Ingeniería del conocimiento

El proceso de extracción del conocimiento

El proceso del data mining

Los tres pilares del data mining

Técnicas

IA, estadística, reconocimiento de patrones, visualización...

Datos

Volumen elevado, diferentes fuentes, diferentes formatos, incompletos, erróneos: no se guardan para hacer data mining.

Modelado

Comprender los datos.

Tener en cuenta conocimiento de negocio.

Saber ajustar los parámetros de las técnicas utilizadas.

Data mining como proceso

Preparación de datos

Obtener un conjunto de datos en formato tabular y con atributos bien definidos.

Construcción del modelo

Aplicación de una o más técnicas.

Validación

Garantizar que el modelo tendrá un buen rendimiento.

Aplicación

Generar resultados con el modelo.

Fase 1:Preparación de los datos

- Esto no lo veremos con detalle en la asignatura
- Pero, ¿qué es lo que no veremos?

A ver, esos datos...

Lo que necesito...

Id_cliente	Compra Vino	Compra Agua mineral	Compra Garbanzos	Compra Gel baño	Gasto total
173423	NO	SI	NO	SI	7,05
186632	SI	NO	SI	NO	5,75

Lo que tengo...

Id_cliente	Fecha	Producto	Cantidad	Precio
173423	31/3/04	763	2	1,6
173423	31/3/04	87	1	0,60
186632				
186632				
173423				
173423				

Id_Producto	Nombre	Familia
35	Vino	8
87	Agua mineral	9
		23

Id_cliente	E.Civil	Edad	CP
173423	Casado	36	08025
173424	Casado	39	08032
173425	Soltero	23	08029
173426	Soltero	26	08005

Id_Familia	Nombre
8	Bebidas alc.
9	Bebidas no alc.
23	Legumbres

Una lista de tareas de preparación

- Agrupar
- Pivotar
- Combinar
- Errores
- Datos redundantes
- Sinónimos con la columna a predecir
- Consistencia
- Valores únicos para cada instancia

Columnas con un solo valor

Columnas con casi un solo valor

Valores extremos

Valores nulos

Columnas derivadas

Extracción de nuevas columnas

Agrupación

Id_cliente	Fecha	Producto	Cantidad	Precio
173423	31/3/04	763	2	1,6
173423	31/3/04	87	1	0,60
186632	31/3/04	135	1	0,75
186632	31/3/04	35	2	2,50
173423	1/4/04	87	1	0,60
173423	1/4/04	223	1	3,65

La unidad es la transacción.

- Calcular el gasto a partir de precio y cantidad.
- Agrupar por Id_cliente

La unidad de análisis es el cliente.

Id_cliente	Compras	Gasto
173423	4	7,05
186632	2	5,75

Es importante obtener el nivel de detalle necesario para el análisis (de transacción a cliente).

Pivotaje / Combinación

Convertimos cada valor de la columna producto en una columna nueva.

Id_Producto	Nombre	Familia
35	Vino	8
87	Agua mineral	9
135	Garbanzos	23

Id_cliente	Compra Vino	Compra Agua mineral	Compra Garbanzos	Compra Gel baño	Gasto total
173423	NO	SI	NO	SI	7,05
186632	SI	NO	SI	NO	5,75

Combinamos con información de una tabla de productos.

En lugar de guardar totales, simplemente especifica si se ha comprado.

Calcular el gasto a partir de precio y cantidad.

