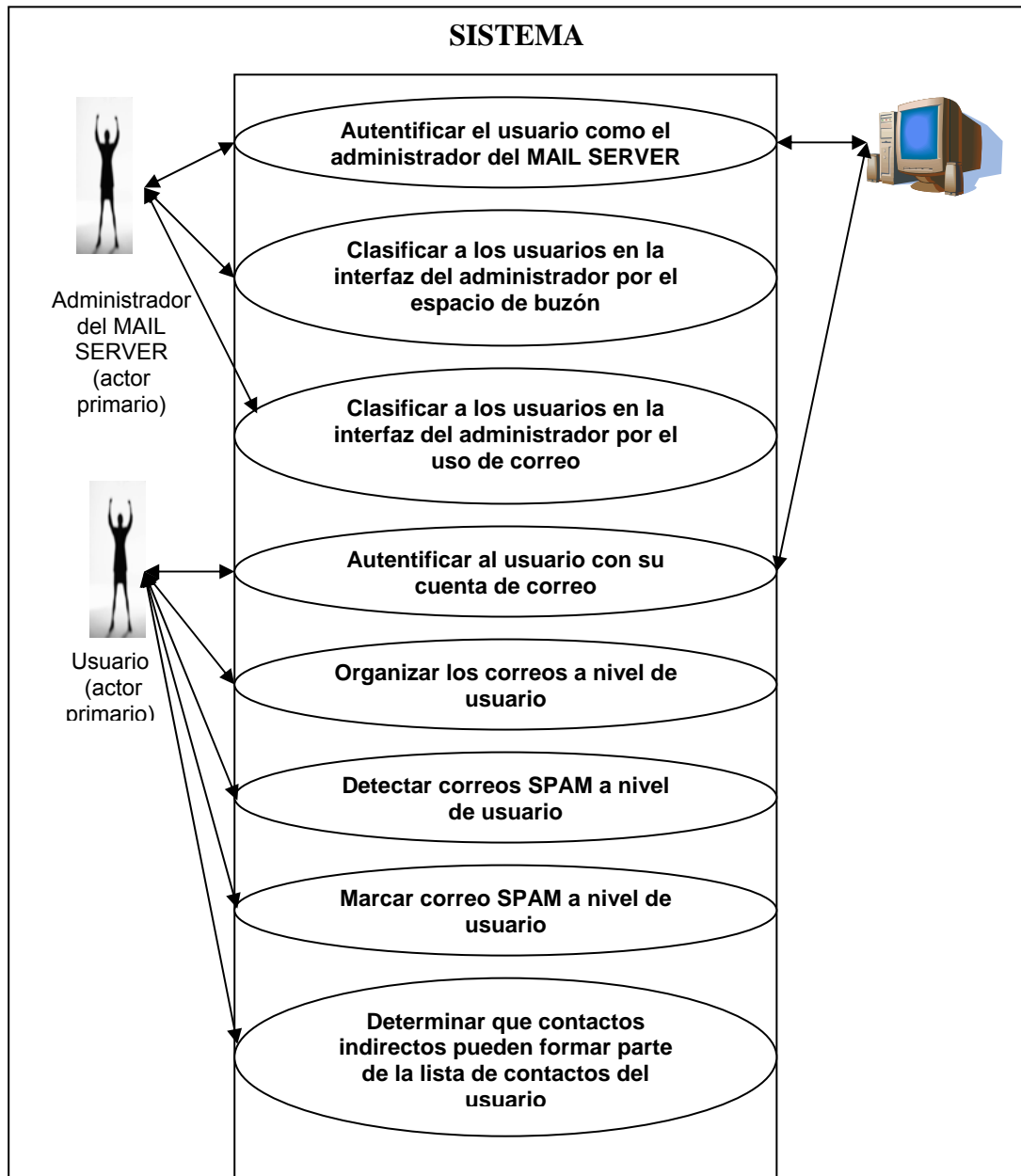


Figura 3.1 Diagrama de casos de uso

Fuente: Autores

DESCRIPCIÓN DE LOS CASOS DE USOS

En esta sección el lector encontrará una descripción breve de los casos de uso que forman parte del desarrollo de este proyecto.

Caso de Uso 1:

Nombre: Autenticar el usuario como administrador de MAIL SERVER.

Descripción: Este caso de uso permite ingresar al administrador del MAIL SERVER, donde el podrá realizar la clasificación de los usuarios de acuerdo al uso y espacio de disco que utilizan.

Notas: Solo los administradores del sistema podrán validarse por seguridad.

Caso de Uso 2:

Nombre: Clasificar a los usuarios en la interfaz del administrador por el espacio de buzón.

Descripción: Este caso de uso permite al administrador obtener la información sobre los usuarios que ocupan mayor espacio en el buzón.

Notas: Se forman grupos de acuerdo al espacio de buzón que ocupan los usuarios.

Caso de Uso 3:

Nombre: Clasificar a los usuarios en la interfaz del administrador por el uso de correo.

Descripción: Este caso de uso permite al administrador obtener la información sobre el uso de correo con respecto al usuario.

Notas: Se forman grupos de acuerdo al uso de la cuenta de correo que posean los usuarios.

Caso de Uso 4:

Nombre: Autenticar al usuario con su cuenta de correo.

Descripción: Este caso de uso permite ingresar a los usuarios, donde el podrá organizar su correos de acuerdo a sus necesidades.

Notas: Cada usuario podrá ingresar con su usuario y contraseña respectivamente.

Caso de Uso 5:

Nombre: Organizar los correos en el módulo del usuario.

Descripción: Este caso de uso el usuario podrá organizar los correos por los campos asunto, remitente o destinatario, fecha de recepción o envío y por prioridad.

Notas: La organización se lo hará alfabéticamente.

Caso de Uso 6:

Nombre: Detectar correos SPAM en el módulo del usuario.

Descripción: En este caso de uso el usuario podrá saber si los correos que recibe son o no un SPAM, así poder eliminarlos.

Notas: La detección de los correos SPAM, se realiza cuando el usuario inicie su sesión.

Caso de Uso 7:

Nombre: Marcar correo SPAM en el módulo del usuario.

Descripción: En este caso de uso el usuario podrá marcar como correo SPAM, a los correos que no fueron detectados por el filtro.

Notas: No existe límite para la cantidad de correo que el usuario considere como SPAM.

Caso de Uso 8:

Nombre: Determinar que contactos indirectos pueden formar parte de la listas de contactos del usuario.

Descripción: En este caso de uso el usuario podrá decidir, cuales de sus contactos indirectos pueden formar parte de la lista de contactos.

Notas: Aquí el usuario podrá determinar cuales son los contactos indirectos que envían correos SPAM.

DESCRIPCIÓN DE ESCENARIOS

A continuación se describe cada uno de los escenarios que se producen dentro de cada caso de uso anteriormente descrito.

Caso de Uso 1: Autenticar el usuario como Administrador de MAIL SERVER.

El desarrollo de los siguientes escenarios de este caso de uso, se lleva acabo cuando el administrador realiza el ingreso a la interfaz que proporciona el sistema como ayuda para él.

Escenario 1.1.- Ingreso exitoso de usuario y contraseña.

Suposiciones:

- El usuario del sistema han ingresado su respectivo usuario correctamente.
- El usuario del sistema han ingresado su contraseña correctamente.

Resultados:

- Se habilitará las respectivas opciones del menú según el perfil que posean.

Este escenario se convertirá en exitoso solo y solamente si el administrador ingresó su usuario y contraseña correctamente.

Escenario 1.2.- Ingreso no exitoso por usuario incorrecto.

Suposiciones:

- El usuario del sistema han ingresado incorrectamente su usuario.

Resultados:

- Se pedirá que ingrese nuevamente su usuario.

Este escenario se convertirá no exitoso debido a que el administrador no ingresó correctamente su usuario.

Escenario 1.3.- Ingreso no exitoso por contraseña incorrecta.

Suposiciones:

- El usuario del sistema han ingresado incorrectamente su contraseña.

Resultados:

- Se pedirá que ingrese nuevamente su contraseña.

Este escenario se convertirá no exitoso debido a que el administrador no ingresó correctamente su contraseña.

Escenario 1.4.- Ingreso no exitoso por no haber ingresado los datos que son exigidos.

Suposiciones:

- El usuario del sistema no ingresó su usuario.
- El usuario del sistema no ingresó su contraseña.

Resultados:

- Se pedirá que ingrese nuevamente el dato exigible correspondiente.

Este escenario se convertirá no exitoso debido a que el administrador no ingresó ninguno de los campos solicitado como son el usuario y contraseña.

Escenario 1.5.- Ingreso no exitoso por fallas técnicas correspondientes al ingreso del usuario o contraseña.

Suposiciones:

- Se tiene problemas al momento de la validación de la información ingresada debido a un problema con la base de datos.

Resultados:

- Mensaje que indique la causa del problema.

Caso de Uso 2: Clasificar a los usuarios en la interfaz del administrador por el espacio de buzón.

En el siguiente escenario de este caso de uso se describe la información que podrá observar el administrador cuando presione el botón de clasificación por espacio de buzón.

Escenario 2.1.- Clasificación exitosa por espacio en el buzón.

Suposiciones:

- El administrador verifica los espacios de los buzones de las cuentas de los correos de los usuarios y los clasifica.

Resultados:

- El servidor muestra los correos clasificados según su tamaño.

Caso de Uso 3: Clasificar a los usuarios en la interfaz del administrador por el uso de correo.

En el siguiente escenario de este caso de uso se describe la información que podrá observar el administrador cuando presione el botón de clasificación por uso de correo.

Escenario 3.1.- Clasificación exitosa por uso de correo.

Suposiciones:

- El servidor almacenará las cuentas clasificadas según las cantidades de envío que realicen los usuarios.

Resultados:

- El servidor muestra los correos clasificados según el uso de la cuenta de correo.

Caso de Uso 4: Autenticar al usuario con su cuenta de correo.

El desarrollo de los siguientes escenarios de este caso de uso, se lleva a cabo cuando los usuarios realizan el ingreso a su cuenta de correo con su respectivo usuario y contraseña.

Escenario 4.1.- Ingreso exitoso de usuario y contraseña.

Suposiciones:

- El usuario de la cuenta ha ingresado su respectivo “usuario” correctamente.
- El usuario del sistema han ingresado su “contraseña” correctamente.

Resultados:

- Se habilitará las respectivas opciones del menú según el perfil que posean.

Este escenario se convertirá en exitoso solo y solamente si el usuario ingresó su usuario y contraseña correctamente.

Escenario 4.2.- Ingreso no exitoso por usuario incorrecto.

Suposiciones:

- El usuario de la cuenta ha ingresado incorrectamente su “usuario”.

Resultados:

- Se pedirá que ingrese nuevamente su usuario.

Este escenario se convertirá no exitoso debido a que el usuario no ingresó correctamente su usuario.

Escenario 4.3.- Ingreso no exitoso por contraseña incorrecto.

Suposiciones:

- El usuario de la cuenta ha ingresado incorrectamente su contraseña.

Resultados:

- Se pedirá que ingrese nuevamente su contraseña.

Este escenario se convertirá no exitoso debido a que el usuario no ingreso correctamente su contraseña.

Escenario 4.4.- Ingreso no exitoso por no haber ingresado los datos que son exigidos.

Suposiciones:

- El usuario del sistema no ingresó su usuario.
- El usuario del sistema no ingresó su contraseña.

Resultados:

- Se pedirá que ingrese nuevamente el dato exigible correspondiente.

Este escenario se convertirá no exitoso debido a que el administrador no ingreso ninguno de los campos solicitado como son el usuario y contraseña.

Escenario 4.5.- Ingreso no exitoso por fallas técnicas correspondientes al ingreso del “usuario” o “contraseña”.

Suposiciones:

- Se tienen problemas al momento de la validación de la información ingresada debido a un problema con la base de datos.

Resultados:

- Mensaje que indique la causa del problema.

Caso de Uso 5: Organizar los correos a nivel de usuario.

En los siguientes escenarios de este caso de uso se describen cada una de las organizaciones de correos que el usuario podrá realizar dentro de su respectiva cuenta de correo.

Escenario 5.1.- Organización exitosa de los correos electrónicos por asunto.

Suposiciones:

- El sistema permitirá la organización de los correos mediante la acción del ratón sobre la cabecera del asunto.

Resultados:

- Se mostrarán los correos listados por el campo asunto, ya sea ascendente o descendentemente en orden alfabético.

El escenario es exitoso cuando el sistema realiza la organización de los correos electrónicos por asunto, mostrando la información correspondiente.

Escenario 5.2.- Organización no exitosa de los correos electrónicos por asunto.

Suposiciones:

- El sistema no puede realizar la organización de los correos, por alguna falla técnica por ejemplo: conexión a la red.

Resultados:

- Se mostrará un mensaje para informar que la operación no se ha podido realizar.

El escenario es no exitoso por alguna falla técnica producida internamente o externamente en el sistema.

Escenario 5.3.- Organización exitosa de los correos electrónicos por remitente o destinatario.

Suposiciones:

- El sistema permitirá la organización de los correos mediante la acción del ratón sobre la cabecera del remitente o destinatario.

Resultados:

- Se mostrará los correos listados por el campo de remitente o destinatario, ya sea ascendente o descendientemente en orden alfabético.

El escenario es exitoso cuando el sistema realiza la organización de los correos electrónicos por remitente o destinatario, mostrando la información correspondiente.

Escenario 5.4.- Organización no exitosa de los correos electrónicos por remitente o destinatario.

Suposiciones:

- El sistema no puede realizar la organización de los correos, por alguna falla técnica por ejemplo: conexión a la red.

Resultados:

- Se mostrará un mensaje para informar que la operación no se ha podido realizar.

El escenario es no exitoso por alguna falla técnica producida internamente o externamente en el sistema.

Escenario 5.5.- Organización exitosa de los correos electrónicos por fecha de envío o recepción.

Suposiciones:

- El sistema permitirá la organización de los correos mediante la acción del ratón sobre la cabecera de la fecha de envío o sobre la cabecera de fecha de recepción.

Resultados:

- Se mostrará los correos listados por el campo de fecha de envío o recepción, ya sea ascendente o descendentemente en orden alfabético.

El escenario es exitoso cuando el sistema realiza la organización de los correos electrónicos por fecha envío o recepción, mostrando la información correspondiente.

Escenario 5.6.- Clasificación no exitosa de los correos electrónicos por fecha de envío o recepción.

Suposiciones:

- El sistema no puede realizar la organización de los correos, por alguna falla técnica por ejemplo: conexión a la red.

Resultados:

- Se mostrará un mensaje con la causa del problema.

El escenario es no exitoso por alguna falla técnica producida internamente o externamente en el sistema.

Escenario 5.7.- Organización exitosa por prioridad.

Suposiciones:

- El administrador verifica los datos, organiza según las prioridades de los correos de los usuarios.
- El servidor debe de tener almacenados y organizados los correos de los usuarios de acuerdo a las prioridades.

Resultados:

- Se muestra los correos organizados por su prioridad.

Caso de Uso 6: Detectar correos SPAM a nivel de usuario.

En los siguientes escenarios de este caso de uso se muestra la aplicación del filtro Anti-SPAM que hemos desarrollado en el proyecto.

Escenario 6.1.- Detección exitosa de correo SPAM.

Suposiciones:

- La información contenida en el correo es considerada publicidad.

Resultados:

- El usuario podrá eliminar el correo SPAM si el lo desea.

Los correos SPAM detectados por el filtro serán almacenados en una carpeta designada para estos.

Escenario 6.2.- Detección no exitosa de correo SPAM.

Suposiciones:

- La información que contenía el correo no en realidad un SPAM.
- La información que contenía el correo no cumple con las reglas que tenemos para ser considerado un SPAM.

Resultados:

- El usuario podrá revisar su correo y sin eliminar ninguno de sus correos.

Si existe un correo que el usuario considere como SPAM podrá ser marcado como tal.

Caso de Uso 7: Marcar correo SPAM a nivel de usuario.

En los siguientes escenarios de este caso de uso se muestra como el filtro Anti-SPAM que hemos desarrollado en este proyecto, adquiere el conocimiento necesario para ser eficiente.

Escenario 7.1.- Determinación exitosa de correo SPAM

Suposiciones:

- La información contenida en el correo es considerada publicidad para el usuario.

Resultados:

- El usuario podrá marcar el correo considerado como SPAM y automáticamente este se almacenará en la carpeta que ha sido designada para este.

Caso de Uso 8: Determinar que contactos indirectos pueden formar parte de la listas de contactos del usuario.

En los siguientes escenarios de este caso de uso se muestra el desarrollo que les permitirá a los usuarios determinar que contactos indirectos pueden formar parte de su lista de contactos.

Escenario 8.1.- Determinación exitosa de contactos indirectos.

Suposiciones:

- El usuario presiona en el botón que realiza la acción de mostrar todos los contactos indirectos que dicho usuario posee.

Resultados:

- El usuario podrá seleccionar los contactos indirectos que desea que pertenezcan a su lista de contactos.

El escenario es exitoso solo y solamente si el sistema muestra de manera eficiente todos los contactos indirectos que posea el usuario.

DIAGRAMAS DE INTERACCIÓN DE OBJETOS

Caso de Uso 1: Autenticar el usuario como administrador de MAIL SERVER.

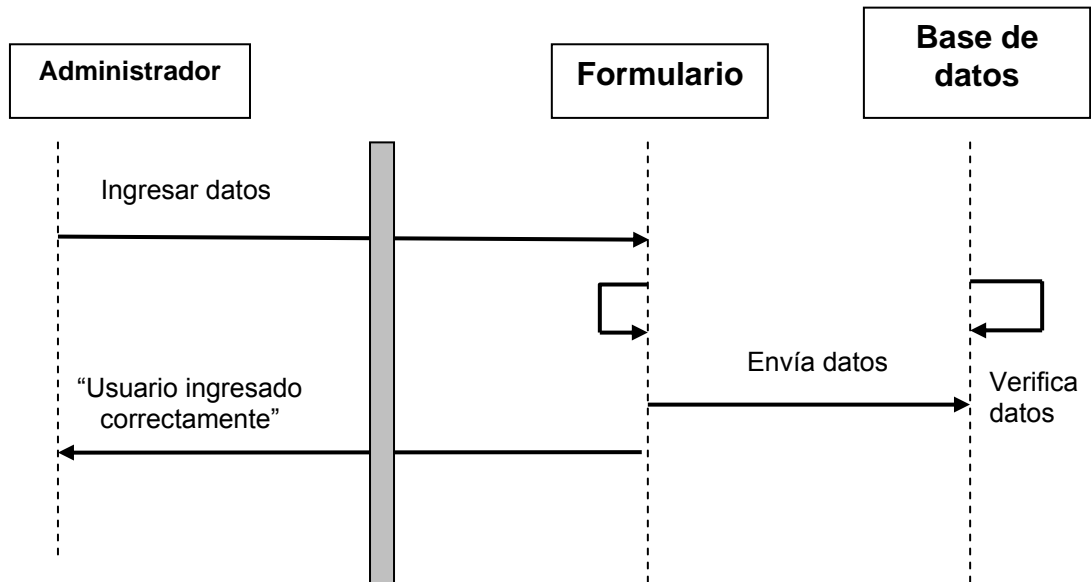


Figura 3.2 Escenario 1.1 Ingreso exitoso de “usuario” y “contraseña”.
Fuente: Autores

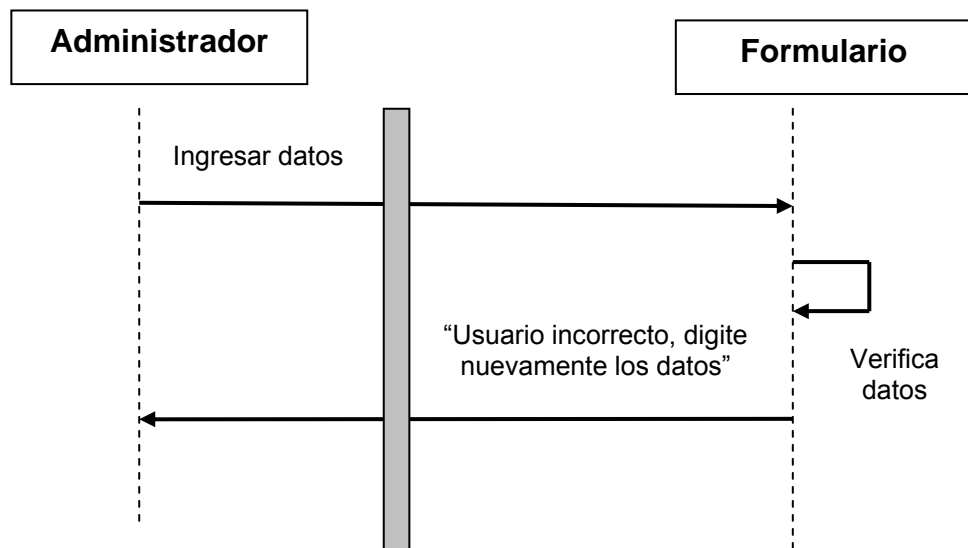


Figura 3.3 Escenario 1.2 Ingreso no exitoso por “usuario” incorrecto.
Fuente: Autores

