CB 32

PROCEDIMIENTOS MATEMÁTICOS PARA EL FILTRADO DE CORREO ELECTRÓNICO

Erica SCHLAPS, Pedro WILLGING

Facultad de Ciencias Exactas y Naturales - Universidad Nacional de La Pampa
Av. Uruguay 151 - Santa Rosa - La Pampa
ericaschlaps@gmail.com pedro@exactas.unlpam.edu.ar

Palabras Clave: spam, filtros Bayesianos, probabilidad condicional, Teorema de Bayes.

RESUMEN

Una de las herramientas más populares, desde los inicios de la difusión y acceso masivos a Internet, ha sido el correo electrónico. Diariamente circulan por la red millones de mensajes, con datos relevantes para las actividades humanas, que se intercambian de manera casi instantánea entre personas de distintas partes del planeta. El correo electrónico es, hoy en día, una herramienta insustituible para muchas personas. Sin embargo, la proliferación de mensajes indeseados que invaden diariamente las casillas de correo de los usuarios, genera fastidio, pérdida de tiempo y dinero. Para evitar el problema del correo no deseado, se han intentado diferentes estrategias, algunas de las cuales se basan en procedimientos matemáticos. Los filtros Bayesianos han demostrado ser eficaces para controlar y clasificar los mensajes electrónicos. Se explicarán los fundamentos matemáticos subyacentes y los procedimientos de mejoramiento de los algoritmos involucrados, y las actuales líneas de investigación sobre la temática. Esta es una propuesta para introducir los conceptos de probabilidad condicional, el Teorema de Bayes y sus aplicaciones en problemas concretos en un curso de probabilidad en el nivel superior.

INTRODUCCIÓN

La gran difusión de Internet en todas las actividades de la sociedad ha conducido al uso extendido de herramientas asociadas con esta tecnología de la comunicación, siendo una de ellas el correo electrónico (e-mail). Desde su creación se convirtió en uno de los métodos más populares de distribución de información, tanto para actividades laborales en empresas como para comunicación inter-personal. En muchos casos se lo utiliza como una alternativa para los envíos postales, mensajes de teléfonos y fax.

Una explicación del rápido crecimiento que ha experimentado el empleo de correo electrónico es que tiene algunas ventajas sobre otras alternativas: es rápido, barato, no contamina el medio ambiente, es seguro y puede trasladar diversos tipos de archivos en múltiples formatos (datos adjuntos, fotos, videos). A medida que se ha ido incorporando como elemento cotidiano de trabajo o socializador en oficinas y hogares, sus prestaciones han ido mejorando, se ha facilitado su uso y se ha convertido en una utilidad indispensable y necesaria. El avance en los dispositivos móviles permite hoy en día, obtener la mensajería electrónica no solo en la computadora de la oficina o el hogar, sino en muchos otros artefactos, como los teléfonos celulares, las agendas electrónicas o las netbooks.

Del mismo modo que se puede enviar propaganda y ofertas de productos a la dirección postal sin requerir un permiso del destinatario, el correo electrónico es un efectivo medio para que

las empresas envíen mensajes no solicitados a potenciales consumidores. Estos mensajes no solicitados, no deseados o de remitente no conocido (correo anónimo), se los llama "spam".
Los mensajes de spam generan pérdidas de tiempo. Además del fastidio y la molestia que producen, a veces contienen virus o programas maliciosos o de espionaje lo que puede ser potencialmente dañino. Los hackers y otros delincuentes cibernéticos atraen al usuario para que visite un enlace y tratar de que ceda información personal, o incluso descarguen códigos maliciosos.
El primer e-mail spam fue enviado en 1978 a 320 miembros de red Arpanet, para promocionar la Corporación Digital Equipment. Según datos de junio del 2008, se enviaron alrededor de 110 mil millones de e-mails spam por día, y el 95.6 por ciento de los e-mails enviados a las empresas son spam. Esto les costó a las compañías 100 mil millones de dólares en el año 2007 (Fish, 2009). En un estudio se ha encontrado que el 35 por ciento de los usuarios de e-mail relató que más de 60 el por ciento de sus mensajes en la bandeja de entrada eran spam, y el 28 por ciento expresó que les insume más de 15 minutos diarios leerlos y eliminarlos (Kong, Rezaei, Sarshar, Roychowdhury, & Boykin, 2006).