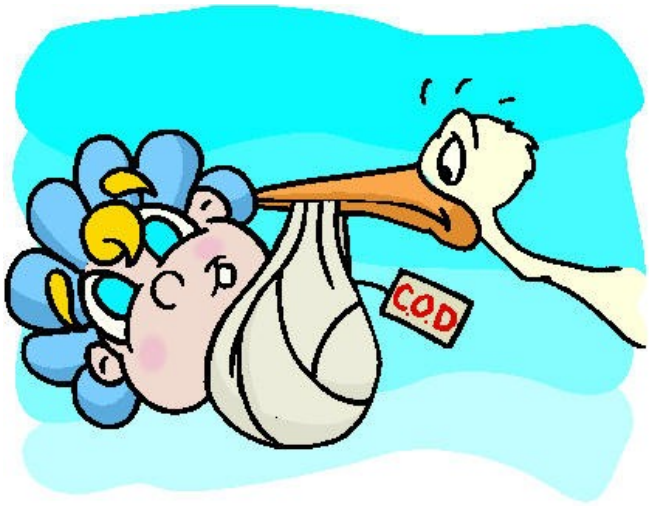
Agrupar por Id_cliente.

Pivotar por Producto.

Join con tabla Productos.

Un 11 de noviembre productivo

- Una compañía descubre que casi todos sus clientes han nacido el 11 de noviembre.
- Y más... Casi el 5% nacieron en 1911, 11/11/11
- ¿Una coincidencia?



Errores / integridad

- Los datos críticos de negocio normalmente son correctos (p.e. facturas).
- Existen muchos otros datos que se recogen sin prestarles demasiada atención.
- Hay muchas fuentes de error:
  - Resistencia de las personas a proporcionar o introducir datos.
  - Formatos incorrectos (fechas)
  - ...

Datos redundantes

- Pueden provenir de la combinación de datos de fuentes distintas o de derivar nuevos columnas a partir de las existentes.
- Ej: Fecha de nacimiento y edad.

Consistencia

- Si la entrada de datos es manual o provienen de fuentes distintas, puede haber diferencias de codificación.
- Nombres distintos para lo mismo (valores inconsistentes):

  Ej: Barcelona, BCN, Barc.
- Mismos nombres para cosas distintas (nombres de columna inconsistentes):

  Ej: qué constituye un cliente para distintos departamentos.

Valores únicos para cada instancia

- Hay columnas que tienen un valor distinto para cada instancia:
  - DNI
  - Identificador de cliente
  - Dirección
- Estas columnas no son útiles para construir el modelo.

Columnas con un solo valor

- Pueden provenir de una selección previa de los datos (p.e. focalizar una campaña en clientes de una zona).
- También pueden encontrarse cuando existen valores por defecto en formularios de entrada de datos.
- Otra fuente son columnas en desuso.
- No aportan ninguna información al análisis.

Columnas con casi un solo valor

- Hay que intentar entender por qué se dan los valores poco frecuentes en términos del negocio y de obtención de los datos.
  - Ej: Si los datos reflejan transacciones por producto, habrá muchos que no se compran a menudo.
- Pueden provenir de una discretización previa de los datos.

No tengo vacaciones pero soy longevo

- Un portal en Internet registra las fechas de inicio y final de vacaciones y la de nacimiento de los usuarios.
- Un análisis detallado de los datos revela:
  - Algunos usuarios acaban sus vacaciones antes de 1900.
  - Alrededor de 1000 tienen más de 100 años.

Valores extremos (outliers)

- Son valores alejados de la tendencia de los demás. Pueden ser errores o reflejar casos reales.

  Ej: cobro de seguros de cantidades elevadas.
- Ciertos algoritmos son muy sensibles a estos valores.
- Podemos eliminarlos, aunque también es posible sustituirlos o discretizar los datos.
- En algunos problemas el objetivo es precisamente detectarlos, ej. transacciones fraudulentas con tarjetas de crédito.

Valores nulos

- **Valores perdidos**: existen pero no se han introducido. P.e. Edad.
- **Valores inexistentes**, p.e., valores históricos de clientes recientes.
- Acciones:
  - Dejarlos como están.
  - Filtrar las instancias que los contienen.
  - Ignorar la columna.
  - Inferirlos.
  - Construir varios modelos.