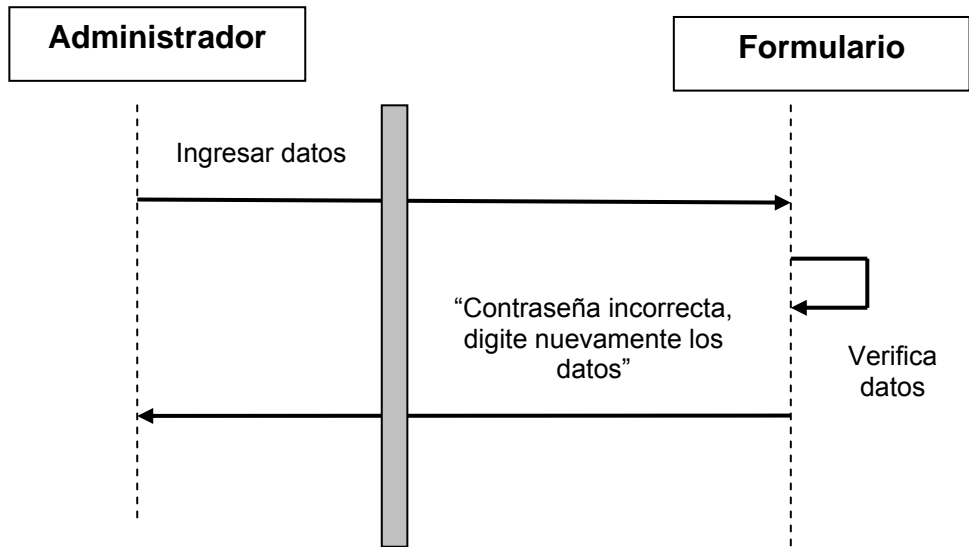


Figura 3.4 Escenario 1.3 Ingreso no exitoso por contraseña incorrecta.
Fuente: Autores

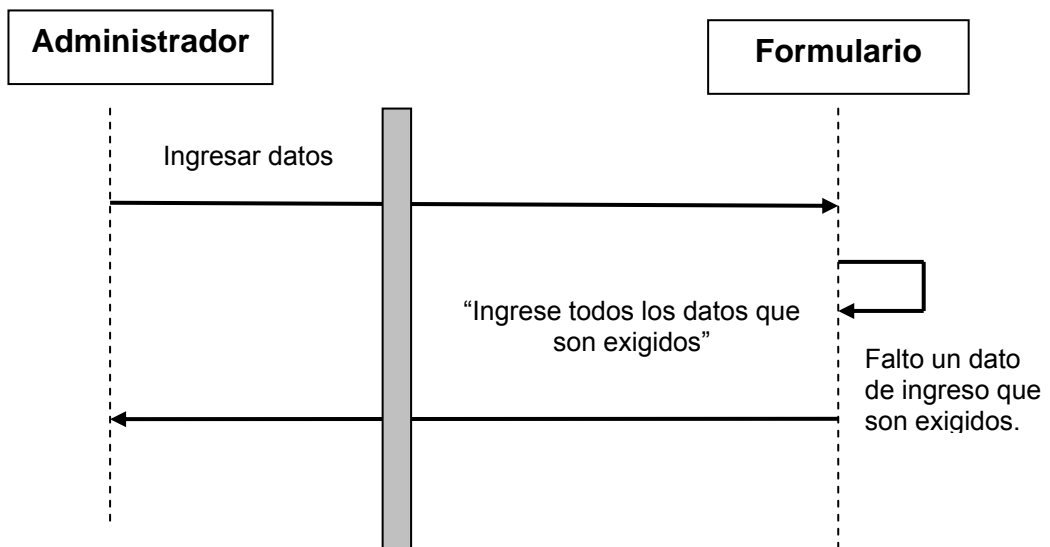


Figura 3.5 Escenario 1.4 Ingreso no exitoso por no haber ingresado los datos que son exigidos.
Fuente: Autores

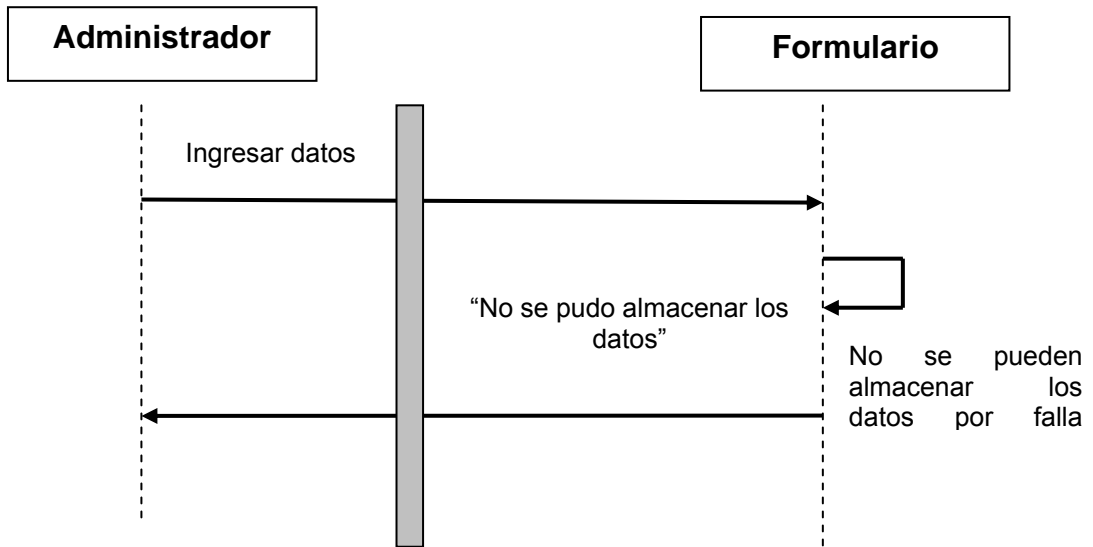


Figura 3.6 Escenario 1.5 Ingreso no exitoso por fallas técnicas correspondientes al ingreso del “usuario” o “contraseña”.
Fuente: Autores

Caso de Uso 2: Clasificar a los usuarios en el módulo de administración por el espacio de buzón.

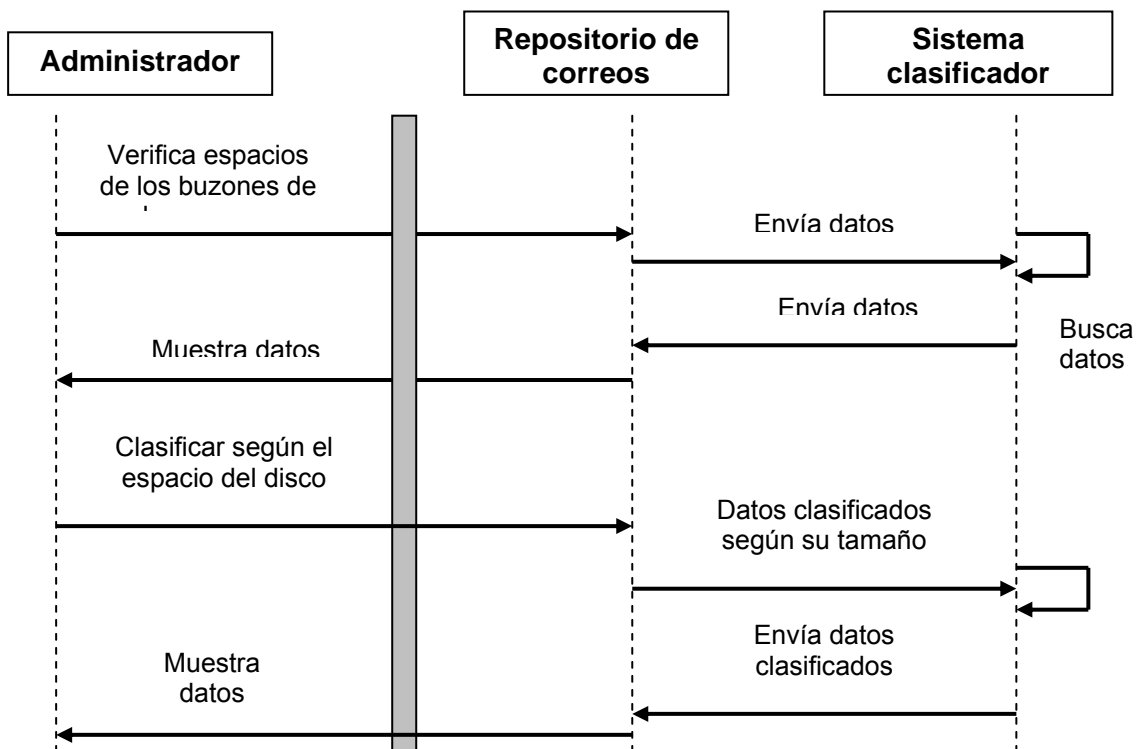


Figura 3.7 Escenario 2.1 Clasificación exitosa por espacio en el buzón.
Fuente: Autores

Caso de Uso 3: Clasificar a los usuarios en el módulo de administración por el uso de correo.

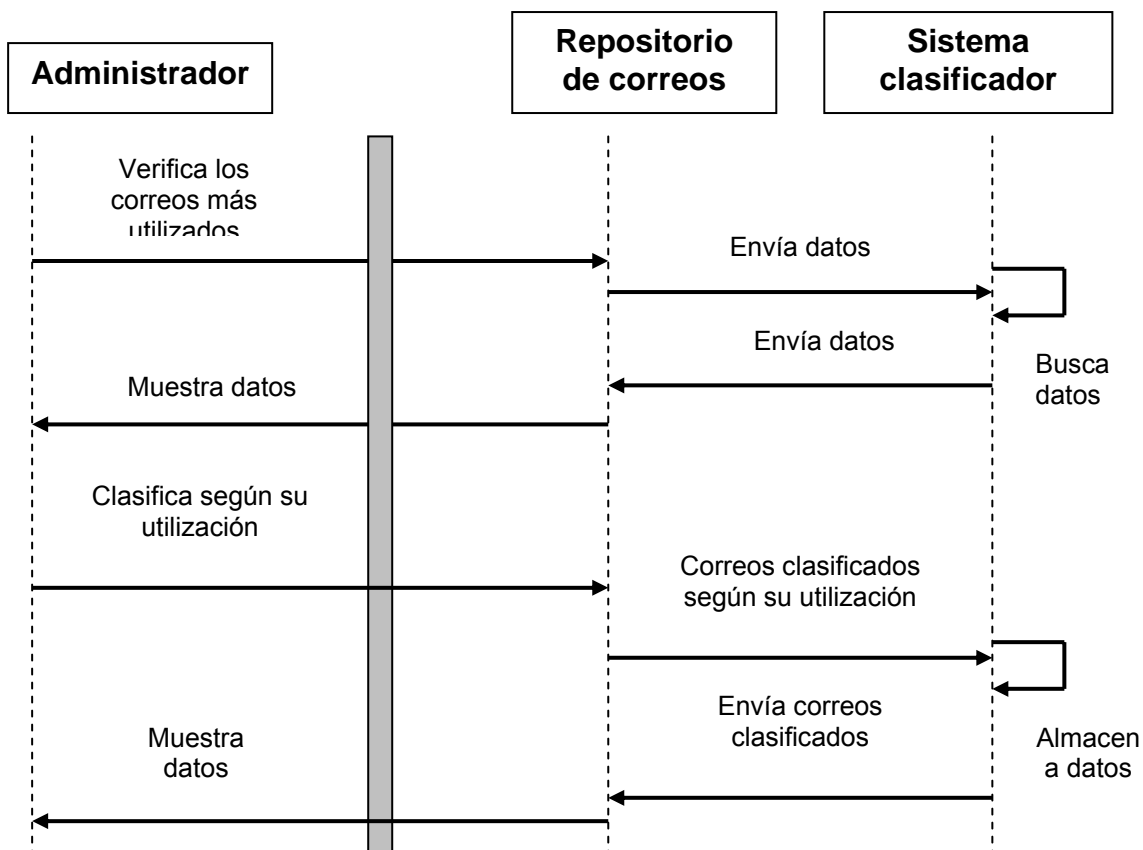


Figura 3.8 Escenario 3.1 Clasificación exitosa por uso de correo.

Fuente: Autores.

Caso de Uso 4: Autenticar al usuario con su cuenta de correo.

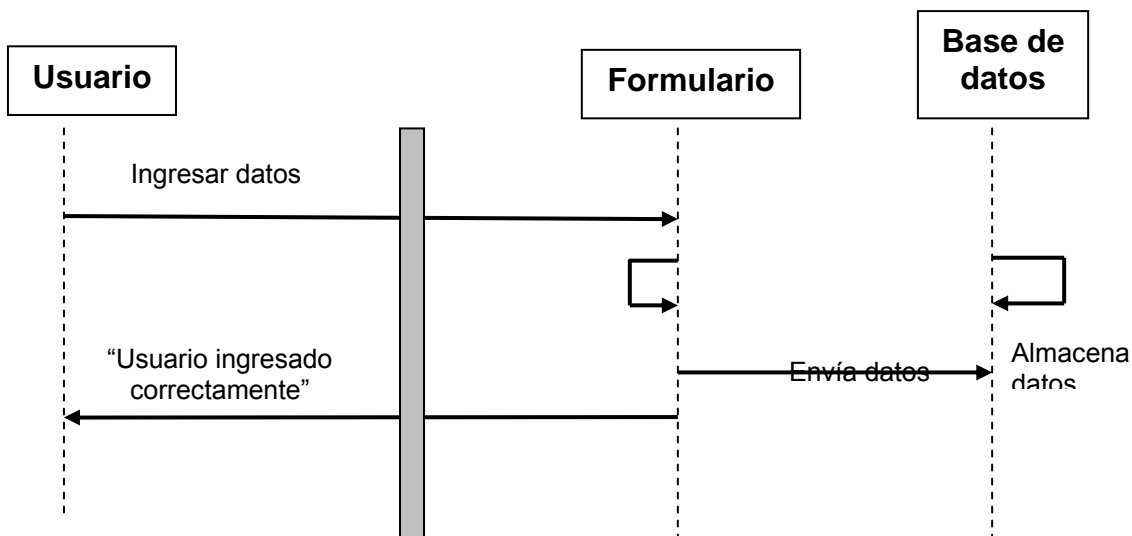


Figura 3.9 Escenario 4.1 Ingreso exitoso de "usuario" y "contraseña".

Fuente: Autores.

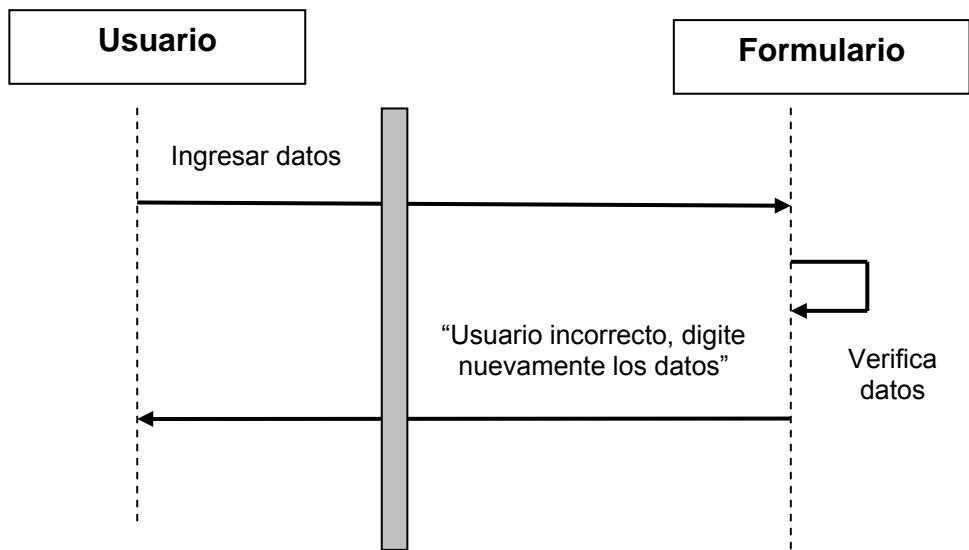


Figura 3.10 Escenario 4.2 Ingreso no exitoso por “usuario” incorrecto.
 Fuente: Autores.

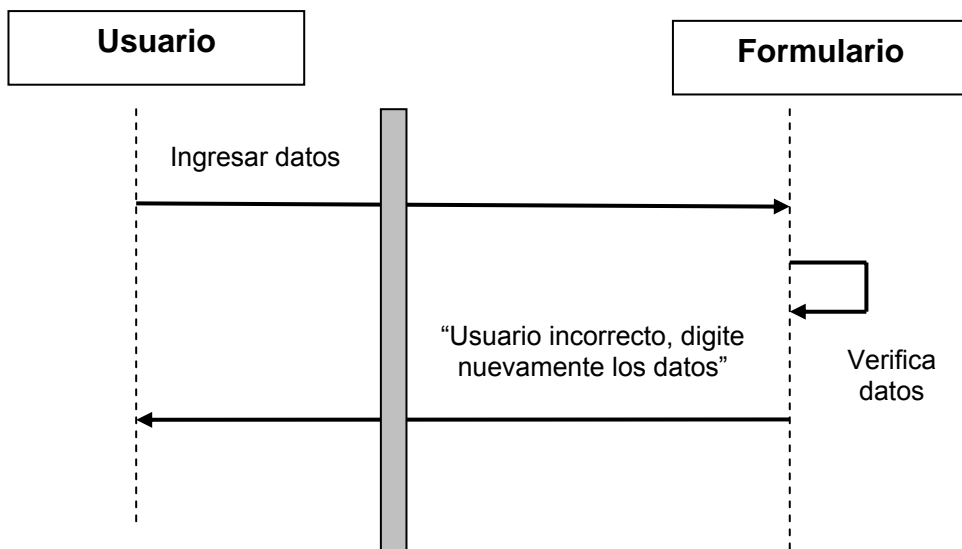


Figura 3.11 Escenario 4.3 Ingreso no exitoso por “contraseña” incorrecta.
 Fuente: Autores.

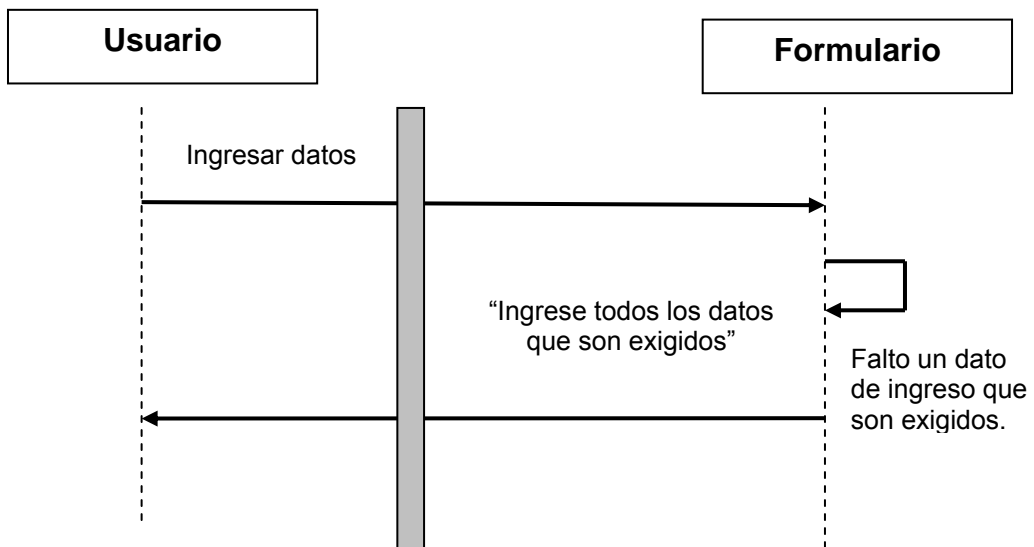


Figura 3.12 Escenario 4.4 Ingreso no exitoso por no haber ingresado los datos que son exigidos.
Fuente: Autores

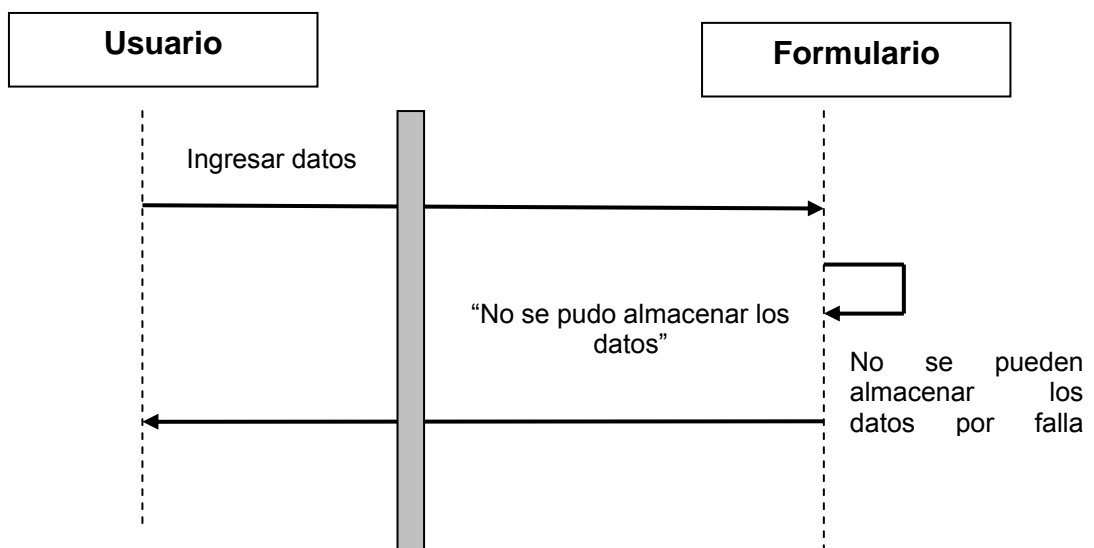


Figura 3.13 Escenario 4.5 Ingreso no exitoso por fallas técnicas correspondientes al ingreso del "usuario" o "contraseña".
Fuente: Autores.