Primeros intentos de controlar el Spam

En un intento de disuadir a los *spammers* (individuos o empresas generadoras de spam), se trató de generar legislación apropiada. Pero el nivel de spam no ha cambiado por la creación de nuevas leyes, ya sea porque la legislación no se aplica o porque las leyes no pueden alcanzar a quienes las infringen si son de otros países.
Tras el fracaso del intento de frenar el spam con las leyes, se comenzaron a crear métodos de filtrado. Una de las ventajas de los filtros es que, a diferencia de las leyes, los filtros son puestos en práctica por el usuario.
Los primeros métodos no estadísticos destinados a poner freno a los crecientes niveles de spam, fueron los llamados "métodos de filtrado heurísticos", los cuales son métodos de filtrado basado en reglas. Ellos trabajan buscando características específicas dentro de un mail y luego evalúan si es un spam. Estas características podrían estar relacionadas con el aspecto de una palabra o frase a partir de listas de palabras claves que indican que el mail es spam, dichas listas son llamadas "listas negras". Las listas negras poseen direcciones de e-mails o nombres de dominio que se cree son usados para enviar mails no solicitados.
Algunos de los filtros de spam que fueron creados basados en reglas son: DSPAM, SpamAssassin, Spam Bayes y ASSP.
El filtro SpamAssassin (http: //spamassassin.apache.org) utiliza todo el contexto de un mail buscando palabras o frases basado en la clasificación realizada por el usuario, pudiendo ser spam o ham (correo legítimo).
El filtro SpamNet (www.cloudmark.com) utiliza un modelo de servidor central, donde los usuarios cargan información sobre los spam a una base de datos central. Una desventaja de estos filtros es que utilizan la memoria colectiva, es decir, para cada nuevo spam enviado, algún usuario debe previamente identificarlo como spam (mediante una lista negra, por ejemplo) de manera tal que el futuro usuario que recibe el mail como sospechoso, lo clasifica como spam o ham, pudiendo cometer errores al realizar esta categorización.
Si bien los filtros heurísticos tuvieron algún éxito inicialmente, los spammers lograron eludirlos y es entonces que se pensó en desarrollar métodos de filtrado a partir de técnicas estadísticas, los cuales deberían ser más efectivos.
En el año 1998 Mehran Sahami fue el primero en hablar de filtros bayesianos. Este método ganó bastante atención en el año 2002 cuando Paul Graham lo describió en un artículo, donde usaba el Teorema de Bayes con el fin de deducir una fórmula para clasificar los e-mails.
A diferencia de los filtros heurísticos, las filtros Bayesianos utilizan un conjunto de reglas que ya no son estáticas, es decir, las reglas cambiarán porque cada usuario las adaptará según sus hábitos, lo cual aumenta la exactitud de la clasificación en un futuro. Una característica

importante del filtro Bayesiano es que hay que "entrenarlo" para mejorar la exactitud del filtro.
Ningún método de filtrado puede ser 100 por ciento eficiente, y debido a que el filtro o el usuario clasifica a un e-mail como spam o ham, existirán dos tipos de errores:

- una clasificación "falsa positiva" cuando se clasifica incorrectamente un e-mail legítimo como spam
- una clasificación "falsa negativa" cuando se clasifica incorrectamente un e-mail como legítimo cuando en realidad es spam.

En la mayoría de las situaciones un falso positivo es mucho más nocivo que un falso negativo debido a la pérdida potencial de información valiosa en la situación falsa positiva.