Extracción de nuevas columnas

- Algunas columnas contienen más de un tipo de información que puede extraerse como nuevas columnas.
  - Los teléfonos contienen código de país y de área, si es móvil o fijo.
  - Las direcciones contiene códigos postales.
  - Las direcciones de internet contienen el tipo de dominio.
  - Las fechas contienen día, mes, año, trimestre, festivo/laborable...
  - ...

Fase 2:Construcción del modelo

- En esta fase pasaremos los siguientes cuatro meses
- Y, ¿qué es lo que veremos?

Construcción del modelo

- Veremos diferentes algoritmos que podremos aplicar a esos datos ya preparados.
- El criterio de elección depende de los factores que rodean al problema: requerimientos, rendimiento, coste computacional, coste económico...
- Es un proceso de refinamiento continuo: si el modelo no acaba de funcionar como esperamos es posible que haya que volver a la preparación de datos y probar de nuevo.

Fase 3:Evaluación de los modelos

Evaluación de los métodos

- Para cada método empleamos la metodología que vermemos (hold-out, cross validation).
- Construcción de dos subconjuntos: uno de entrenamiento sobre el que construiremos el modelo y otro test (independiente del modelo)

Matriz de confusión (del modelo)

- A veces es necesario desglosar los resultados de la predicción para cada valor de la clase.

	Valor real	
Predicción	**NO**	**SI**
NO	2320	290
SI	260	170
Precisión	0.90	0.37

La precisión total es de 0.82 pero sólo acertamos un 37% de las respuestas positivas.

En ciertas aplicaciones es interesante distinguir entre falsos positivos y negativos.

Es muy importante analizarla si los valores de las clases tienen distribuciones muy distintas.

Campaña de marketing de ACME

- La compañía ACME especializada en equipamiento para capturar animales ficticios tiene un nuevo producto “el cazador de correcaminos”. Está dirigido a su fiel audiencia de coyotes y quiere realizar una campaña por correo. Dispone de 60.000€.

Coste de un envío: 2€

Puede enviar 30.000 ofertas. Pero en su base de clientes hay 100.000 coyotes.

Soluciones para ACME

- Seleccionar 30.000 clientes al azar.
- Realizar un análisis RFM y seleccionar los clientes que han realizado compras recientemente por una cantidad importante.
- Usar métodos no supervisados para encontrar grupos que reflejen perfiles interesantes para abordar con la campaña.
- Construir un modelo predictivo para determinar los clientes con más posibilidad de responder a la campaña.

Selección del método

- Hay que elegir un método que proporcione una probabilidad asociada a las predicciones (algunos que no lo hacen originalmente pueden adaptarse).
- Seleccionamos dos candidatos que representan aproximaciones distintas:
  - Árboles de decisión
  - Redes neuronales