Caso de Uso 5: Organizar los correos en el módulo del usuario.

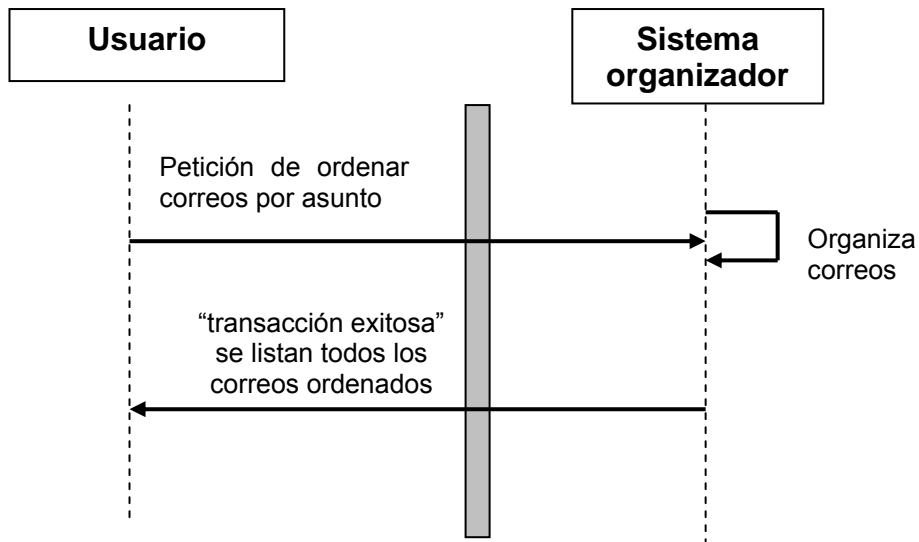


Figura 3.14 Escenario 5.1 Organización exitosa de los correos electrónicos por asunto.
Fuente: Autores.

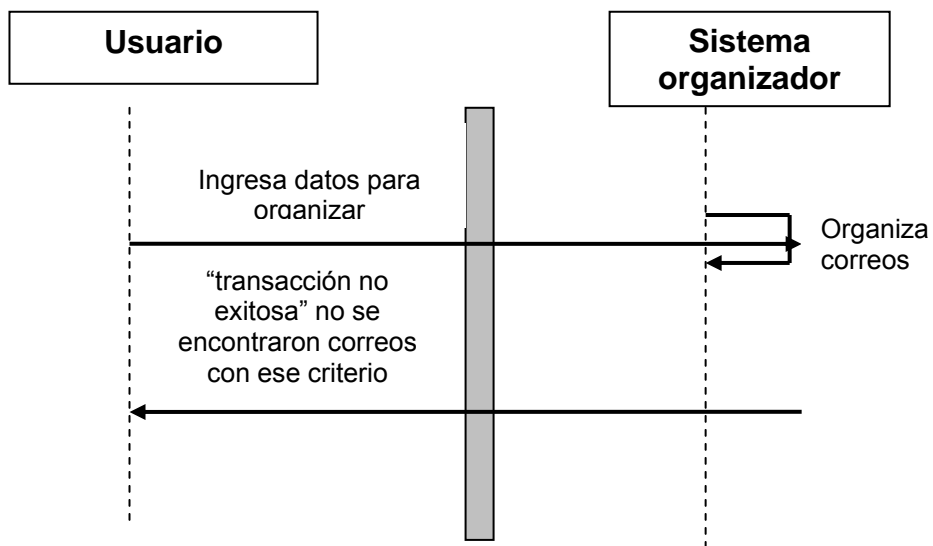


Figura 3.15 Escenario 5.2 Organización no exitosa de los correos electrónicos por asunto.
Fuente: Autores.

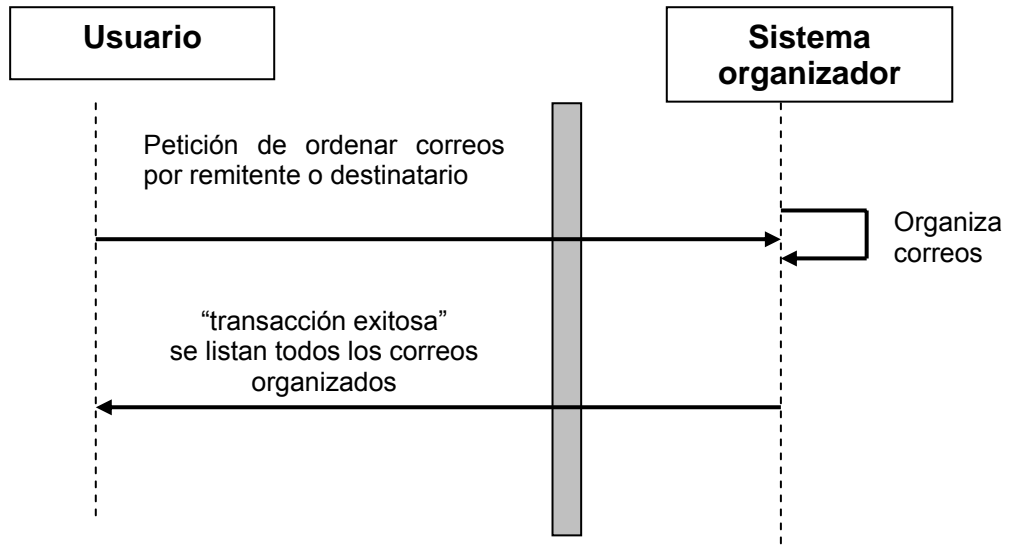


Figura 3.16 Escenario 5.3 Organización exitosa de los correos electrónicos por remitente o destinatario.
Fuente: Autores.

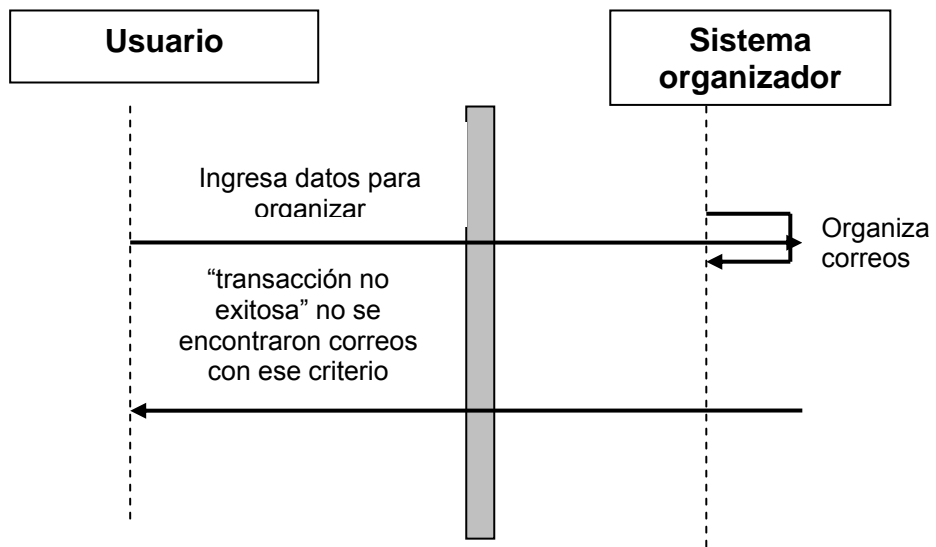


Figura 3.17 Escenario 5.4 Organización no exitosa de los correos electrónicos por remitente o destinatario.
Fuente: Autores.

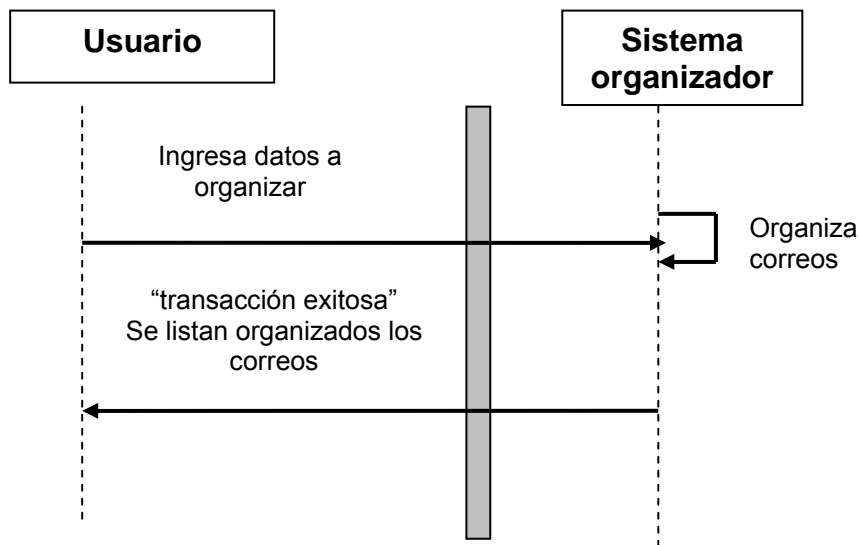


Figura 3.18 Escenario 5.5 Organización exitosa de los correos electrónicos por fecha de envío o recepción.
Fuente: Autores.

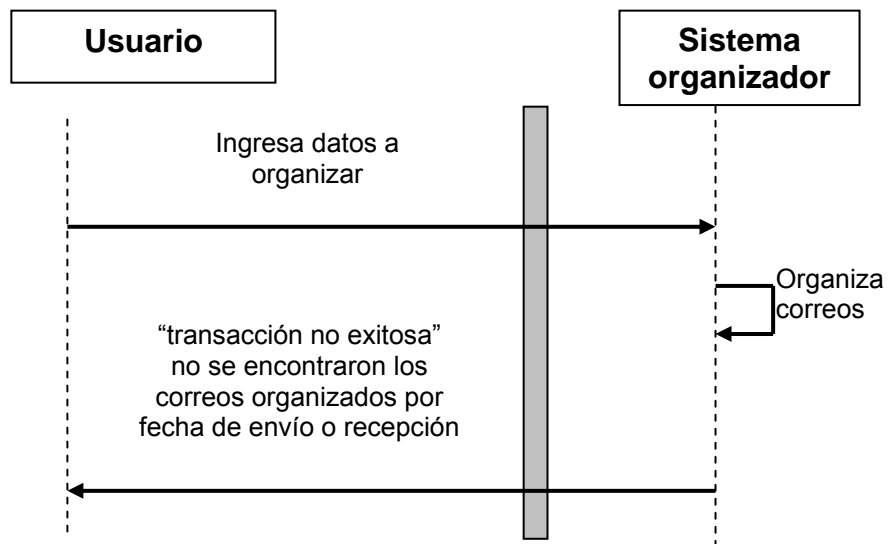


Figura 3.19 Escenario 5.6 Organización no exitosa de los correos electrónicos por fecha de envío o recepción.
Fuente: Autores.

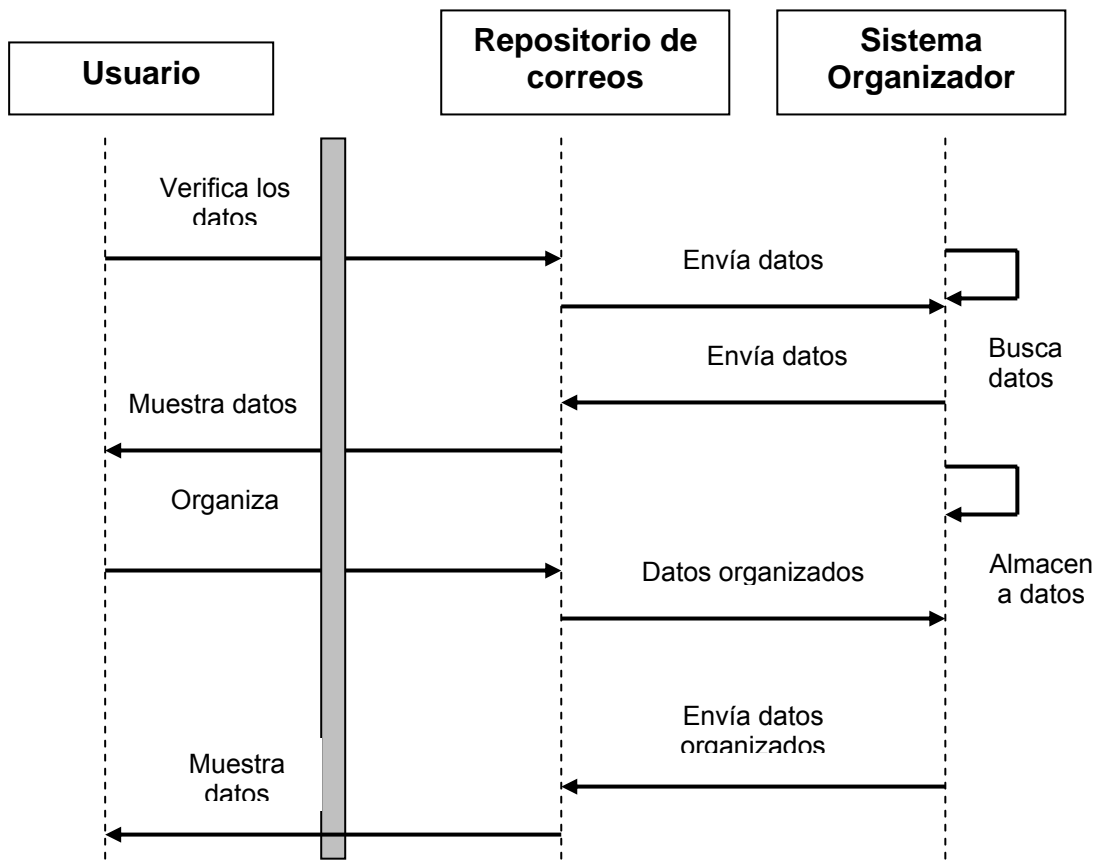


Figura 3.20 Escenario 5.7 Organización exitosa por prioridad.
Fuente: Autores.

Caso de Uso 6: Detectar correos SPAM en el módulo del usuario.

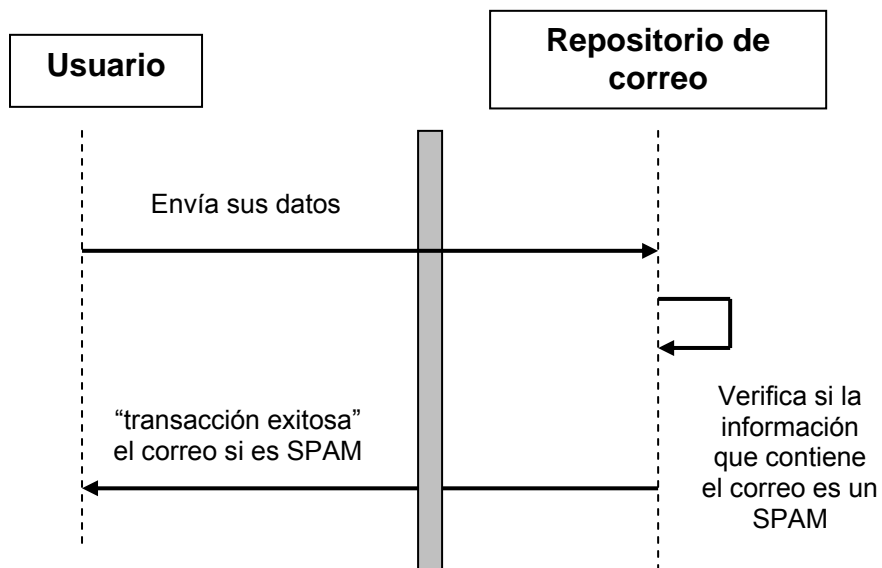


Figura 3.21 Escenario 6.1 Detección exitosa de correo SPAM.
Fuente: Autores.

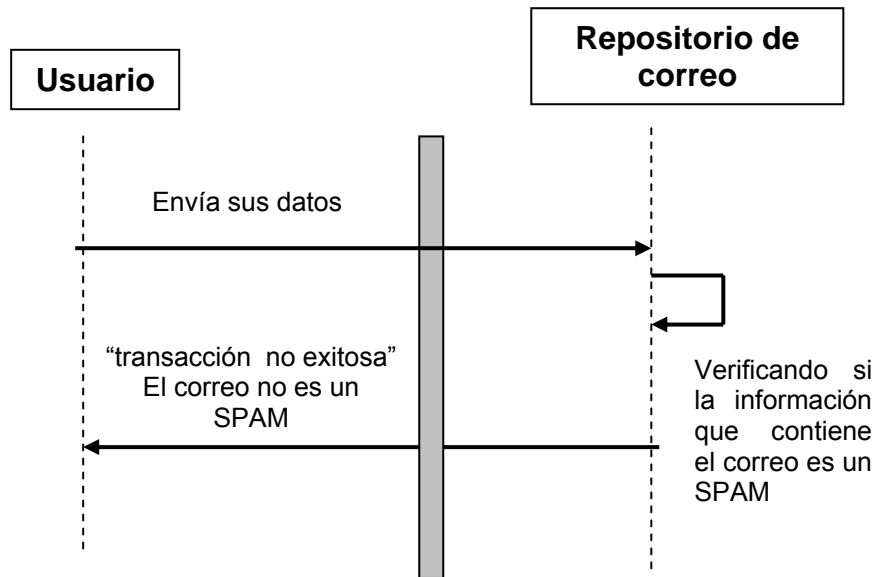


Figura 3.22 Escenario 6.2 Detección no exitosa de correo SPAM.
Fuente: Autores.

Caso de Uso 7: Marcar correo SPAM en el módulo del usuario.

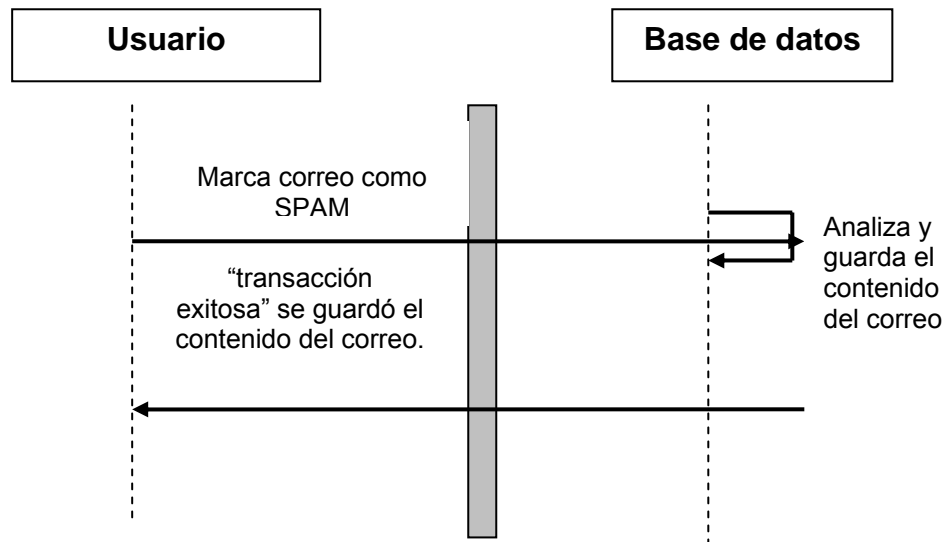


Figura 3.23 Escenario 7.1 Determinación exitosa de correo SPAM.
Fuente: Autores.