EL MÉTODO BAYESIANO

La lógica Bayesiana, creada a partir del trabajo del matemático inglés Thomas Bayes, se basa en las estadísticas y las probabilidades para predecir el futuro. El filtro Bayesiano determina la categoría (spam o ham) de un e-mail.
La clasificación puede ser entrenada sobre el cuerpo del spam *S* y otra en el cuerpo del ham *H*, desde los cuales se aprenden las apariencias de cada uno. Cuando un e-mail es recibido y analizado por el filtro, el primer paso, es realizar el proceso de "señalización" (conocido en inglés como *tokenization*), el cual consiste en dividir un e-mail compuesto por palabras. Y luego de este proceso, el filtro identifica las características del e-mail. Cada palabra (o *token*) es tomada individualmente y se calcula la probabilidad de que el e-mail sea spam dado que el e-mail contiene esa palabra o token. Por último, el filtro involucra todas las probabilidades individuales con el fin de *clasificar* un e-mail como spam o ham, y toma esa decisión calculando la probabilidad de que ese e-mail es spam dado que contiene un conjunto de palabras o tokens.

Teorema de Bayes[1]**:** *Sea* $B=\{B_1, B_2, ..., B_n\}$ *una partición del espacio muestral* Ω*, donde* $P(B_i)>0$ *para todo i. Si A es un suceso cualquiera contenido en el espacio muestral, se tiene:*

$$P(B_k / A) = \frac{P(A/B_k)\,P(B_k)}{P(A)} \qquad \textbf{(1)}$$

donde $P(A)=\sum_{i=1}^{n} P(A/B_i)\,P(B_i)$ *es el Teorema de la Probabilidad Total, en el cual* $P(B_i)$ *son las probabilidades a priori y* $P(B_i / A)$ *son las probabilidades a posteriori.*

Clasificación Bayesiana

La Clasificación Bayesiana usa probabilidades estimadas y el Teorema de Bayes para calcular la probabilidad de que el mail sea spam (o ham) suponiendo que los e-mails son generados por una función de distribución de probabilidad específica a su categoría con el parámetro desconocido θ. La consideración del parámetro θ se debe a que el mail está *condicionado* por el autor y por su categorización (ham o spam). Por ejemplo, si un e-mail contuviera un número significativo de características maliciosas tendería a ser clasificado como spam.

1 Thomas Bayes (1702-1761) fue uno de los seis primeros reverendos protestantes ordenados en Inglaterra. Comenzó como ayudante de su padre hasta que en 1720 fuera nombrado pastor en Kent. Abandonó los hábitos en 1752. Sus controvertidas teorías fueron aceptadas por Laplace, y posteriormente cuestionadas por Boole. Bayes fue elegido miembro de la Royal Society of London en 1742.

Un e-mail es una lista ordenada de palabras, donde W_i es la palabra en la *i*-ésima ubicación del e-mail. Si *D* es el suceso formado por todas las palabras contenidas en un e-mail, definimos los siguientes sucesos: $D=\{W_1,W_2,...,W_n\}$ (donde W_i es la palabra en la i-ésima ubicación del e-mail), S = "*el mail es spam"* y H = *"el mail no es spam (es ham)"*. Luego, $P(S/\theta)$ es la probabilidad de que el e-mail sea spam y, $P(H/\theta)$ es la probabilidad de que el e-mail sea ham; de manera tal que $P(S/\theta)=1-P(H/\theta)$.

La longitud del e-mail es una variable aleatoria denotada por *M*. Definimos $W=\{w_1,w_2,...\}$ como el conjunto de palabras o tokens que se consideran relevantes para la clasificación de un e-mail como spam.
La probabilidad de que las palabras o tokens estén en un e-mail, dado que éste es un spam es:

$$P(D/S;\theta)=P(W_1,W_2,...,W_M,M/S;\theta) \quad =P(M/\theta)\prod_{i=1}^{M}P(W_i/S;\theta;W_{i-1},...,W_1) \qquad \textbf{(2)}$$

donde $P(W_i/S;\theta;W_{i-1},...,W_1)$ para cada $i=1,2,...,M$, dependen de la categoría del e-mail, en este caso del spam *S*, y de las palabras precedentes $W_1,...,W_{i-1}$.
La probabilidad de que las palabras estén en un e-mail, es decir, del suceso $D=\{W_1,W_2,...,W_n\}$, sabiendo que θ es el parámetro desconocido, está dado por:

$$P(D/\theta)=P(D/S;\theta)P(S/\theta)+P(D/H;\theta)P(H/\theta) \qquad \textbf{(3)}$$