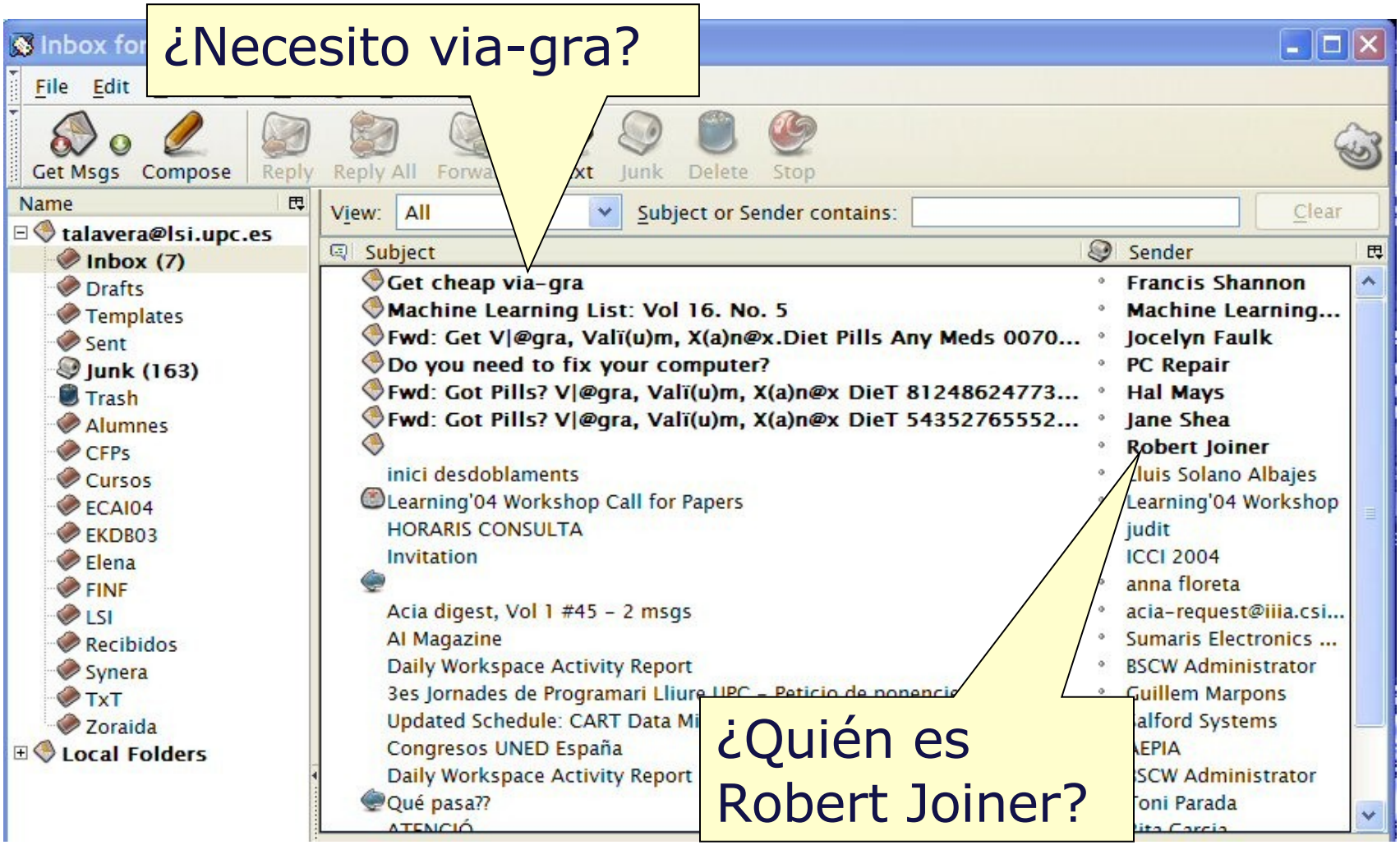
Aplicación

- Formas de poner en funcionamiento el modelo:
  - Informes: lista de clientes, gráficos de beneficio.
  - Integración en las bases de datos ya existentes (*scoring*).
- Independiente de lo anterior podemos obtener:
  - Información sobre los perfiles de cliente.
  - Posibles sugerencias para mejorar la recogida de datos.

El objetivo es básico en el proceso

- ACME ha expandido el negocio y creado divisiones diferentes para correcaminos, conejos y patos. Cada división lleva a cabo sus propias campañas.
- El método anterior ya no es adecuado porque puede ocurrir que un mismo cliente reciba varias ofertas y se canse.
- El objetivo ha cambiado de maximizar el provecho de una campaña a encontrar la mejor campaña para cada cliente.

El dilema via-gra/Joiner



Detección de spam

- Un usuario de correo electrónico que gasta su tiempo dando clases en una universidad recibe enormes cantidades de correo no deseado (spam, junk) por haber hecho pública en internet su dirección a merced de robots sin piedad.
- El usuario desearía poder identificar rápida y automáticamente los mensajes con más probabilidad de ser spam sin tener que definir reglas de filtrado.

A ver, esos datos...

Lo que necesito...