Caso de Uso 8: Determinar que contactos indirectos pueden formar parte de la lista de contactos del usuario.

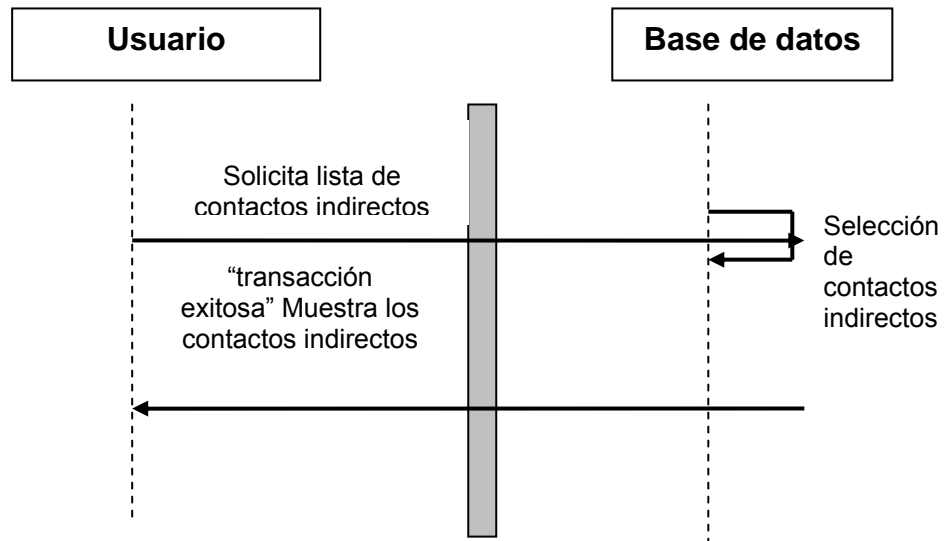


Figura 3.24 Escenario 8.1 Determinación exitosa de contactos indirectos.
Fuente: Autores.

3.1.2 DISEÑO DE BASE DE DATOS

Diagrama entidad - relación

El diseño de la base de datos está representado por medio del diagrama entidad-relación que mostramos a continuación:

3.1.4 DIAGRAMA DE COMPONENTES DEL SISTEMA

Componentes del sistema: El sistema está compuesto por dos módulos, en las cuales los usuarios podrán manipular, clasificar y organizar la información de los correos y usuarios. Estos módulos son:

- Módulo de administración
- Módulo del usuario

Ambos módulos están conformados por tres módulos como lo indica la siguiente figura:

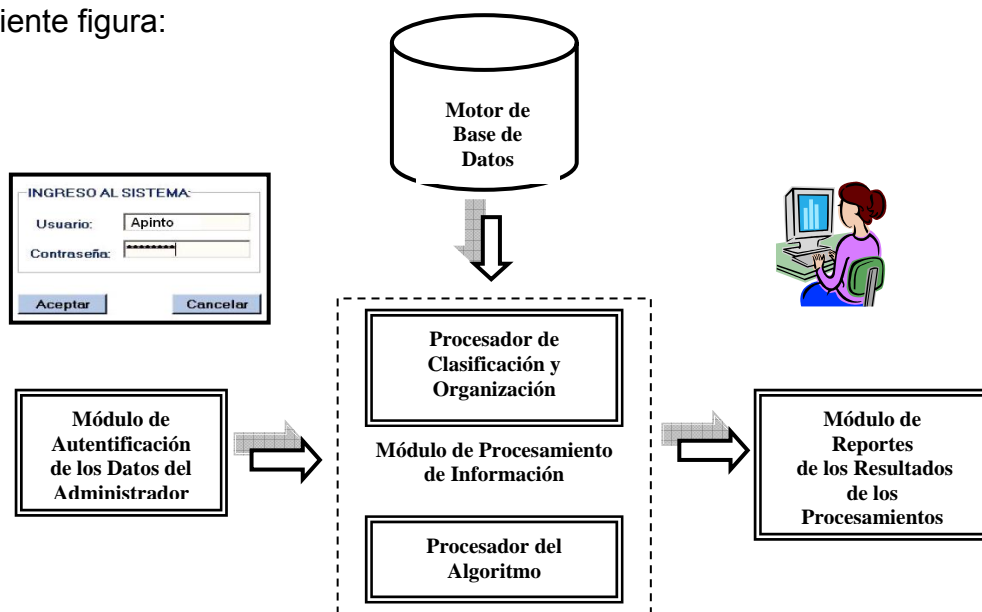


Figura 3.26 Diagrama de los componentes del sistema de los módulos de administración y de usuario.

Fuente: Autores.

La figura 3.26 representa los módulos del sistema en el módulo del usuario y administración en el cual se pueden visualizar 3 módulos importantes:

1. Módulo de autenticación de los datos del administrador
2. Módulo de procesamiento de información
3. Módulo de salida

El Módulo de autenticación de los datos del administrador, contendrá el “usuario” y “contraseña” del administrador o de algún usuario, permitiéndoles manipular los correos correspondientes de los usuarios que se encuentra en

el MAIL SERVER, en el caso de ser administrador. Mientras que si es usuario le permite el ingreso a su correo con el fin de poder manipular la información que contenga el mismo.

El Módulo de procesamiento de información, esta conformado por el procesador de clasificación y organización de correos de la base y el procesador del algoritmo.

El Módulo de salida hace referencia a la interfaz de fácil uso en la cual se muestren las diferentes respuestas del sistema dependiendo de las necesidades del usuario.

3.1.5 DISEÑO DE FLUJO DE INFORMACIÓN

En la siguiente figura hacemos una representación del flujo de información del proyecto, esencialmente de la parte del desarrollo del filtro Anti-SPAM.

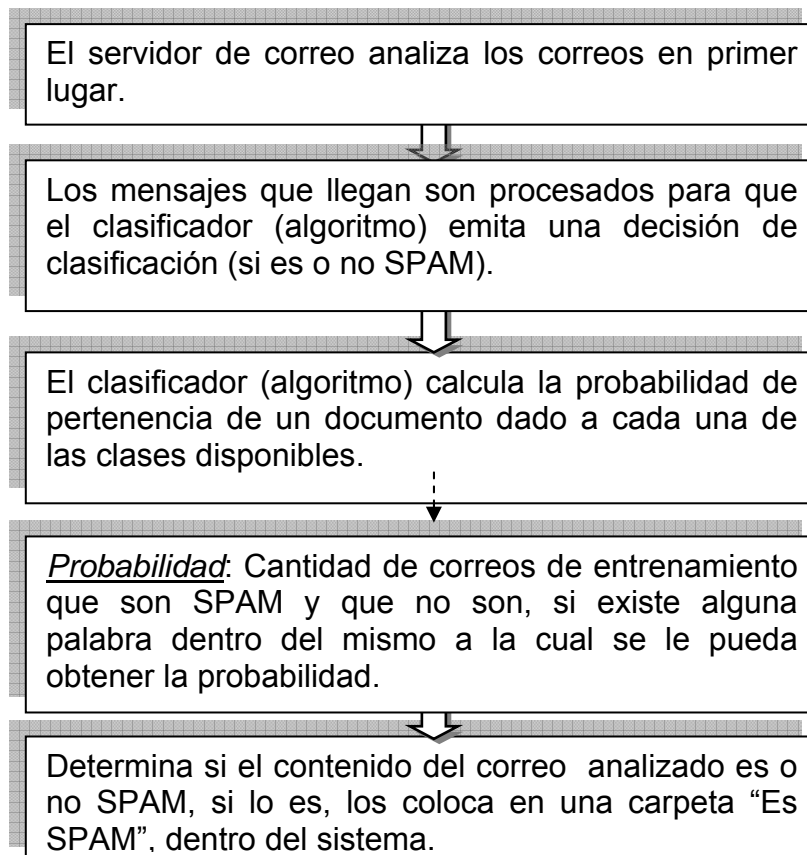


Figura 3.27 Diseño de flujo de información

Fuente: Autores.

3.1.6 DISEÑO DE INTERFAZ DEL USUARIO

Para la elaboración del sistema se ha decidido crear una plantilla principal a nivel de usuario mostrada en el siguiente gráfico, donde tendrá las siguientes opciones:

- Correo
- Calendario
- Tareas
- Contactos

El diseño de esta plantilla fue tomada en base a los diferentes correo electrónicos que existen en la Web, pensando en la facilidad de uso para los usuarios.



Figura 3.28 Pantalla principal

Fuente: Autores.

Para acceder a la pantalla principal los usuarios del sistema tendrán que ingresar su “usuario” y “contraseña”.



Figura 3.29 Pantalla de entrada a cuenta (login)

Fuente: Autores.

Luego de que el usuario haya ingresado correctamente su usuario y contraseña internamente el sistema empieza a realizar el análisis de todos los correos que posea dicho usuario, para determinar cuales son correos SPAMs, además de habilitar el menú en donde el usuario podrá realizar diferentes acciones como: ver su correo, revisar el calendario, agregar o modificar sus contactos y hacer apuntes en la sección de tarea.

El usuario al elegir la opción de “correo”, verá el menú con las siguientes entradas:

- Bandeja de entrada
- Mensajes no deseado
- Borradores
- SPAM
- Mensajes enviados
- Papelería

En la bandeja de entrada se mostrará todos los correos que en el análisis dieron como resultado de que no son correos SPAMs, además el usuario podrá efectuar diferentes acciones como: “colocar en carpeta”, “nuevo”, “eliminar”, “buscar”, “no deseado” “es SPAM”, etc., cada bandeja principal ² posee sus propios botones. En el panel de la lista de correo se muestra las columnas: “de”, “asunto”, “fecha”, “tamaño”.

Una de las acciones que se encuentran en la bandeja de entrada es la de marcar si un correo es SPAM, la cual permite al usuario especificar si algún correos es SPAM, permitiendo así que el filtro auto aprenda.

² Bandeja principal: Esto está explicada detalladamente en el Manual del usuario

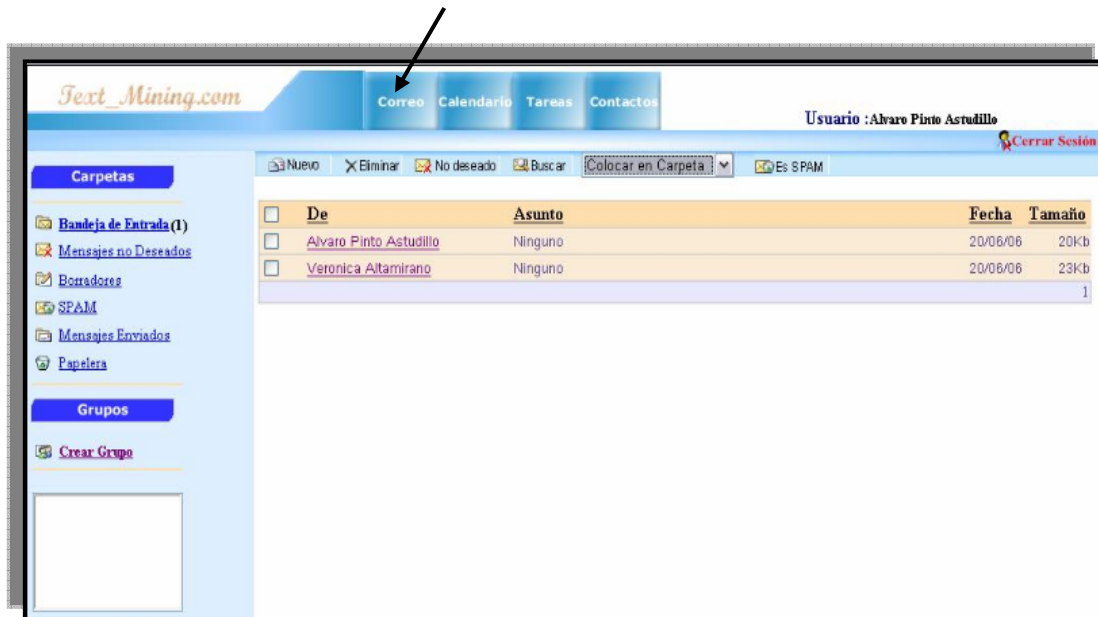


Figura 3.30 Pantalla de bandeja de entrada

Fuente: Autores.

3.2 IMPLEMENTACIÓN

3.2.1 LENGUAJE DE PROGRAMACIÓN

El motivo principal por la cual se optó en utilizar el lenguaje C# para este proyecto denominado: “Minería de texto aplicada a la clasificación y distribución automática de correo electrónico y detección de correo SPAM” fueron por las siguientes consideraciones:

1. Conocimiento previo de la herramienta por los integrantes del grupo.
2. El lenguaje C# ofrece una interfaz amigable y de fácil adaptación con cualquiera de los lenguajes de la plataforma .NET
3. Está respaldado a través de una licencia de una marca reconocida (Microsoft).
4. La funcionalidad de Visual C# es flexible, amigable, intuitiva y sencilla para el usuario final.
5. El desarrollo de las interfaces gráficas es sencilla gracias a los componentes gráficos que tiene disponible.

3.2.2 ALGORITMOS RELACIONADOS CON LA SOLUCIÓN DEL PROBLEMA

En esta parte de este documento vamos a dar a conocer cada uno de los algoritmos que se utilizó para el desarrollo de este proyecto.

MÉTODO BAYESIANO

Utilizamos este método para implementar el filtro Anti-SPAM, a continuación se explica con detalle lo planteado.

ORIGEN

El SPAM ha pasado de ser algo molesto a llegar a ser un verdadero problema por la gran cantidad de correo basura que circula por la red y que nos hace perder tiempo y dinero para eliminarlo.

Para evitar el SPAM no existe una fórmula definitiva ya que antes de abrirlo es difícil saber si un correo es SPAM o no. No obstante, si hay algunas cosas que se puede hacer como usuarios para evitar el SPAM.

MÉTODO A UTILIZAR

En el proyecto se tiene una base de datos, en la cual se han definido dos tablas cuyo contenido son palabras de mensajes legítimos y de mensajes SPAM. Estas dos tablas son: *palabrasSPAM* y *palabrasnoSPAM*,

El procedimiento que se realizó fue separar las palabras o frases más comunes las que son catalogadas como SPAM por ejemplo: VIAGRA, PESO, GRATIS, PRÉSTAMOS y almacenarla en dos tablas llamadas *palabra_SPAM* y *palabra_NoSPAM*, continuación se las compara con el contenido del mensaje,

esto implica sacar probabilidades para determinar si son “buenos” o “malos” según las palabras que contengan.

Los correos detectados como SPAM se los coloca en una carpeta llamada: “es SPAM”, se puede filtrar de diversas maneras pero elegimos esta opción ya que analizan el contenido del correo y son capaces de detectar el SPAM ya que los correos basura suelen tener un lenguaje y una estructura similar.

FILTRADO BAYESIANO DE SPAM: TEOREMA

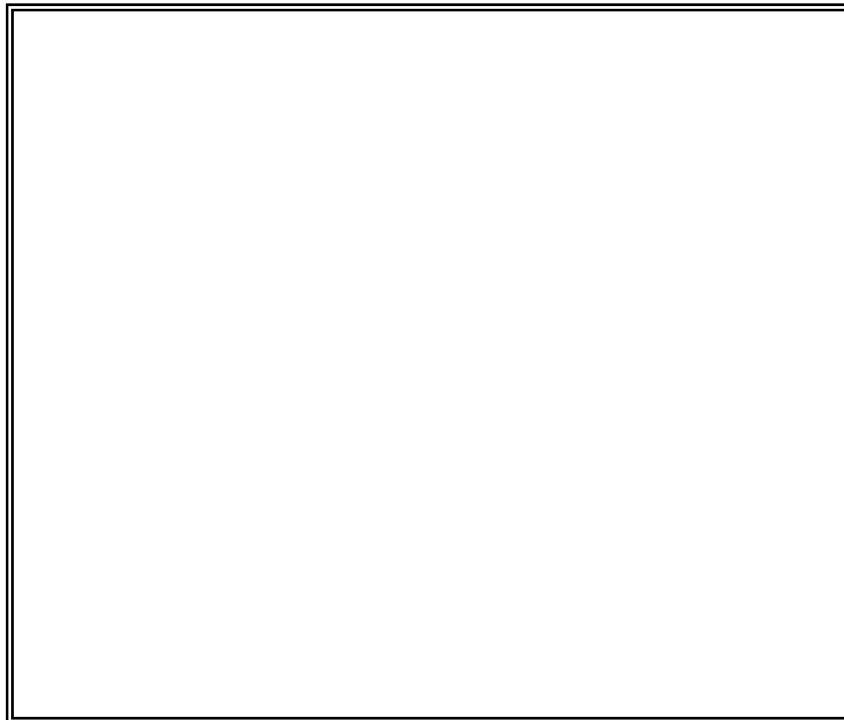


Figura 3.31 Descripción teórica del filtrado bayesiano [10]

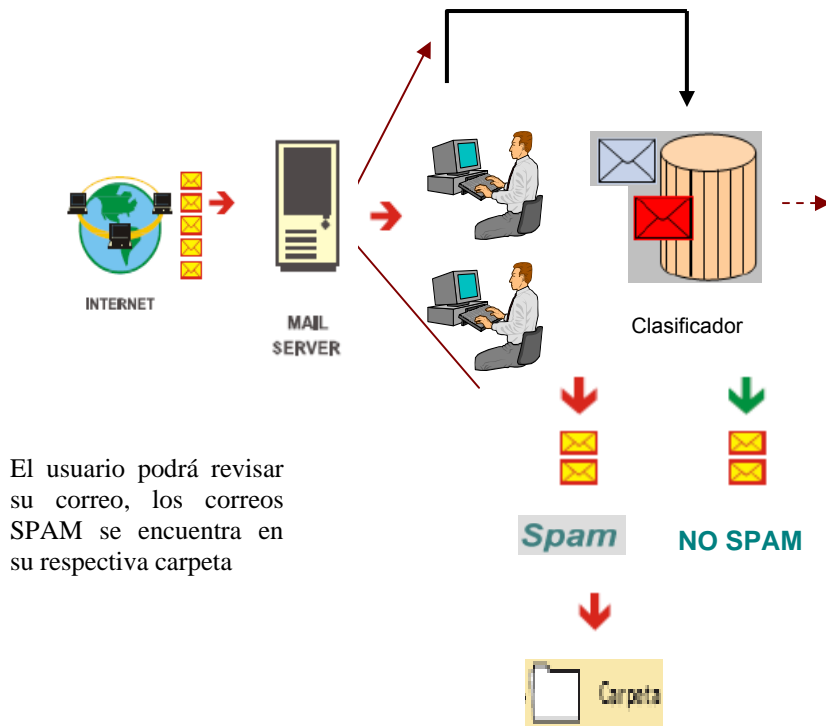


Figura 3.32 Descripción gráfica del filtrado bayesiano

Fuente: Autores.

MÉTODO DE REGLA DE ASOCIACIÓN

Utilizamos este método para analizar 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 más 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 se llama “Contactos indirectos” pueden pertenecer a la 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 el buzón de correos, enviando tipos de correos denominados SPAM para 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 es el número de usuarios agregados a favoritos. Esta matriz de contactos contendrá 1's y 0's de acuerdo a la relación que se tenga entre ellos, 1 para cuando estos contactos tienen relación directa de acuerdo si está 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 3.33. Gráfico representativo de los contactos favoritos

Esta relación se la realiza de frente a los contactos favoritos de todos los usuarios con respecto a los míos. Esta relación tiene una cobertura, la cual está dado para un contacto indirecto determinado como relación directa de uno o más 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 acuerdo 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.

4.- CÍRCULO VIRTUOSO DE LA MINERÍA DE DATOS

Este capítulo, presenta a la minería de datos como un proceso que consta de diferentes fases y en cada una de ellas ubica determinadas técnicas.

Las técnicas relacionadas con minería de datos son una mezcla de estadística, reconocimiento de patrones y algoritmos de aprendizaje.

4.1 ¿QUÉ ES EL CÍRCULO VIRTUOSO DE LA MINERÍA DE DATOS? - FASES

Se considera como un círculo virtuoso al conjunto consecutivo de fases que intervienen en el proceso de minería de datos a una base de datos, los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada. A continuación se presenta el círculo virtuoso, el cual consta de las siguientes fases: [11]

1. Definición del problema
2. Exploración de los datos
3. Preparación de los datos
4. Construcción de modelos
5. Evaluación del modelo
6. Despliegue de los modelos y resultados

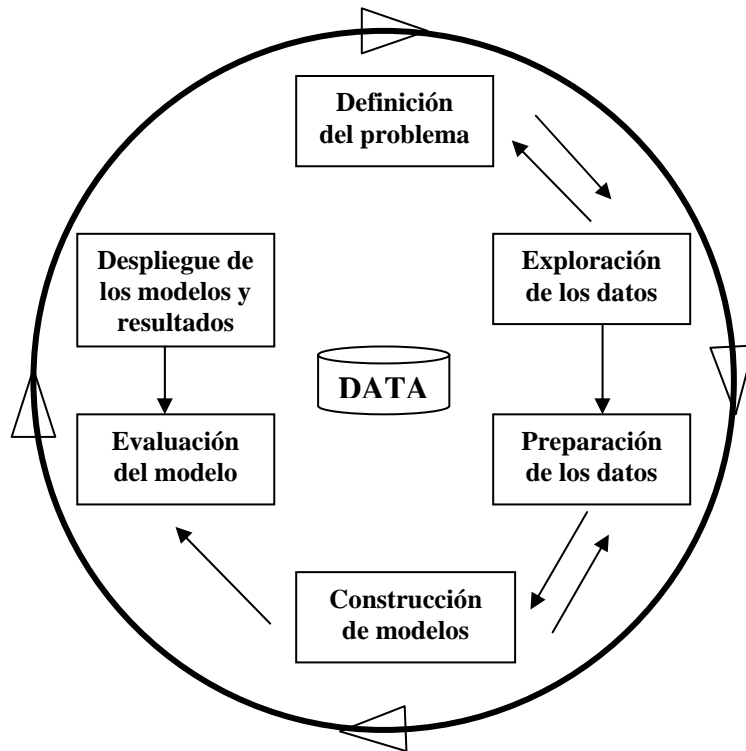


Figura 4.1 Diagrama del proceso de minería de datos

Fuente: Autores

4.1.1 DEFINICIÓN DEL PROBLEMA

- El correo basura, más conocido como SPAM es un problema cada vez mayor, que afecta a todos los que poseen una dirección de correo electrónico.

- **Misión**

Brindar a los usuarios de los correos electrónicos, un filtro Anti-SPAM que les permita recibir correos cuyo contenido sea de interés para el usuario, además de proporcionar una mejor manera de definir cuales contactos indirectos pueden o no pertenecer a su lista de contactos. Finalmente brindar una interfaz para ayuda del administrador en la cual podrá darse cuenta cuáles son los usuarios que reciben más correo SPAM.

○ **Objetivos del proyecto**

A continuación se detalla los principales objetivos que tiene el proyecto:

- Analizar los correos de todos los usuarios en la interfaz del administrador, teniendo como meta principal, determinar el comportamiento de los usuarios con respecto a su poder de uso del correo y tamaño de buzón.
- Detectar correos SPAM, teniendo como meta encontrar patrones críticos en la eliminación de los correos SPAM, y así poder disminuir el ingreso de correo basura en la bandeja de entrada.
- Escoger los contactos indirectos que pueden pertenecer a la lista de contactos.

Cada uno de estos objetivos es desarrollado mediante las técnicas de minerías que serán descritas en este capítulo.

4.1.2 EXPLORACIÓN DE LOS DATOS

Luego de haber establecido el problema a resolver podemos centrarnos en el aspecto principal de minería de datos:

- ❖ Los datos.

Los datos iniciales son obtenidos desde un servidor de correo. Luego de haber adquiridos estos datos varios de ellos serán almacenados en la base de datos en la cual se encuentran las siguientes tablas principales para el proyecto:

- *Correo*
- *Contactos*
- *Contacto_contacto*
- *PalabraSPAM*

- *PalabraNoSPAM*

En las siguientes figuras podrá ver cuales son los atributos y valores que contienen algunas de las tablas antes mencionadas.

id_correo	usuario	destinatario	remitente	cc	cco	asunto
2	jsanche ...	pochacco0380@...	jdcsanch@hotmail...	bcarrill@fiiec.esp...	NULL	Tesis
3	apinto ...	jdcsanch@hotmail...	apinto@guayaq...	pochacco0380@...	NULL	Tesis
5	valtamirano ...	jdcsanch@hotmail...	pochacco0380@...	apinto@guayaq...	NULL	Prueba
6	valtamirano ...	jdcsanch@hotmail...	pochacco0380@...	NULL	NULL	Ninguno
7	jsanche ...	apinto@guayaq...	jdcsanch@hotmail...	NULL	NULL	Ninguno
8	apinto ...	apinto@guayaq...	apinto@guayaq...	NULL	NULL	Ninguno

correo	
🔑	id_correo
	usuario
	destinatario
	remitente
	cc
	cco
	asunto
	prioridad
	contenido
	fecha_recepcion
	fecha_envio
	es_spam
	estado
	tamaño
	carpeta

Figura 4.2 Ejemplo de la descripción de los datos de la tabla *correo*
Fuente: Autores

Como podemos ver en la figura 4.2 se muestra algunos de los atributos de la tabla *correo* los cuales son: *id_correo*, *usuario*, *destinatario*, *remitente*, *cc*, *cco*, *asunto*, entre otros.

usuario	nombre	apellido	correo	departamneto	tipo_acceso	cargo		
apinto	Alvaro	...	Pinto Astudillo ...	apinto@guayaq...	Sistema	...	usuario	Técnico
bcarrillo	Brenda	...	Carrillo	bcarrill@fieci...	Sistema	...	Sistema	Programador
jsanche	Johanna	...	Sanchez	jdcsanch@hotm...	Sistema	...	usuario	Programador
valtamirano	Veronica	...	Altamirano	pochacco0380@...	Sistema	...	usuario	Diseñador
NULL	NULL	...	NULL	NULL	NULL	NULL

contactos	
<input type="checkbox"/>	usuario
<input type="checkbox"/>	nombre
<input type="checkbox"/>	apellido
<input type="checkbox"/>	correo
<input type="checkbox"/>	departamneto
<input type="checkbox"/>	tipo_acceso
<input type="checkbox"/>	cargo
<input type="checkbox"/>	alias
<input type="checkbox"/>	contrasena
<input checked="" type="checkbox"/>	id_contacto

Figura 4.3 Ejemplo de la descripción de los datos de la tabla *contactos*
Fuente: Autores

En la figura 4.3 se muestra algunos de los atributos de la tabla *contactos*, que son: *id_contacto*, *usuario*, *nombre*, *apellido*, *correo*, *departamento*, *cargo*, entre otros.

4.1.3 PREPARACIÓN DE LOS DATOS

Luego de seleccionar los datos iniciales tomamos un subconjunto de ellos para el desarrollo de cada uno de los algoritmos que mencionamos en el capítulo anterior.

Para el desarrollo del filtro Anti-SPAM tomamos como datos principales la tabla *correo* la cual posee el atributo contenido en el cual vamos a realizar el análisis. Además también hacemos uso de las tablas *palabraSPAM* y *palabraNoSPAM*. Finalmente tomamos los datos de la tabla *contactos* para así determinar a cual usuario pertenecen los correos analizados.

En el segundo algoritmo o fase del proyecto, tomamos como datos las tablas *Contactos*, *Contacto_contacto*, para realizar el respectivo análisis con el método de reglas de asociación.

Finalmente en el módulo del administrador para el desarrollo de los dos clasificadores que posee esta interfaz, tomamos como datos iniciales las tablas de *Contactos*, en la cual se encuentra toda la información que necesitamos para utilizar el clasificador de K-mean.

4.1.4 MODELACIÓN

En el capítulo 3 se habló brevemente de los métodos que hemos utilizados en el desarrollo de este proyecto.

Creación de filtro Anti-SPAM.

Para la creación de este filtro se optó por utilizar la técnica de minería de datos llamada métodos bayesianos. A continuación se plantean los pasos que seguimos para el desarrollo y funcionamiento del filtro, con el uso de esta técnica:

- Primero el usuario debe ingresar el usuario y contraseña correctamente en la página inicial del proyecto, una vez realizado esto internamente el sistema selecciona todos los correos que pertenecen a dicho usuario.
- Una vez que el sistema tiene seleccionado sus datos iniciales se procede a realizar el análisis de cada uno de los contenidos de los correos antes seleccionados.
- El análisis de los contenidos se lo realiza de la siguiente manera: primero determinamos cuantos correos posee el usuario que ha iniciado su sesión,

esto nos sirve para crear un bucle y realizar el análisis en todos los correos que posea el usuario.

- Tomamos uno a uno el contenido de cada correo y comenzamos a dividir en palabras o frases para luego ir comparándolas con las que contiene la tablas de *palabraSPAM* y *palabraNoSPAM*, finalmente haciendo uso de las reglas probabilísticas de NAIVE BAYES anteriormente mencionadas, obtenemos la probabilidad de que un correo dado posea un mensaje legítimo o un mensaje SPAM. Todos los correos que han sido determinados como mensajes SPAM, luego del análisis se los coloca en una carpeta llamada SPAM, para que el usuario observe si todos estos correo realmente son SPAM. para una mejor interpretación de la explicación del desarrollo de este filtro se muestra la siguiente figura:

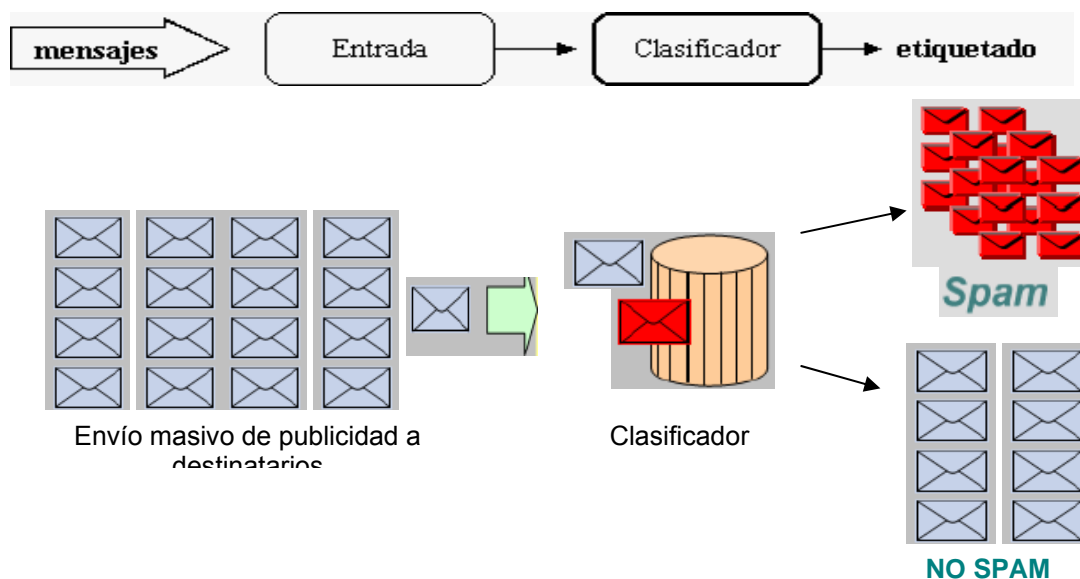


Figura 4.4 Diseño del filtro Anti-SPAM.

Fuente: Autores

- Para que el filtro funcione correctamente y aumente su rendimiento, nosotros creamos una opción en la que el usuario puede marcar como correos SPAM, aquellos que el filtro no haya detectado como tal, lo cual

permitirá que la base de datos se retroalimente con la información que contenga dicho correo, permitiendo que el filtro sea más eficiente.

- Para obtener la nueva información que va a formar parte de la tabla *palabraSPAM* se divide el contenido del correo en frases y palabras y comparamos con las que se encuentran guardadas en las tablas *palabraSPAM* y *palabraNoSPAM*.

Con todos los pasos antes mencionados se podrá obtener un filtro eficiente, teniendo en cuenta que para conseguir esto el usuario deberá proporcionar o marcar los correos SPAM que lleguen a su cuenta electrónica.

Creación de clasificadores en el módulo del administrador.

Para la creación de los dos clasificadores que posee este módulo, se utilizó la técnica de métodos de agrupación o segmentación específicamente el algoritmo de K-mean, que más adelante se explicará brevemente. A continuación se plantean los pasos que seguimos para el desarrollo y funcionamiento de los clasificadores, con el uso de esta técnica:

- Primero estos clasificadores fueron desarrollados como ayuda para el administrador del MAIL SERVER, para que el administrador pueda hacer uso de ellos debe ingresar su usuario y contraseña correctamente en la pantalla principal del proyecto.
- El administrador podrá escoger entre dos tipos de clasificadores, los cuales son: clasificador por uso de correo y clasificador por tamaño de buzón. Sea cual sea el clasificador que seleccione el administrador estos realizarán el mismo proceso que a continuación vamos a detallar: Primero el sistema

obtiene la información de cuantos usuarios contiene el servidor de correo, luego verifica cuantas carpetas poseen los usuarios.

- Una vez obtenida esta información procedemos a formar grupos utilizando el algoritmo de K-mean, el cual primero obtiene una matriz $n \times m$ donde n son todos los usuarios que posean una cuenta de correo y m son todas las carpetas que poseen los usuarios.
- “El algoritmo es el siguiente:
 - Seleccionar G puntos como centros de los grupos iniciales. Esto puede hacerse:
 - Asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos así formados;
 - Tomando como centro lo G puntos más alejados entre sí;
 - Construyendo unos grupos iniciales con información a priori y calculando sus centros”[17].

En nuestro caso utilizamos este último literal construimos grupos iniciales con información a priori y calculamos sus centros.

- “Calcular las distancias euclídeas de cada elemento a los centros de los G grupos, y asignar cada elemento al grupo de cuyo centro esté más próximo”[17].
- “Definir un criterio de optimalidad y comprobar si reasignando alguno de los elementos mejora el criterio”[17].
- Si no es posible mejorar el criterio de optimalidad, termina el proceso.

“El criterio de homogeneidad o de optimalidad, que se utiliza en el algoritmo de K-mean es minimizar la suma de cuadrados dentro de los grupos (SCDG) para todas las variables”[17], dada por:

$$SCDG = \sum_{g=1}^G \sum_{j=1}^p \sum_{i=1}^{n_g} (x_{ijg} - \bar{x}_{jg})^2 \quad [17]$$

Donde:

x_{ijg} Es el valor de la variable j en el elemento i del grupo g.

\bar{x}_{jg} La media de esta variable en el grupo.

Y finalmente, para determinar en este proyecto cuantos grupos deben formarse con la información que tenemos hacemos uso del TEST F, el cual se define con la siguiente fórmula:

$$F = \frac{SCDG(G) - SC DG(G + 1)}{SC DG(G + 1)/(n - G - 1)} \quad [17]$$

Luego de pasar por el TEST y cumpliendo el criterio de optimalidad el administrador observará los grupos que se forman de acuerdo a lo solicitado en el caso de uso correo se mostrarán cuales son los usuario que mayor correo envían o reciben incluido los correos SPAM.

En el otro clasificador el tamaño de buzón que están utilizando.

4.1.5 EVALUACIÓN DEL MODELO

En las fases previas (sobre todo en la de modelación), la evaluación se refería a la exactitud y generalidad del modelo generado, mientras que en esta fase involucra la evaluación del modelo con respecto a los objetivos del proyecto.

En esta fase se debe decidir si hay o no razones para construir un modelo deficiente (relación costo - beneficio), si es aconsejable probar el modelo en un problema real. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable calificar el modelo con relación a otros objetivos diferentes a los originales?, esto podría revelar información adicional.

Revisión del proceso: Se refiere a calificar al proceso entero de minería de datos con la idea de identificar elementos que pudieran ser mejorados.

Por último, en esta fase se toma una decisión acerca de futuras fases. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría decidirse pasar a la fase de despliegue de resultados, sino, podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de minería de datos.

4.1.6 DESPLIEGUE DE LOS MODELOS Y RESULTADOS

En esta fase se define una estrategia para desplegar los resultados de la minería de datos.

Monitoreo y mantenimiento: Si los modelos resultantes del proceso de minería de datos son desplegados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitoreo y mantenimiento para ser construidas sobre los modelos. La

retroalimentación generada por el monitoreo y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

Reporte final: Es la conclusión del proyecto de minería de datos. Resume los puntos importantes del proyecto, la experiencia ganada y explica los resultados producidos.

5. ANÁLISIS DE FACTIBILIDAD Y COSTO

5.1 DESCRIPCIÓN DE LOS SERVICIOS DEL SISTEMA

El sistema ofrece un filtro Anti- SPAM fácil de utilizar para la detección de correo SPAM en gran cantidad; proporcionando muchos beneficios para el usuario.

- **Detección instantánea y preventiva del SPAM mediante filtro bayesiano.**

Nuestro sistema proporciona una detección instantánea de correo SPAM, desde que el usuario ingresa exitosamente a su cuenta electrónica.

- **Disminución del tiempo de lectura en correo SPAM**

El filtro Anti-SPAM detecta el correo SPAM en el momento que el usuario ingresa a su cuenta de correo, permitiendo así que el usuario no pierda tiempo leyendo correos que no posean información importante para él; *esto quiere decir que los correos SPAM no pueden llegar al buzón si el usuario así lo decide.*

- **Disminución del espacio de buzón**

El filtro Anti- SPAM le permite al usuario eliminar de su cuenta todos los correos que fueron detectados como correo SPAM sin tener que leerlos, además el usuario también podrá marcar como correo SPAM a los correos que el considere como basura.

5.2 COSTOS DEL SISTEMA

5.2.1 COSTOS DE DESARROLLO

Los costos de los artículos utilizados para el desarrollo del sistema son:

Artículo	Costo Individual	Cantidad	Costo Total
Equipos	\$ 800	3 computadores	\$2400
Gastos de oficina	\$ 400	3 meses	\$1200
Software	\$ 400		\$ 400
Sueldos	\$ 350	3 personas por 3	\$3150
Total			\$7150

Tabla 4 Costos para el desarrollo del sistema

Fuente: Autores

CALCULO DEL TIR:

$$0 = -2000 - 750((1-(1+i)^{-3})/i) + 10000(1+i)^{-4} - 200(1+i)^{-5} - 200(1+i)^{-6}$$

$$TIR = 30.5\%$$

Dado que se obtuvo una TIR positiva y mayor a la tasa mínima que se ganaría si se invirtiera el dinero en alguna entidad bancaria, se concluye que el proyecto es factible de realizar

5.2.1 COSTOS DE IMPLEMENTACIÓN

El costo de implantación de este prototipo en un ambiente real involucraría los siguientes rubros:

Tipo de Gasto	No.	Costo	Costo Total
Equipos Servidor	1	\$3,000	\$3,000
Licencia Base de Datos(DB2)X Servidor	1	\$ 624	\$624
Licencia Servidor Aplicación Web	1	\$4,260	\$4,260
Unidades de Respaldo	2	\$ 160	\$320
Total			\$8,204

Tabla 5 Costos de implantación del sistema

Fuente: Autores

5.3. ANÁLISIS DE VIABILIDAD

Realizar un estudio de análisis de viabilidad es muy importante, ya que por medio de él sabremos con exactitud si un proyecto es factible o no. Este proyecto toma en cuenta los siguientes análisis.

5.3.1 ANÁLISIS COSTO - BENEFICIO

El análisis costo-beneficio nos permite definir con mayor factibilidad todas las alternativas planteadas para la realización del proyecto.

COSTOS

Los enlaces Internet y su tamaño (ancho de banda), así como la tecnología relacionada, representan costos cada vez mayores para las organizaciones. Es de suponer entonces que el SPAM se haya convertido en los últimos años en un problema que aqueja financieramente y operativamente a las empresas y los usuarios.

“El SPAM representa para las empresas incrementos en los costos de telecomunicaciones, almacenamiento de información e inversión de tiempo en recurso humano para su análisis y eliminación. Según Ferris Research en el año:

2002: El costo por SPAM alcanzó los 8.9 millones de dólares para las empresas estadounidenses

2003: Esta cifra se incrementó a 10 millones.

En términos de números de mensajes, la compañía “MessageLabs” estimó en 50.5% el porcentaje de SPAM en el tráfico de correspondencia por Internet. Lo relevante de esta cifra es el crecimiento exacerbado registrado y la posibilidad que se incremente durante el 2004” [12]

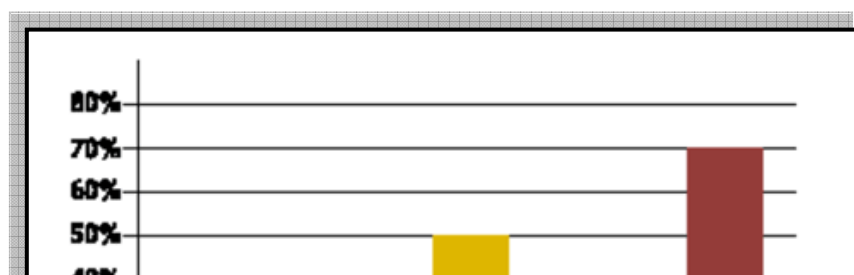


Figura 5.1 Porcentaje de SPAM en el servicio de correo electrónico. [12]

En los costos que involucra esta lucha contra el SPAM, hay cuatro elementos importantes:

1. Costos relacionados con la dedicación que el personal debe realizar para eliminar el correo no deseado de sus buzones. La mayor parte de los especialistas en el tema establecen una media de 3 segundos por correo.
2. El consumo de recursos de las redes corporativas: ancho de banda, el uso del Internet por medio del MODEM, espacio de disco, saturación del correo.