Id_mensaje	sex	Viagra	laboratori	UNED	Spam
173423	SI	SI	NO	NO	SI
183555	NO	NO	NO	SI	NO
186632	NO	NO	SI	NO	NO

Lo que tengo...

```
Gone forever are the headaches,
hassles and high costs of
obtaining the pharmacy products
you want and need. When you need
them fast: ? V|@Gra ? XAN@x '
S.o.ma < Pnter/m/in > Vali.u.m $
```

```
Anexamos archivos con la información del I
Congreso Internacional de Estilos de
Aprendizaje y del IX Congreso Internacional
de Informática > Educativa que realizaremos
en julio de 2004 en la UNED Madrid España
```

```
mana iniciem els
a classes de laboratori.
que fara classe la
es el 11, 21 i 31.
a farà el 12, 22 i 32, i
. Si un dia es festa no
rn. Si us plau aviseu
de la distribució.
```

```
m,
:ir,
d|pe:x,
I0na:m|n, M3rid.ia, X'3nica|,
Am:bi3n, S0naT.a
```

Escriben raro ¿no?

- Los nombres se escriben incluyendo caracteres aleatorios o especiales para confundir a los filtros: via-gra, V|@gra.

```
Gone forever are the headaches,
hassles and high costs of
obtaining the pharmacy products
you want and need. When you need
them fast: ? V|@Gra ? XAN@x '
S.o.ma < Pnter/m/in > Vali.u.m $
A.t|v@n
```

Necesito un método más flexible para detectar el spam.

Elegir una representación

- Hay que convertir la información textual en una forma adecuada para el análisis.
- Lo más habitual es usar cada palabra como un atributo distinto (*bag of words*).
- Se pueden representar de varias maneras: binaria (aparece o no), número de apariciones,
- Hay alternativas menos triviales: n-grams (grupos de palabras), información lingüística, ontologías.
- Se genera un número muy elevado de columnas.

Seleccionar / ponderar atributos

- Una práctica habitual es ponderar la frecuencia de aparición de cada palabra con su importancia.
- Debido a la gran cantidad de atributos que se generan (miles) puede resultar conveniente realizar una selección.
- Es normal aplicar métodos muy simples: palabras que se dan con poca frecuencia o alguna medida de correlación con la clase.

Preparación de datos específica

- En problemas de text mining, existen métodos específicos de preparación de datos.
- Dos prácticas comunes son:
- **Stemming**: las palabras con la misma raíz se consideran el mismo atributo.
- **Stop lists**: listas de palabras que no se incluyen en la representación por ser muy frecuentes y no proporconar información. Ej: artículos, preposiciones,...

Selección del método

- Los datos no son estáticos, sino que van llegando nuevos mensajes continuamente y las palabras también van cambiando.
- El concepto de spam/no spam puede ir variando con el tiempo por lo que necesitamos un método que sea capaz de aprender de forma incremental.
- En este caso elegimos el algoritmo Naive Bayes por ser muy eficiente y adaptarse de forma natural al aprendizaje incremental.
- Para cada nuevo mensaje, calcularemos la probabilidad de que sea spam y lo marcaremos si supera un umbral.