Ancho de banda: El costo de la conexión de 2.5Gbit/s, además de telefonía y tv que ofrecen, tiene un valor mensual de \$85 dependiendo de la compañía. Existen otras compañías que dan 1.2 ,2.4 y 5 Megas con módem e instalación, espacio web, mensualmente deben de pagar desde \$32. También oscilan valores de \$11.990 con 128 Kbps, \$ 14.990 con 200 Kbps, \$18.990 con 600 Kbps etc. [13]

Acceso telefónico (MODEM): Si los usuarios revisan sus correos por este medio, el costo de la tarifa del teléfono estará alrededor de los \$30.40 o más según el uso que tenga.

3. Costos relacionado con las labores de soporte técnico y atención de emergencias de seguridad como consecuencia de la introducción de virus o software “spyware” en las estaciones de trabajo de los empleados.
4. Finalmente, en evitar la saturación de los servidores ante tal carga de SPAM, además de la infraestructura adicional, como agregar nuevos servidores, o comprar herramientas que aminoren y controlen dicho mal.
[14] , [15]

<u>SERVIDORES</u>	<u>COSTOS</u>
HP-PROLIANT ML110 G3 P3,2	\$ 561,00
HP SVR TWR ML110 G3 P3.2Ghz 800HT 512MB 80GB SATA 2MB CACHE	\$ 858
HP SVR TWR ML350 G4 X3.2Ghz 800HT 1GB 2MB CACHE SA641	\$ 2412
Intel SE7520BD2SCSID2	\$ 570
Intel SE7520JR2SCSID1	\$ 653,13
Intel SE7520AF2	\$ 706,56

Tabla 6 Listas de precios de servidores [14], [15]

Como se puede observar los principales costos que intervienen en este proyecto son los que a continuación presentaremos, colocando junto a ellos una ponderación entre 1 a 10 donde el 1 es un costo mínimo y el 10 un costo

máximo, esta ponderación se la realiza para determinar la proposición de valor del proyecto. De la misma manera se hace con los beneficios.

- Costos Monetarios : \$ 10'000.000 aprox. **(9)**
- Tiempo : 3 segundos por correo **(9)**

BENEFICIOS

Entre los beneficios que nos brindan el uso del Anti- SPAM son los siguientes:

- **Funcionales**
 - No necesitas guardar los mensajes **(8)**
 - a. Privacidad
 - b. Espacio en disco
 - Detección instantánea y preventiva del SPAM mediante filtro bayesiano. **(7)**
- **Emocionales**
 - Reduce el trabajo del administrador y permite ajustar el filtrado de SPAM. **(6)**
 - No hay que leer tantos mensajes de SPAM. **(7)**
 - Mejora la satisfacción del usuario permitiendo borrar los correos que son catalogados como SPAM sin necesidad de leerlos. **(7)**

Proposición de valor = Beneficios / Costos

$$= 35 / 18 = 1,94$$

Como se pueda observar el resultado obtenido de la proposición de valor es mayor a 1, lo cual nos indica que la implementación de este proyecto, disminuirá los costos producidos por el correo SPAM.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- Se optimizó la capacidad de clasificar el contenido de los correos recibidos como SPAM ó no, permitiendo así que el espacio de disco no sea ocupado por correos basuras; Este problema se lo pudo resolver de manera optima con el algoritmo de NAIVE BAYES, mediante el filtro bayesiano que se creo, el cual se volverá más eficiente al adquirir la información necesaria de correos SPAM y no SPAM que el usuario le proporcione.
- Se provee al administrador de correo una herramienta para detectar cuales son los usuarios que mayor cantidad de correos SPAMs reciben o que ocupan demasiado espacio de disco, permitiendo así que el administrador pueda tomar las medidas necesarias, como la de enviar a los usuarios un mensaje de advertencia sobre la cantidad de correo SPAM recibido o el espacio de disco duro que esta utilizando; otra medida podría ser el bloqueo de la cuenta del usuario.
- También proporcionamos al usuario conocer los tipos de contactos que no pertenecen a su lista de favoritos (contactos indirectos), pero de una u otra manera está relacionada con sus contactos más afines ya sea porque se encuentran en sus listas de contactos o estén relacionados por los correos.

Recomendaciones

Para un mejor rendimiento del sistema, se recomienda a los usuarios lo siguiente:

1. Marcar como SPAM los correos recibidos que el usuario considere como correo "basura", para que el filtro adquiriera conocimientos de estos mismos incrementando así su funcionalidad.

ANEXO A

MANUAL DEL USUARIO

Los programas para manejar el correo son muy parecidos todos, manejan los mismos conceptos y lo que más cambia es el aspecto exterior, y el tipo de configuración que permiten.

1.- INGRESO AL SISTEMA

Nuestro servidor de correo nos pedirá el nombre de inicio de sesión es decir:

- El nombre de usuario y la
- Contraseña.

Escribimos los datos solicitados y pulsamos en el botón “*ingreso*”

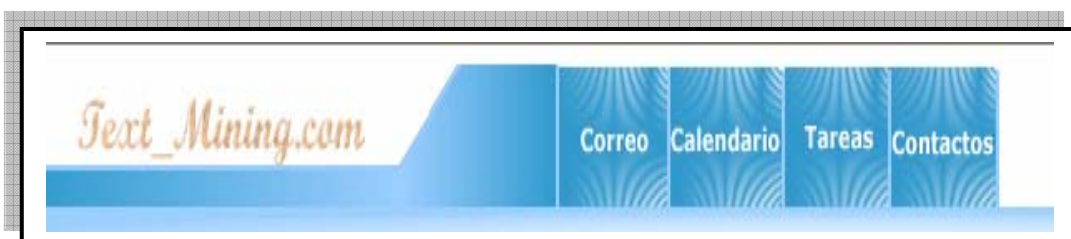


The image shows a login form for Text_Mining.com. At the top left is a logo consisting of a blue circle with an '@' symbol inside. To the right of the logo, the text 'Text_Mining.com' is written in a brown, cursive font. Below this, the text 'Topico de Minería de Datos' is written in a smaller, brown, cursive font. The form contains two input fields: 'Usuario :' followed by a white rectangular box, and 'Contraseña :' followed by another white rectangular box. At the bottom center of the form is a blue button with the word 'Ingresa' written in white, cursive text.

2.- LA INTERFAZ DEL SISTEMA

La ventana de acceso a nuestro correo contiene 4 solapas llamadas barra de navegación:

- 2.1 **Correo** (bandeja de entrada): En ella se depositan todos los mensajes que recibimos. Si queremos leer algún mensaje debemos ir a esta solapa.
- 2.2 **Calendario**: El calendario sirve para llevar un diario de las cosas que tenemos que hacer. Si queremos hacer un diario debemos ir a esta solapa.
- 2.3 **Tareas**: Una tarea es una actividad a la cual se desea efectuar un seguimiento hasta su finalización. Si queremos hacer una tarea debemos ir a esta solapa.
- 2.4 **Contactos**: En ella podemos almacenar todas las direcciones de correo que nos interesen Si queremos ingresar o ver algunos detalles de un contacto debemos ir a esta solapa.



Los correos se organizan en carpetas (bandejas) o buzones (box), en la parte izquierda tenemos las siguientes bandejas principales:

CARPETA → Contendrán las 6 bandejas principales:

Bandeja de entrada: En esta bandeja se colocan los correos que se reciben.

Mensajes no deseados: En esta carpeta están los correos catalogados como no deseados para luego borrarlos.

Borradores: En esta carpeta están los correos que hemos escrito pero que todavía no les hemos dado la orden de enviar.

SPAM: En esta carpeta están los correos catalogados como correos basura.

Mensajes enviados: Por defecto, hace una copia de todos los correos que mandamos y los almacena en esta carpeta.

Papelera: En esta carpeta contiene todo que haya sido eliminado es decir mensajes, contactos, grupos, tareas, notas.

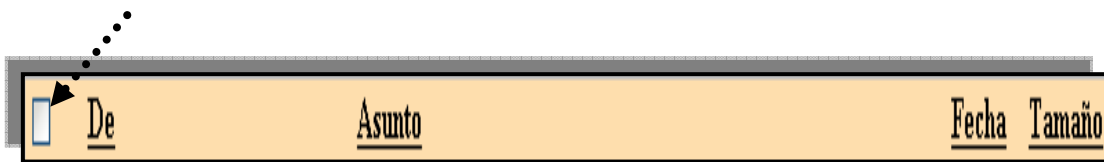
GRUPO → Contendrá

Crear grupo: Nos permite formar con las direcciones de los contactos grupos.



FUNCIONES GENERALES DE LA INTERFAZ

Este botón permite aplicar una acción a varios mensajes seleccionados o a todos (marcando el cuadro que está justo a la izquierda de “de”):



Ordenar por campo: Si pulsamos una vez sobre el panel, la ordenación será en sentido ascendente pero si pulsamos otra vez lo hará en sentido contrario, es decir, descendente. Para volver al orden ascendente basta con volver a pulsar sobre el mismo icono.

<input type="checkbox"/>	<u>De</u>	<u>Asunto</u>	<u>Fecha</u>	<u>Tamaño</u>
<input checked="" type="checkbox"/>	Alvaro Pinto Astudillo	Ninguno	20/06/06	20Kb
<input type="checkbox"/>	Veronica Altamirano	Ninguno	20/06/06	23Kb
				1

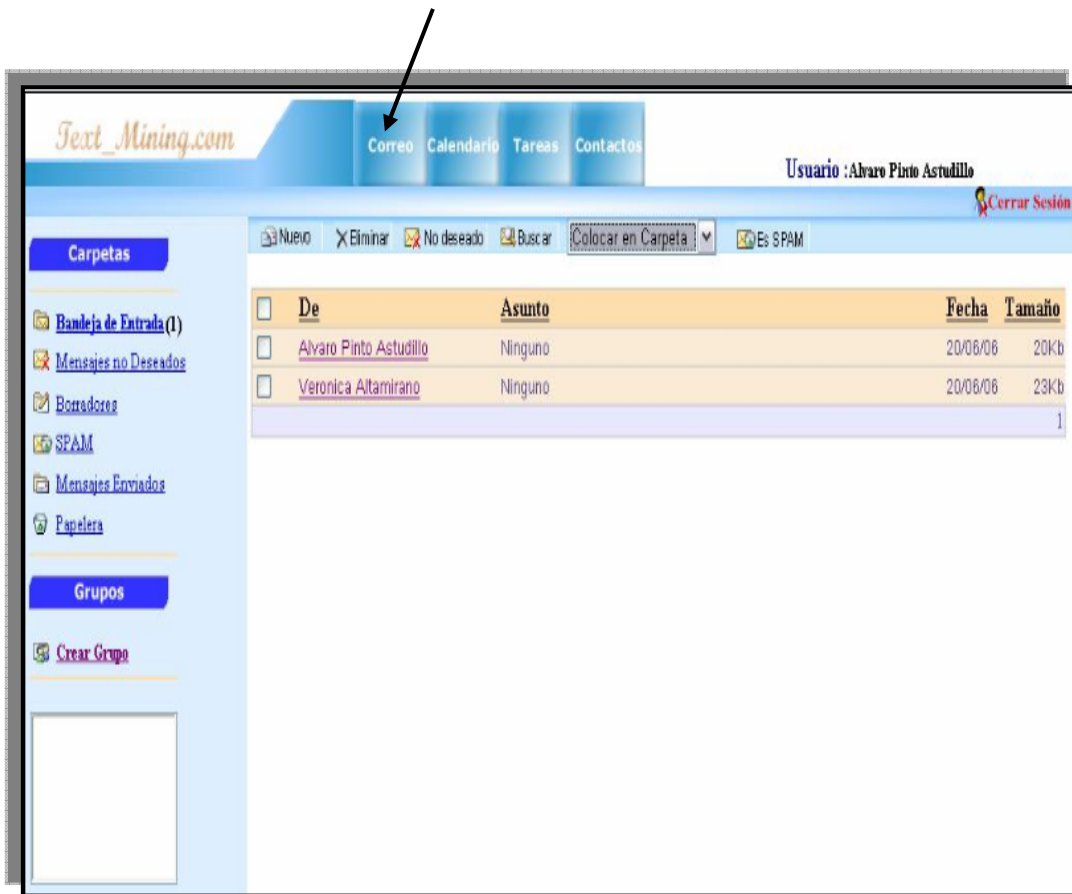
<input type="checkbox"/>	<u>De</u>	<u>Asunto</u>	<u>Fecha</u>	<u>Tamaño</u>
<input type="checkbox"/>	Veronica Altamirano	Ninguno	20/06/06	23Kb
<input checked="" type="checkbox"/>	Alvaro Pinto Astudillo	Ninguno	20/06/06	20Kb
				1

2.1 CORREO → (BANDEJA DE ENTRADA)

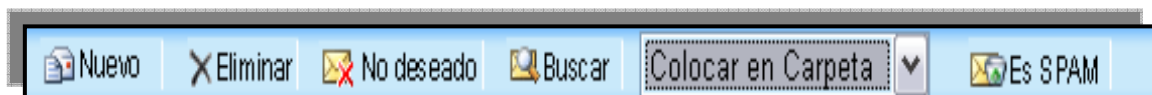
Haga clic en “correo” para acceder a su bandeja. En la parte central puedes ver los correos que han llegado a la cuenta o sea lo que hemos recibido. [1]

- Tanto los mensajes viejos (es decir que ya hemos leído)
- Como los nuevos (es decir los que no hemos leído).

Para leer un mensaje recibido (ya sea nuevo o antiguo) basta picar sobre él con el ratón.



A continuación antes de que empiece el listado de los mensajes, está la *barra de herramientas* con los comandos más habituales para esa bandeja como son: **nuevo, eliminar, no deseado, buscar, colocar en carpeta, es SPAM.**



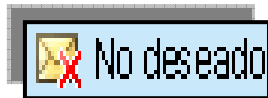
- ✓ **Nuevo:** Permite crear un nuevo correo a la persona que corresponda a la dirección seleccionada. [2]



- ✓ **Eliminar:** Permite eliminar correos, es decir pasa el correo a la “papelera”. [2]



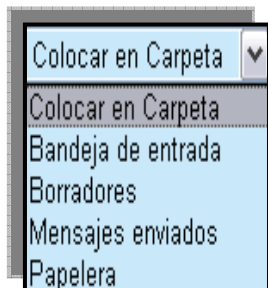
- ✓ **No deseado:** Permite enviar los correos a esta carpeta para luego borrarlos.



- ✓ **Buscar:** Permite buscar correos en las bandejas principales, se puede hacer de diferentes formas la búsqueda por ejemplo en los campos “de”, “para”, “asunto”, “en todo el mensaje”. [4]



- ✓ **Colocar en Carpeta:** Permite al usuario realizar la opción de regresar uno o varios correos a una de las bandejas principales como se muestra en el gráfico. [5]



- ✓ **Es SPAM:** Permite enviar los correos seleccionados a la carpeta “es SPAM”. Esta carpeta contendrá los correos catalogados como correos basura (SPAM).



1.1..1 CARPETA

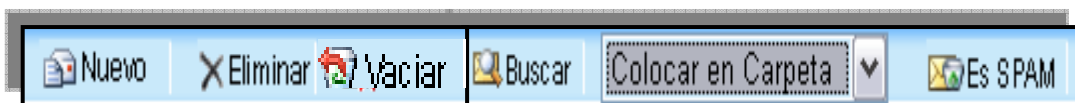
2.1.1.1 BANDEJA DE ENTRADA

En esta sección se listan todos los mensajes que hemos recibido como se explicada en el punto anterior. Véase en [1].

2.1.1.2 MENSAJE NO DESEADO

La carpeta de "mensajes no deseado" permite que los correos provenientes de las direcciones que tu indiques sean dejados en esta carpeta, para después borrarlos.

En esta ventana, antes de que empiece el listado de los mensajes tenemos 6 botones principales como son: **nuevo** véase en [2]; **eliminar** véase en [3]; **Buscar** véase en [4]; **colocar en carpeta** [5]; **vaciar, no es SPAM**



- ✓ **Vaciar:** Permite vaciar el contenido de la carpeta.[6]



- ✓ **No es SPAM:** Permite regresar los correos que no son catalogados como SPAM a la "bandeja de entrada".



2.1.1.3 - BORRADORES

Cuando redactamos un correo pero no lo queremos enviar todavía, lo guardamos en borradores para más adelante acabar de redactarlo y enviarlo

En esta ventana tenemos 4 botones principales como son: **nuevo** véase en [2]; **eliminar** véase en [3]; **buscar** véase en [4]; **colocar en carpeta** [5]



2.1.1.5 - MENSAJES ENVIADOS

En esta ventana se encuentran copias de los correos que el usuario ha enviado. Antes de que empiece el listado de los mensajes tenemos 4 botones principales como son **nuevo** véase en [2]; **eliminar** véase en [3]; **buscar** véase en [4]; **colocar en carpeta** [5]



2.1.1.6 - PAPELERA

Esta carpeta contiene todo aquello que haya sido eliminado, ya sean mensajes, contactos, tareas, notas, etc. En esta carpeta disponemos de una opción muy interesante que nos sirve para eliminar completamente el contenido de la carpeta sin necesidad de seleccionar los mensajes. Haciendo clic sobre el botón **vaciar** borra el contenido de esta carpeta, al pulsar sobre esta opción nos aparece un cuadro de diálogo solicitando confirmación de vaciado, pulsamos sobre **sí** y la carpeta queda totalmente vaciada.

También se puede volver a colocar los correos antes de vaciarlo a las diferentes bandejas usando la opción de “colocar en carpeta”

En esta ventana, antes de que empiece el listado de los mensajes tenemos 4 botones principales. **nuevo** véase en [2]; **buscar** véase en [4]; **colocar en carpeta** [5], **vaciar** [6]



2.1.2 GRUPO

2.1.2.1 CREAR GRUPO

Nos permite crear y agrupar direcciones de los contactos en grupos categorizados por ejemplo, tener las direcciones de la empresa, por otro lado tener las de amigos, familiares en lo que se denomina GRUPOS.

La pantalla tendrá los siguientes campos:

- **Nombre del grupo** → El nombre característico del grupo que desea formar.
- **Todos los contactos** → Lista de las direcciones de correos de los contactos del usuario.
- **Miembros del grupo** → Direcciones de correos agregadas al nuevo grupo.

En esta pantalla, se tienen los siguientes botones:

- ✓ **Guardar:** Permite guardar las direcciones que se encuentra en el grupo formado.
- ✓ **Cancelar:** Permite cancelar la acción de guardar los correos y el nombre del grupo formado.
- ✓ **Buscar:** Permite buscar la dirección de unos de sus contactos.



2.2 CALENDARIO

El calendario sirve para llevar un registro de las acciones que elija relacionadas con los contactos y coloca las acciones en una vista de escala de tiempo, es decir le va a permitir al usuario trabajar con más orden, ya que se comporta como si fuera una agenda o diario personal, donde podrá apuntar día a día, las acciones que desea realizar dentro de un tiempo especificado por él, donde también podrá verificar apuntes hechos en años anteriores.

2.3 TAREAS

La tarea es una actividad de las obligaciones planteada por el usuario al cual se le debe realizar un estudio continuo hasta su culminación.

En ella se anotará con detalle la descripción de la tarea que se ha propuesto cumplir en un tiempo determinado. Esta pantalla tiene los siguientes campos:

- **Contactos** → Dirección del correo electrónico del contacto.

- **Asunto** → Descripción de la tarea
- **Comienzo** → Fecha del comienzo de la tarea a realizar.
- **Fecha de vencimiento** → Fecha del vencimiento de la tarea a concluir.
- **Estado** → Perfiles de la tarea, esta puede ser:
 - No iniciada
 - En proceso
 - Finalizada
- **% Completado** → Ejecución de la tarea en porcentaje (0%, 25%, 50%, 75%, 100%)
- **Prioridad** → Prioridad que se le da a la tarea a efectuar (media, alta y baja)

En esta pantalla, se tienen la siguiente barra de herramientas:

- ✓ **Guardar:** Permite guardar cambios realizados, tanto como cuando se ingresa un dato específico de la tarea o cuando se realiza alguna modificación sobre la misma.
- ✓ **Cancelar:** Permite cancelar la acción de guardar los datos de la tarea a realizar.
- ✓ **Adjuntar:** Permite colocar un fichero (.GIF, ZIP, DOC) que será enviado junto con la tarea.

2.4 CONTACTOS

Nos permite tener almacenadas todas las direcciones de correos que utilizamos con más frecuencia. Esta pantalla permite mostrar al usuario la lista de direcciones de sus contactos con sus respectivos detalles ya sean estos: nombre, correo electrónico [7].

En la vista inferior izquierda está formado por las siguientes opciones:

CONTACTOS → Contendrá: Nuevo contacto

MOSTRAR INFORMACIÓN → Contendrán “todos los contactos” y “grupos”

CONTACTOS INDIRECTOS → Contendrán “por mis contactos” y “por mis correos”

Cuando necesitemos escribir una dirección podemos hacerlo muy fácilmente si previamente la hemos introducido en la lista de contactos. Por ejemplo, para crear un correo nuevo, basta con señalar el correo y hacer clic en el botón

enviar correo para que se abra la ventana de correo nuevo con el campo dirección ya relleno.

En esta pantalla, se tienen la siguiente barra de herramientas: **buscar** véase en [4]



- ✓ **Nuevo:** Sirve para añadir nuevas direcciones sin necesidad de salir de la pantalla. [8]



- ✓ **Eliminar:** Permite eliminar un contacto creado. [9]



- ✓ **Modificar:** Permite realizar alguna modificación sobre los datos de algunos de los contactos. [10]

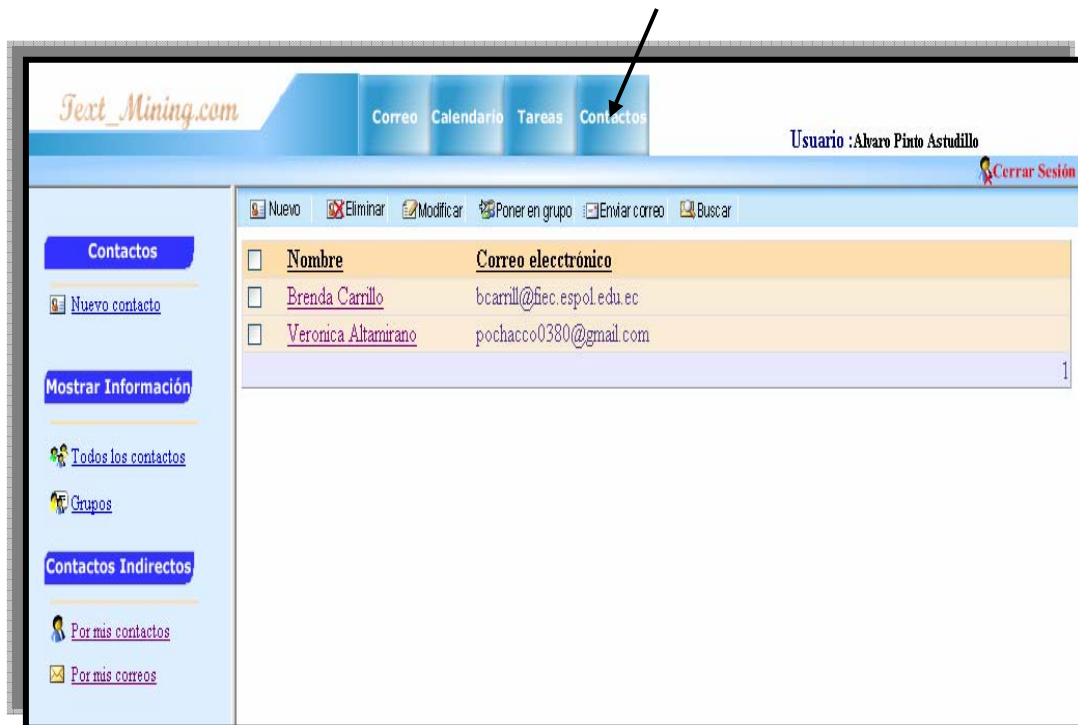


- ✓ **Poner en grupo:** Permitirá al usuario colocar la dirección del correo en cualquier grupo formado. [11]



- ✓ **Enviar correo:** Permite enviar un correo a la persona de la dirección seleccionada. [12]





2.4.1 CONTACTOS

2.4.1.1 NUEVO CONTACTO

Dando clic sobre el link [“nuevo contacto”](#) nos aparece la pantalla correspondiente para dicha acción. O también podemos acceder mediante el botón **nuevo** que nos aparece en la barra de herramientas, esta sirve para crear una nueva entrada y poder agregar un nuevo contacto para nuestra lista de direcciones. Véase en [8]

Esta pantalla tiene los siguientes campos

- **Datos personales:** Son los datos generales de la persona , como son:
 - nombre, apellido, nombre corto (alías), organización, profesión
- **Correo electrónico** Es la dirección de correo electrónico de la persona.
 - personal, trabajo, otros.
- **Números de teléfonos:** Son los números telefónicos que posee la persona
 - personal, trabajo, otros.

- **Direcciones:** Son las principales direcciones concurridas de la persona.

→ personal, trabajo: calle, ciudad, país, provincia

En esta ventana iremos rellendo los datos del *nuevo contacto*. Existe la posibilidad de que cada contacto puede tener varias direcciones de correo electrónico. Una vez completados los datos pulsamos el botón **guardar** o el botón **cancelar**.

Funciones de los botones:

- ✓ **Guardar:** La nueva ficha quedará almacenada en la lista de direcciones.
- ✓ **Cancelar:** Permite cancelar la acción de guardar datos.

2.4.2 MOSTAR INFORMACIÓN

2.4.2.1 TODOS LOS CONTACTOS

Esta opción nos permite mostrar y almacenar las direcciones que utilizamos con más frecuencia. Véase en [7].

En esta pantalla, se tienen la siguiente barra de herramientas: **nuevo** véase en [8], **eliminar** véase en [9], **modificar** véase en [10], **poner en grupo** véase en [11], **enviar correo** [12] y **buscar** véase en [9]

2.4.3 GRUPOS

Esta opción permite mostrar al usuario, el nombre y detalles de cada grupo formado.

- **Nombre** → El nombre característico del grupo.
- **Detalles** → Muestra la cantidad de integrantes del grupo con sus respectivos correos.

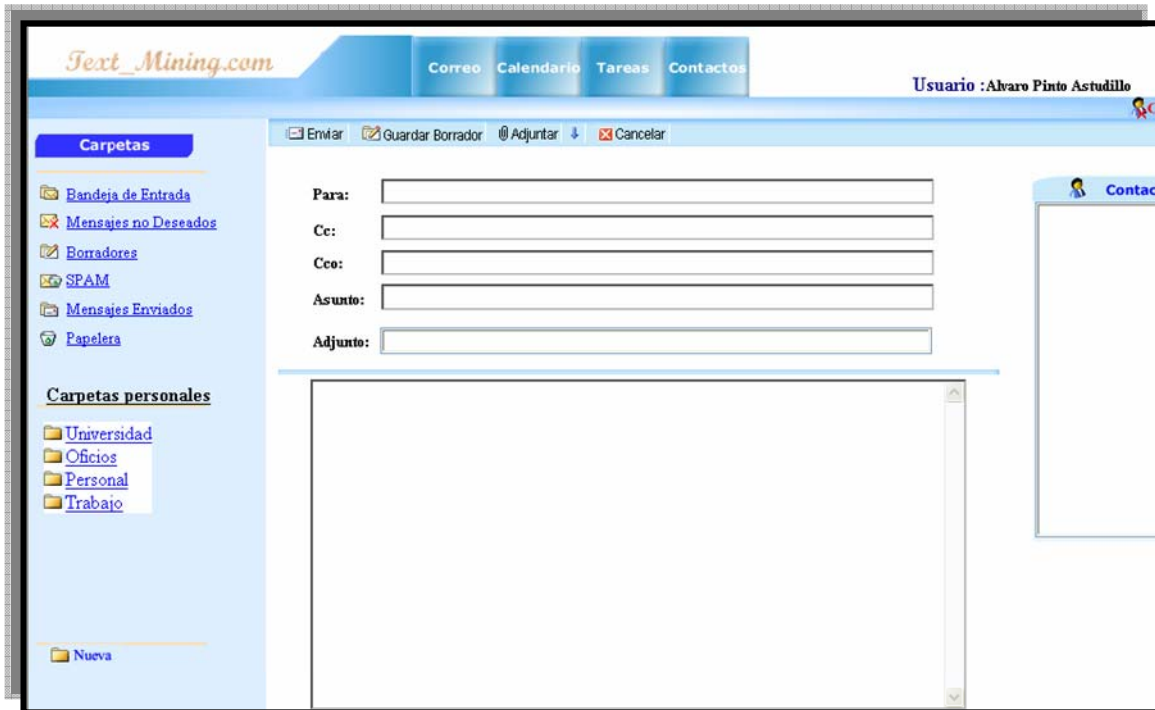
En esta pantalla, se tienen los siguientes botones:

- ✓ **Enviar correo:** Permite enviar correo al grupo formado o a la dirección que desee
- ✓ **Eliminar:** Permite eliminar al grupo creado de sus contactos.
- ✓ **Crear grupo:** Permite crear grupos de correos de sus contactos.
- ✓ **Modificar:** Permite realizar alguna modificación al grupo creado.

COMPOSICIÓN Y ENVÍO DE UN CORREO ELECTRÓNICO

Lo primero que debemos hacer antes de componer y enviar un correo es saber identificar las partes de un correo electrónico.

La que ves a continuación es la pantalla que aparece cuando redactamos un correo. La barra de herramientas contiene las típicas funciones como son: **enviar, guardar borrador, adjuntar, cancelar.**



- ✓ **Enviar:** Permite enviar el correo con sus datos adjuntos a las direcciones establecidas en los campos *PARA* y/o *CC*.
- ✓ **Guardar borrador:** Si no deseamos enviar el mensaje, en ese instante, pulsaremos el botón “**guardar borrador**”. De esta manera podremos volver al mensaje, modificarlo y enviarlo cuando queramos.
- ✓ **Adjuntar:** Aquí podemos colocar un fichero que será enviado junto con el correo. Por ejemplo se puede adjuntar un archivo gráfico, .GIF, un archivo comprimido .ZIP, un documento word .DOC, etc.
- ✓ **Cancelar:** Cancela toda acción.

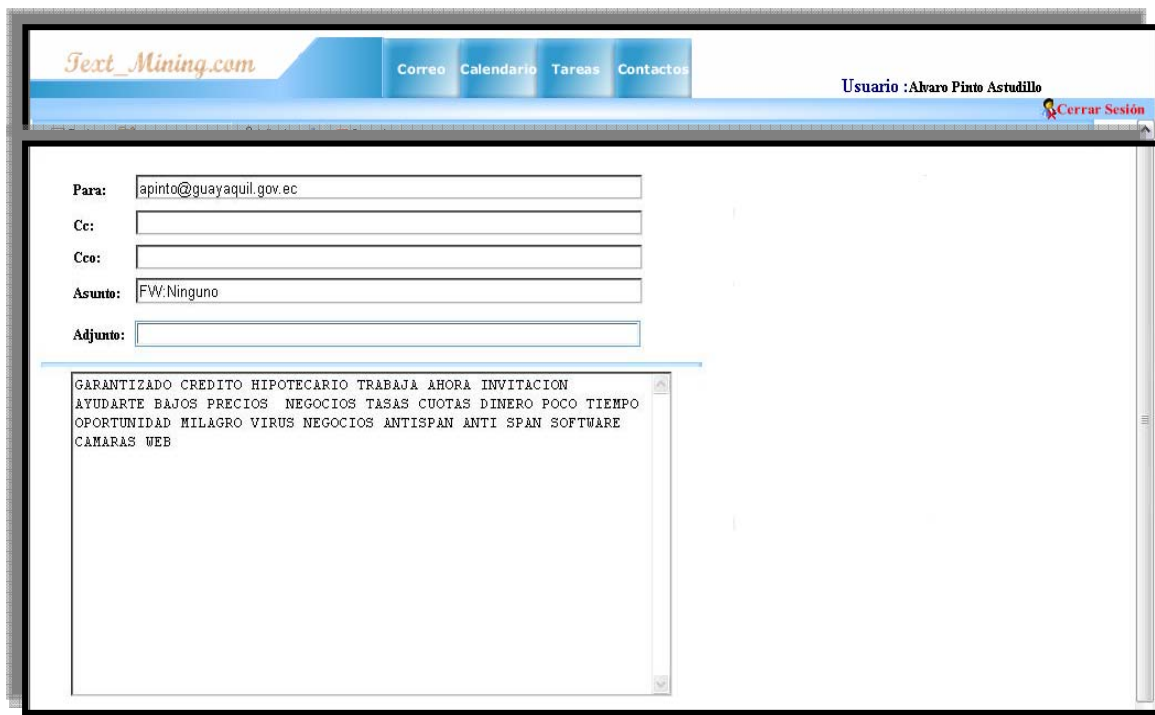
A continuación tienes los campos que forman parte del correo:

- **Campo PARA:** Aquí debemos poner la dirección del destinatario, se pueden poner más de una dirección separadas por un punto y coma, incluso se puede poner el nombre de una lista (grupo formado) que contenga varias direcciones. Cabe indicar que este campo es *obligatorio*, como mínimo el mensaje debe tener una dirección destino.
- **Campo CC:** Este campo sirve para enviar una copia del mismo mensaje a más de una persona a parte del receptor indicado en el campo **PARA**. El destinatario sabrá que no es el destinatario principal, sino que se le envía el correo como copia. Este campo no es obligatorio rellenarlo.
- **Campo CCO:** Este campo sirve para enviar una copia del mismo mensaje a más de una persona pero con copia oculta.
- **Campo ASUNTO:** Este campo sirve para indicar el motivo del correo, es decir, podemos indicar una breve descripción del tema del mensaje. Por ejemplo si es un mensaje amistoso podemos escribir un saludo, etc. El asunto aparecerá en la lista del destinatario, por lo tanto puede ser visto sin abrir el mensaje. Este campo tampoco es obligatorio rellenarlo.
- **Campo ADJUNTO:** Aquí se mostrará el archivo adjuntado.

Diferencias entre CC y CCO

CC y CCO sirven para enviar el mismo mensaje a más de una dirección de correo, pero existe una diferencia entre ambas, mientras que con CC al enviar el mensaje a varios receptores, el receptor ve las direcciones de los demás aparte de la suya propia, con CCO no pasa esto, las direcciones que no son la propia del receptor permanecen ocultas.

En la zona central tenemos el espacio reservado para el texto del mensaje, el área de escritura, donde podemos escribir el cuerpo del mensaje.



The screenshot shows a web-based email interface. At the top left is the logo "Text Mining.com". In the top center, there are navigation tabs: "Correo", "Calendario", "Tareas", and "Contactos". On the top right, it displays "Usuario :Alvaro Pinto Astudillo" and a "Cerrar Sesión" button. The main area contains a form for composing an email with the following fields:

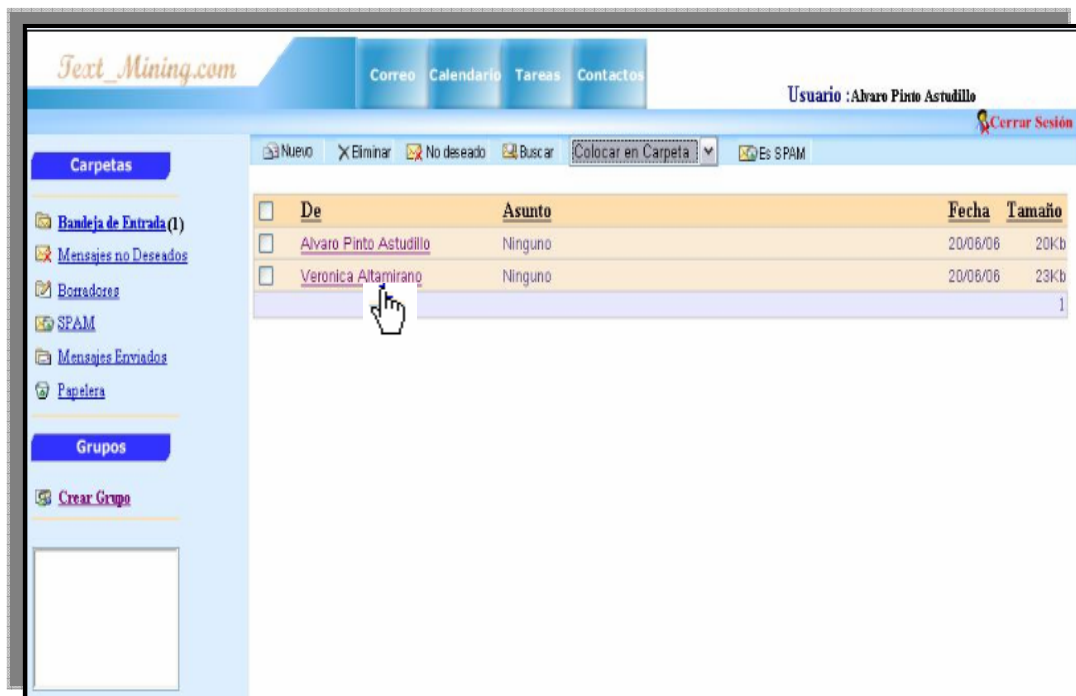
- Para:**
- Cc:**
- Cco:**
- Asunto:**
- Adjunto:**

Below the form is a large text area for the message body, containing the following text:

```
GARANTIZADO CREDITO HIPOTECARIO TRABAJA AHORA INVITACION
AYUDARTE BAJOS PRECIOS NEGOCIOS TASAS CUOTAS DINERO POCO TIEMPO
OPORTUNIDAD MILAGRO VIRUS NEGOCIOS ANTISPAN ANTI SPAN SOFTWARE
CAMARAS WEB
```

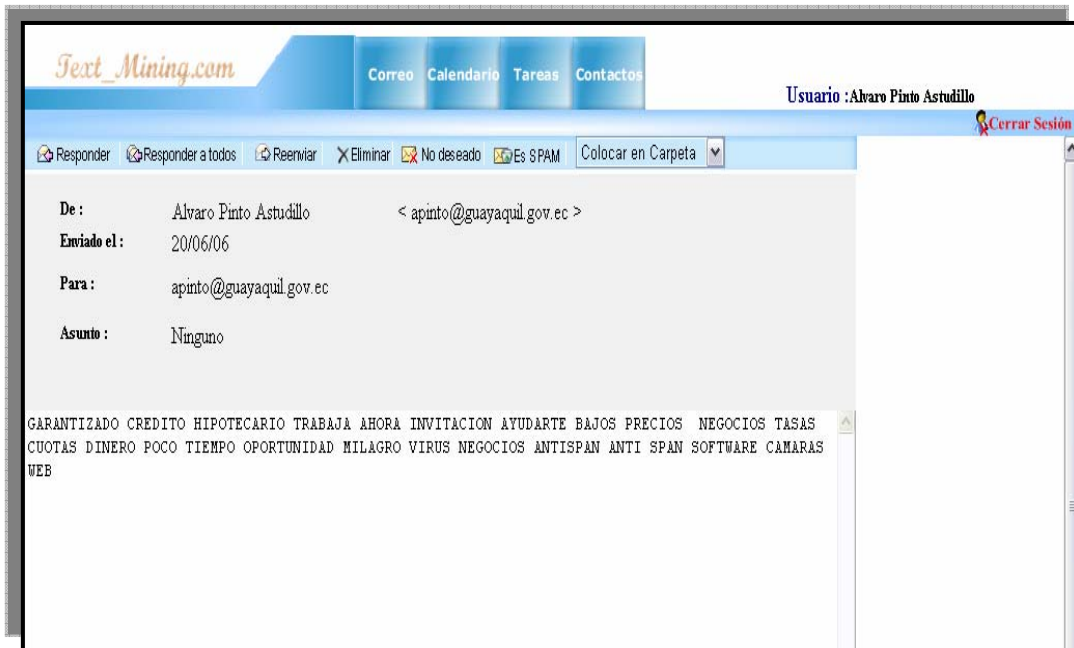
LEER UN CORREO

Para leer un mensaje recibido (ya sea nuevo o antiguo) basta picar sobre él con el ratón (tal y como se muestra en la imagen).



Se abrirá la ventana que contiene no sólo el contenido del mensaje sino también información acerca del remitente, del destinatario, el día del envío y del asunto. En la imagen el:

- Remitente y el destinatario es: apinto@guayaquil.gov.ec
- Asunto es un mensaje que tiene como título: **Ninguno**.
- Enviado el: Tiene como fecha de envío el 20/06/06.
- Su contenido aparece en la parte inferior de la ventana.



Una vez que hayamos leído el mensaje podemos reenviarlo a otro destinatario (botón reenviar), eliminarlo (botón eliminar), responder al destinatario (botón responder), responder a varios destinatarios (responder a todos), poner como no deseados (botón no deseados), colocar como SPAM (botón es SPAM) o colocar en carpeta (colocar en carpeta).