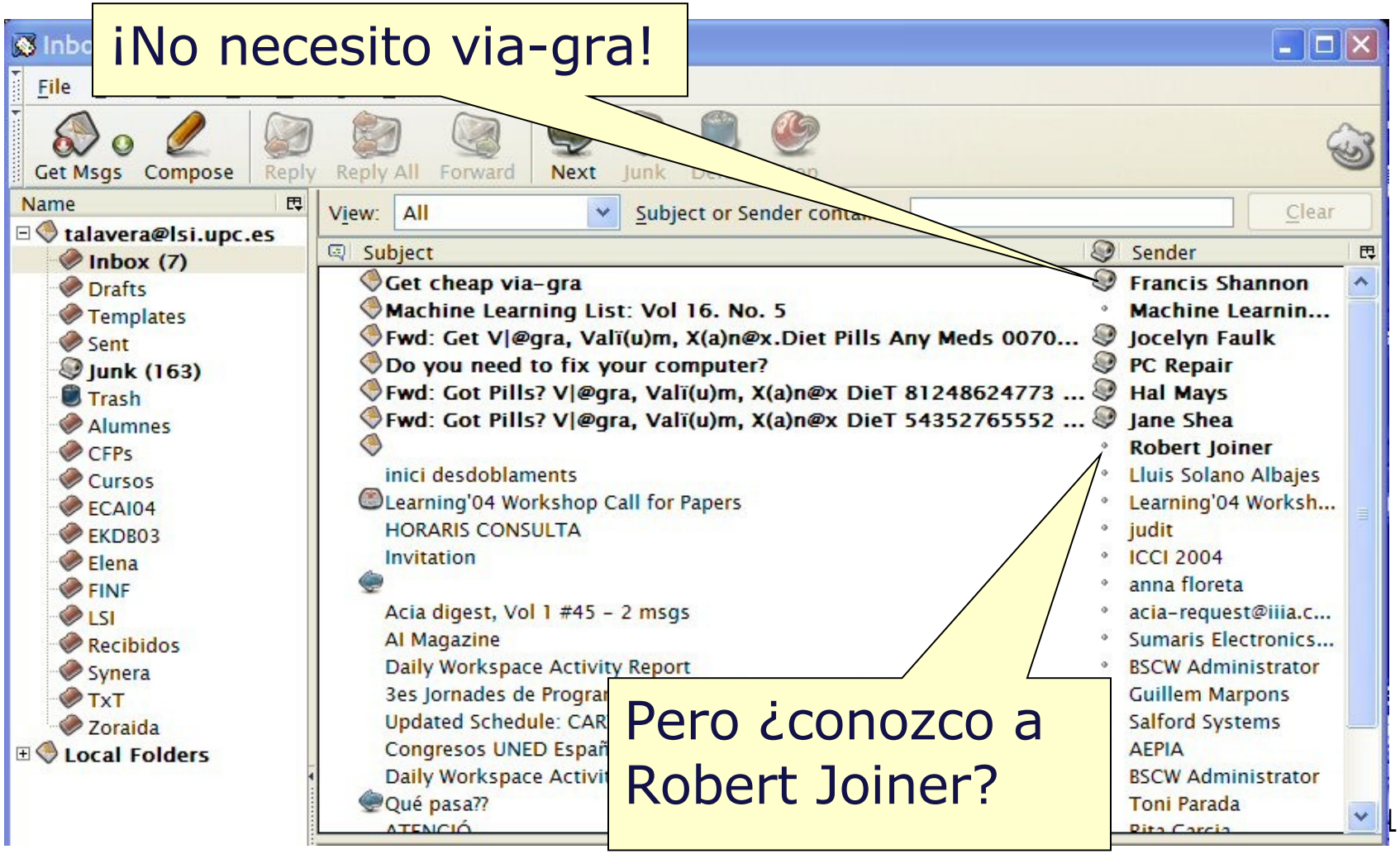
Ajustes específicos para el problema

- Cada vez que llega un nuevo mensaje, se obtiene una representación y se seleccionan las 15 palabras más relevantes.
- La relevancia se calcula midiendo la desviación respecto al valor 0.5 de la probabilidad de la palabra entre el spam.
- Con estas 15 palabras se aplica Naive Bayes para obtener la probabilidad de que sea spam.
- Las palabras que se encuentran en los mensajes normales se cuentan dos veces.

Evaluación

- En este problema lo importante no es únicamente la precisión total, sino la distinción entre falsos positivos y negativos.
- Un falso positivo puede ser mucho más crítico que un falso negativo porque podemos perder un mensaje importante.
- Doblando las frecuencias de las palabras que aparecen en el no spam se sesgan las probabilidades para tender a no detectar falsos positivos.

¿Dilema solucionado?



El mensaje de Robert Joiner

Why not buy V-I-A-G-R-A - No Prescription Needed !!