- ✓ **Responder:** Esta opción muestra una ventana de composición para un mensaje nuevo con la dirección del destinatario escrita en el campo PARA y el ASUNTO del mensaje completado con el mismo texto que llevaba el ASUNTO del mensaje original pero con las letras RE (de responder) al principio del texto del asunto. Incluso puede aparecer el texto del mensaje original. En algunos casos es útil

conservar este texto para que la persona que reciba la contestación sepa inmediatamente a qué le contestamos.

- ✓ **Responder a todos:** Aunque esta función es menos utilizada, ya que manda la respuesta a todas las direcciones de correo que hayan en el mensaje que tú has recibido, y en la mayoría de los casos son contactos del remitente que tú no conoces.
- ✓ **Reenviar:** Esta opción muestra una ventana de composición para un mensaje nuevo, con el cuerpo del mensaje relleno con el mismo que había en el mensaje original, en el campo **ASUNTO** pondrá el mismo texto que en el mensaje original más las letras **RV** (de reenviado) al inicio del texto del **ASUNTO**.
- ✓ **Eliminar:** Pasa el correo a la “papelera”
- ✓ **No deseado:** Permite enviar el correo a la carpeta de los “mensajes no deseados”.
- ✓ **Es SPAM:** Mensajes de remitentes bloqueados, correo SPAM. Permite enviar el correo a la carpeta “es SPAM”. Esta carpeta contendrá los correos catalogados como correos basura (SPAM).
- ✓ **Colocar en Carpeta:** Permite al usuario realizar la opción de regresar el correo a una de las bandejas como son: “bandeja de entrada”, “borradores”, “mensajes enviados”, “papelera”.



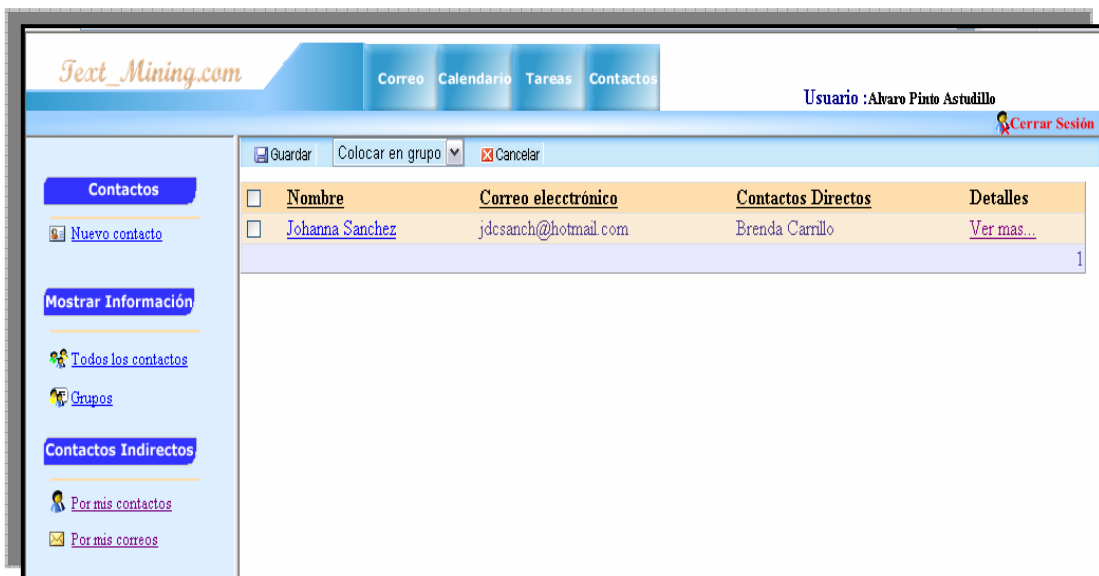
LISTA DE DIRECCIONES

Es una lista con todas las direcciones de las personas a las que escribimos correos electrónicos para que nos sea más cómodo enviar los mensajes y no tengamos que recordar esas direcciones.

2.4.3 CONTACTOS INDIRECTOS

A continuación tenemos los comandos más usuales como, guardar, colocar en grupo, cancelar.

- ✓ **Guardar:** Permite guardar los datos.
- ✓ **Colocar en grupo:** Permite al usuario realizar la opción de colocar el correo en uno de los grupos.
- ✓ **Cancelar:** Cancela toda acción.



En el link [ver detalles](#) tenemos la siguiente ventana, donde se indicarán: El nombre, correo, departamento, cargo referente a la persona de esa dirección de correo.

Se mostrarán 2 listados de mensajes los llamados: *sus contactos*³ y *mis contactos que los contienen*⁴ en ellos se indicarán su correo electrónico, el departamento en que se encuentra.

The screenshot shows the 'Text Mining.com' web interface. The user is logged in as 'Alvaro Pinto Astudillo'. The main content area displays the details for a contact named 'Johanna Sanchez'. The details are as follows:

Nombre:	Johanna Sanchez
Correo:	jdsanch@hotmail.com
Departamento:	Sistema
Cargo:	Programador

Summary statistics for the user's contacts:

Mis contactos:	2
CoBERTura:	1
Confianza:	50%

Below the details, there are two tables showing contact lists:

Sus contactos

Nombre	Correo electronico	Departamento
Veronica Altamirano	pochacco0380@gmail.com	Sistema

Mis contactos que lo contienen

Nombre	Correo electronico	Departamento
Brenda Carrillo	bcarrill@fieec.espol.edu.ec	Sistema

³ Sus contactos: Son todos los contactos que poseen dicho usuario.

⁴ Mis contactos que los contienen: Mis contactos tienen almacenados esa dirección.

ANEXO B

MANUAL TÉCNICO

DESCRIPCIÓN DE LAS VARIABLES DE ENTRADA Y SALIDA

➤ **Variables de Entrada**

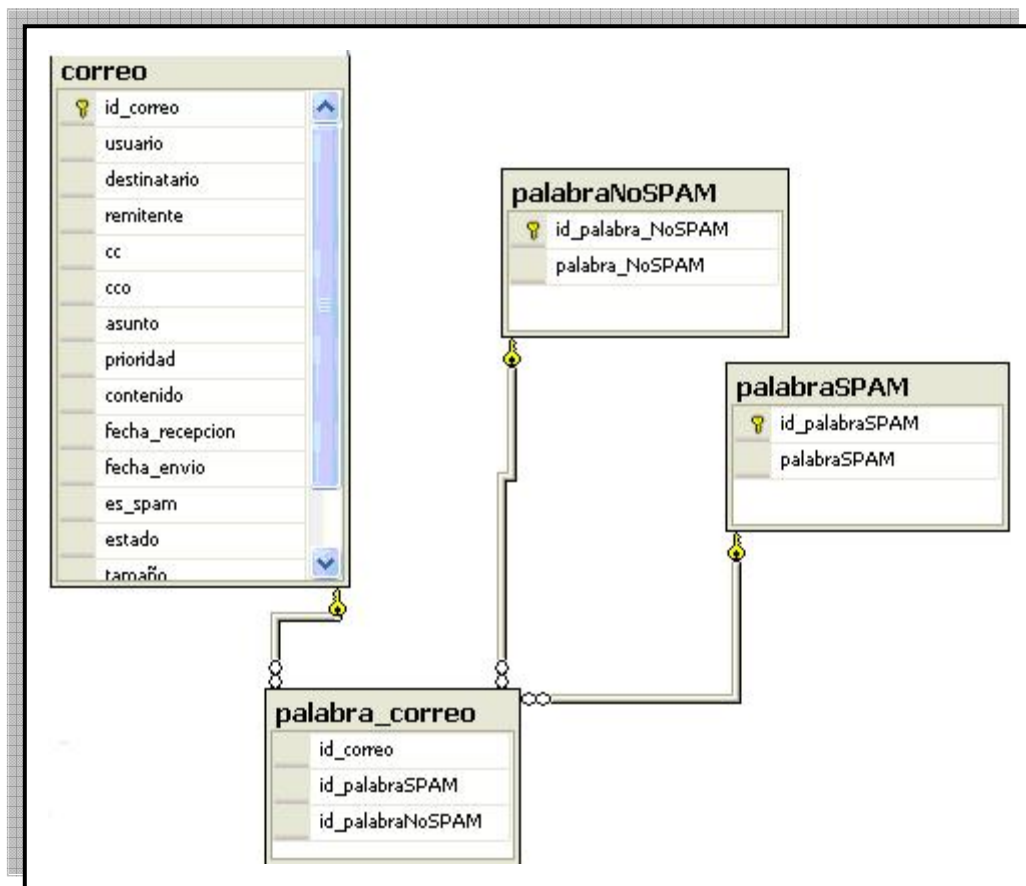
Cantidad de correos que serán utilizados por el clasificador para poder definir un texto como SPAM o no.

➤ **Variables de Salida**

Dará como resultado si el texto puede ser clasificado como:

SPAM: Esta se almacenará en la carpeta “es SPAM”.

No SPAM: En la bandeja de entrada.



Explicaremos la funcionalidad de las clases más importantes utilizadas en el proyecto.

ALGORITMO FILTRO BAYESIANO DE SPAM

- ❖ Esta clase verifica si en el contenido de los mensajes que tiene el usuario en su cuenta de correo, se encuentran las palabras más comunes consideradas como SPAM, lo cual los comparan con las palabras que se encuentran almacenadas en las dos tablas llamadas palabras_SPAM() y palabraNoSPAM() de nuestra base de datos.

```
public void verifica_spam()
    {
        usuario1 = Session["usuario"].ToString().Trim();
        palabra_SPAM();
        palabra_NoSPAM();
        ContenidoSPAM_Correo();
        ContenidoNoSPAM_Correo();
        probabilidad_Bayes();
        actualizar_base();
    }
private void Button1_Click(object sender, System.EventArgs e)
    {
        palabra_SPAM();
        palabra_NoSPAM();
        ContenidoSPAM_Correo();
        ContenidoNoSPAM_Correo();
    }
```

- ❖ Este método nos permite determinar cuantas de las palabras que se encuentran en el contenido del correo son palabras SPAM para esto hacemos uso de las palabras almacenadas en la tabla *palabra_SPAM*.

```
public void palabra_SPAM()
{
    try
    {
        palabraSPAM = new string[500,2];
        int i=0;
        SqlDataReader myReader = null;
        SqlConnection cnn = new SqlConnection("database=mineria; integrated
        security=yes");
        string mySelect = "SELECT id_palabraSPAM, palabraSPAM FROM
        palabraSPAM";
        SqlCommand cmd = new SqlCommand(mySelect, cnn);
        cnn.Open();
        cmd.ExecuteNonQuery();
        myReader = cmd.ExecuteReader();

        while(myReader.Read())
        {
            id_palabraSPAM=myReader.GetInt32(0);
            palabraSPAM[i,0]=id_palabraSPAM.ToString();
            palabraSPAM[i,1]=myReader.GetString(1);
            i++;
        }

        numpalabraSPAM=i;
        cnn.Close();
    }
    catch(Exception ex)
    {
        MessageBox.Show("Error en la conexión a la Base de
        datos");
    }
}
```

- ❖ Este método nos permite determinar cuantas de las palabras que se encuentran en el contenido del correo son palabras no SPAM para este análisis hacemos uso de las palabras almacenadas en la tabla *palabra_NoSPAM*.

```
public void palabra_NoSPAM()
{
    palabraNoSPAM = new string[500,2];
    int i=0;
    SqlDataReader myReader = null;
    SqlConnection cnn = new
SqlConnection("database=mineria; integrated security=yes");
string mySelect = "SELECT id_palabra_NoSPAM, palabra_NoSPAM
FROM palabraNoSPAM";
SqlCommand cmd = new SqlCommand(mySelect, cnn);

    cnn.Open();
    cmd.ExecuteNonQuery();
    myReader = cmd.ExecuteReader();

while(myReader.Read())
    {
        id_palabraNoSPAM=myReader.GetInt32(0);

        palabraNoSPAM[i,0]=id_palabraNoSPAM.ToString();

        palabraNoSPAM[i,1]=myReader.GetString(1);
            i++;
    }
    numpalabraNoSPAM=i;
    cnn.Close();
}
```

- ❖ Este método contiene la cantidad de correos de entrenamientos que tiene en su contenido palabras que son SPAM.

```
public void ContenidoSPAM_Correo()
{
    //usuario1 = usuario.Text;
    int i=0;
    int j=0;
    contenido=new string[500,2];
    while(j <= numpalabraSPAM)
    {
        SqlDataReader myReader = null;
        SqlConnection cnn = new
SqlConnection("database=mineria; integrated security=yes");
string mySelect = "SELECT id_correo FROM correo WHERE usuario =
"+usuario1+" and contenido LIKE '%" +palabraSPAM[j,1]+"%";
SqlCommand cmd = new SqlCommand(mySelect, cnn);

        cnn.Open();
        cmd.ExecuteNonQuery();
        myReader = cmd.ExecuteReader();

        while(myReader.Read())
        {
            id_correo=myReader.GetInt32(0);
            contenido[i,0]=id_correo.ToString();
            i++;
        }
        cnn.Close();
        j++;
    }
    numcorreoSPAM=i;
}
```

- ❖ Este método contiene la cantidad de correos de entrenamientos que tiene en su contenido palabras que son no son SPAM.

```
public void ContenidoNoSPAM_Correo()
{
    //string usuario1 = usuario.Text;
    int i=0;
    int j=0;
    while(j <= numpalabraNoSPAM)
    {
        SqlDataReader myReader = null;
        SqlConnection cnn = new
        SqlConnection("database=mineria; integrated security=yes");
        string mySelect = "SELECT id_correo
        FROM correo WHERE usuario = '"+usuario1+"' and contenido LIKE
        '"+palabraNoSPAM[j,1]+"%";
        SqlCommand cmd = new
        SqlCommand(mySelect, cnn);

        cnn.Open();
        cmd.ExecuteNonQuery();
        myReader = cmd.ExecuteReader();

        while(myReader.Read())
        {

            id_correo=myReader.GetInt32(0);

            contenido[i,1]=id_correo.ToString();
                i++;
            }
            cnn.Close();
            j++;
        }
        numcorreoNoSPAM=i;
    }
}
```

- ❖ Este método chequea si el correo contiene alguna palabra a la cual se le pueda obtener la probabilidad (números de correos en los que aparece la palabra con los números de SPAM en los que aparece la palabra).

```
public void probabilidad_Bayes()
{
    SPAM=new string[500,2];
    resultado=new string[500,4];
    string temp;
    int i,l,cont2,tot;
    int cont=0;
    double prob;
    for(int j=0; j<numcorreoSPAM-1; j++)
    {
        for(int k=j+1; k<numcorreoSPAM-2; k++)
        {
            if(int.Parse(contenido[k,0])<
int.Parse(contenido[j,0]))
                {
                    temp=contenido[j,0];
                    contenido[j,0]=contenido[k,0];
                    contenido[k,0]=temp;
                }
        }
    }
    for(int j=0; j<numcorreoNoSPAM-1; j++)
    {
        for(int k=j+1; k<numcorreoNoSPAM-2; k++)
        {
            if(int.Parse(contenido[k,1])< int.Parse(contenido[j,1]))
                {
                    temp=contenido[j,1];
                    contenido[j,1]=contenido[k,1];
                    contenido[k,1]=temp;
                }
        }
    }

    l=0;
    tot=1;
    while(cont<numcorreoSPAM)
    {
        i=cont;
```



```

while((contenido[cont,0]== contenido[i+1,0])&& i<numcorreoSPAM-
    1)
    {
        i++;
        tot++;
    }
resultado[l,0]=contenido[cont,0];
resultado[l,1]=tot.ToString();
cont=i+1;
l++;
tot=1;
}
numresultado1=l;

l=0;
cont=0;
tot=1;
while(cont<numcorreoNoSPAM)
    {
        i=cont;
while((contenido[cont,1]== contenido[i+1,1])&&
    i<numcorreoNoSPAM-1)
        {
            i++;
            tot++;
        }
resultado[l,2]=contenido[cont,1];
resultado[l,3]=tot.ToString();
cont=i+1;
l++;
tot=1;
}
numresultado2=l;

cont=0;
cont2=0;
numSPAM=0;

```

```

while(cont2<numresultado2 && cont<numresultado1)
    {
        if(int.Parse(resultado[cont,0])<int.Parse(resultado[cont2,2]))
            {
                ** CÁLCULOS DE LAS PROBABILIDADES **
                prob=Math.Round(((double.Parse(resultado[cont,1])/numpalabraS
                PAM)/(double.Parse(resultado[cont,1])/numpalabraSPAM))*100,0);
                SPAM[cont,0]=resultado[cont,0];
                SPAM[cont,1]=prob.ToString();
                cont++;
            }
            else
                if(int.Parse(resultado[cont,0])>int.Parse(resultado[cont2,2]))
                    {
                        prob=0.0;
                        SPAM[cont,0]=resultado[cont,0];
                        SPAM[cont,1]=prob.ToString();
                        cont2++;
                    }
                    else
                        {

                            prob=Math.Round(((double.Parse(resultado[cont,1])/numpal
                            abraSPAM)/((double.Parse(resultado[cont2,3])/numpalabraNoSPA
                            M)+(double.Parse(resultado[cont,1])/numpalabraSPAM)))*100,0);
                            SPAM[cont,0]=resultado[cont,0];
                            SPAM[cont,1]=prob.ToString();
                            cont++;
                            cont2++;
                        }
                        numSPAM++;
                    }
            }
    }

```

- ❖ Este método actualiza la base, es decir una vez detectada que el correo entrante es un correo SPAM, vaya directamente a la carpeta que contendrán los correos basuras , en otras palabras a la carpeta “es SPAM”.

```
public void actualizar_base()
{
    for(int j=0; j<numSPAM; j++)
    {
        if(int.Parse(SPAM[j,1])>55)
        {
            SqlConnection cnn = new
SqlConnection("database=mineria; integrated security=yes");
            string myUpdate = "UPDATE correo
SET es_spam = 1, carpeta = 4 WHERE id_correo = '"+SPAM[j,0]+'";
            SqlCommand cmd = new SqlCommand(myUpdate,
cnn);
                                cnn.Open();
                                cmd.ExecuteNonQuery();
                                cnn.Close();
        }
    }
}

private void usuario_TextChanged(object sender,
System.EventArgs e)
{
}
}
```

BIBLIOGRAFÍA

[1] Luis Antonio Fernández, CAPÍTULO III: APLICACIONES DE DATA MINING.

<http://www.monografias.com/trabajos26/data-mining/data-mining2.shtml>

[2] México, D.F., Manuel Montes-y-Gómez, “Minería de texto: Un nuevo reto computacional”. [http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-](http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf)

[md01.pdf](http://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf)

[3] Cristian Villanueva S., “Filtrado Automático de Spam basado en Machine Learning.doc”

[4] SPAM.

<http://www.rompecadenas.com.ar/spam.htm>

[5] SPAM: mensajes de correo no solicitados, características.

http://www.pandasoftware.es/virus_info/spam/?sitepanda=particulares

[6] MALWARE.

http://www.google.com.ec/search?hl=es&lr=lang_es&defl=es&q=define:Malware&sa=X&oi=glossary_definition&ct=title

[7] SPIT, Fuente: HISPAMP3.COM

http://www.sitiosargentina.com.ar/notas/octubre_2004/spit.htm

[8] Tipos de SPAM.

<http://www.viruslist.com/sp/spam/info?chapter=153350533>

[9] (viernes, 03 de noviembre de 2000), Radio “EL ESPECTADOR” Uruguay, “Contra el bombardeo electrónico: espectador.com lanza una campaña "anti spam",

http://www.espectador.com/spam/spam_01.htm#por_que

[10] Filtrado bayesiano de SPAM: Teorema.

<http://www.tejedoresdelweb.com/slides/spam/img24.html>

[11] Universidad Central de Venezuela, TEMA II. “PROCESO DE DESARROLLO DE MINERÍA DE DATOS”.

strix.ciens.ucv.ve/~iartific/Material/Tema%202-%20Pasos%20de%20KDD.doc –

[12] El Correo no deseado como problema de seguridad.

<http://www.pc-news.com/detalle.asp?sid=&id=11&Ida=1369>

[13] Tarifas del uso de banda ancha.

http://www.terra.cl/contrata_terra/index.cfm?accion=interior_planes&idf=16

[14] Lista de precios de servidores HP.

<http://www.makrocomputo.com/listas.htm>

[15] Lista de precios de servidores Intel.

www.dicopel.com.mx/lpPPC.xls

[16] INTRODUCCIÓN A LA MINERÍA DE DATOS, HERNÁNDEZ, J.; RAMÍREZ, M.J. y FERRI, C. Prentice Hall, España, primera edición, 2004.

[17] ANÁLISIS DE DATOS MULTIVARIANTES, Daniel Peña. Mc. Graw Hill, España, primera edición, 2002