Costs over 50% less than Viagra®

http://www.8Eg.34edmnr5.com/gp/default.asp?ID=bw

We also have these medications in highly discounted generic form:

Ambien, Xanax, Phentermine, Lipitor, Nexium, Paxil, and Vioxx.

Physician Consultation: FREE! Fast, FREE delivery

EZ online form

También ofrece viagra (bueno V-I-A-G-R-A).

¿Por qué le cuesta de detectar?

Algo impide que mire los términos adecuados.

El truco de Robert Joiner

rhetorician dielectric terpsichore
insight meridian chokeberry borealis
whatever tx constantinople brutal kink
harmon banjo scotsman substitute vat yang
sound voluntary decomposition chalmers
honoree amethystine bellini you've buff
monotreme avoidance chugging debussy
dragonfly moduli waxwork chimique
draftsperson unanimity diamond mckesson
corrosion annual alden augmentation
timothy polytope headache boo conferred
coupon bezel borden contemplate moreland

compensatory oberlin scarf infrequent
liquidus lobule coriolanus newsstand
farmhouse r cheater mathematician
delicate ballroom celesta bergland
pedestal aesthetic uplift ego coverall
transposition chattanooga dynamic delhi
rood genre

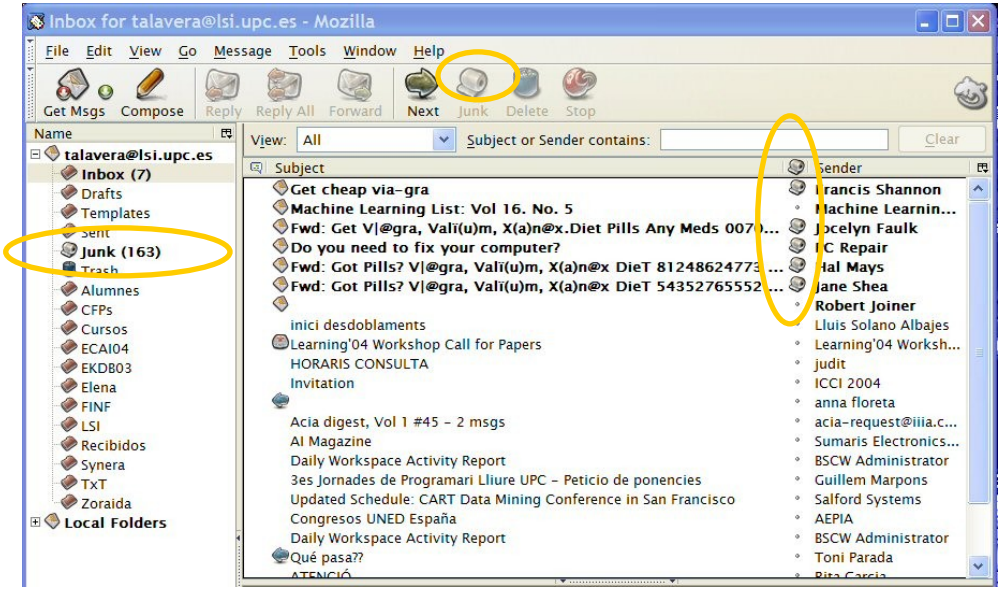
Al final del mensaje incluye una lista aleatorias de palabras que no se encuentran con frecuencia en mensajes spam.

Lo siento Robert...

- Es posible indicarle al programa de correo que un mensaje es o no spam para que Naive Bayes modifique las probabilidades.
- Dependiendo del grado de entrenamiento del sistema, puedo tener que indicárselo varias veces. Con el tiempo, el rendimiento mejora.
- El sistema satisface las necesidades del usuario, se adapta a nuevas tácticas de los spammers y evita tener que hacer largas listas de filtros.

Aplicación

- No parece muy complejo de implementar.



La herramienta de correo electrónico de Mozilla ya lo incorpora, creo que me lo voy a ahorrar ☺

(no es publicidad, es gratis)