Volviendo al Teorema de Bayes, podemos calcular la probabilidad de que un e-mail sea spam dado que contiene la palabra W_i (o un conjunto de palabras, es decir, D) y considerando la distribución de la categoría con el parámetro desconocido θ, de la siguiente manera:

$$P(S/D;\theta)=\frac{P(S/\theta)P(D/S;\theta)}{P(D/\theta)} \qquad \textbf{(4)}$$

pues la partición del espacio muestral en **(1)** la generan las categorías $B_1=S$ y $B_2=H$ y D es el conjunto de palabras del e-mail conocidas a priori, por lo cual, $A=D$.
También calculamos la probabilidad de que un e-mail sea ham, sabiendo que contiene un conjunto de palabras, a partir de: $P(H/D;\theta)=1-P(S/D;\theta)$.
Luego, reemplazando **(3)** en **(4)**, la probabilidad de que un e-mail sea spam sabiendo que contiene un conjunto de palabras D es:

$$P(S/D;\theta)=\frac{P(S/\theta)P(D/S;\theta)}{P(D/S;\theta)P(S/\theta)+P(D/H;\theta)P(H/\theta)} \qquad \textbf{(5)}$$

Clasificación Ingenua Bayesiana
Para simplificar la ecuación **(5)**, supondremos que las posiciones de palabras del e-mail spam son independientes entre sí y del contexto (es decir., de las otras palabras), entonces:

$$P(W_i/S;\theta;W_{i-1},...,W_1)=P(W_i/S;\theta) \quad \text{con} \quad i=1,2,...,M \qquad \textbf{(6)}$$

Esta suposición hace que el Clasificador Bayesiano se llame Clasificador Bayesiano *Ingenuo*, lo cual conduce a reescribir la ecuación **(2)**:

$$P(D/S;\theta) = P(M/\theta)\prod_{W\in D} P(W/S;\theta) \quad \textbf{(7)}$$

Observemos que en **(7)** la notación "$W \in D$" indica aquellas palabras que se aparecen en el conjunto de todas las palabras del e-mail. Reemplazando **(7)** (y su equivalente para *H*) en la ecuación **(5)**, se tiene:

$$P(S/D;\theta) = \frac{P(S/\theta)P(M/\theta)\prod_{W\in D} P(W/S;\theta)}{P(S/\theta)P(M/\theta)\prod_{W\in D} P(W/S;\theta) + P(H/\theta)P(M/\theta)\prod_{W\in D} P(W/H;\theta)} \quad \textbf{(8)}$$

Lógicamente, $P(M/\theta)$ es cancelada, debido a que la longitud de un e-mail es independiente de si el e-mail es spam o ham, obteniendo:

$$P(S/D;\theta) = \frac{P(S/\theta)\prod_{W\in D} P(W/S;\theta)}{P(S/\theta)\prod_{W\in D} P(W/S;\theta) + P(H/\theta)\prod_{W\in D} P(W/H;\theta)} \quad \textbf{(9)}$$

Entrenamiento del Filtro Bayesiano

Con el fin de clasificar un e-mail como spam o no, todas las probabilidades individuales de cada palabra se calculan. Para calcular la ecuación **(9)** conociendo el parámetro θ, necesitamos saber la probabilidad de cada palabra en el vocabulario conocido $W = \{w_1, w_2, ...\}$, y también se conocen $P(W/S;\theta)$, $P(W/H;\theta)$ y la probabilidad de que un e-mail es spam $P(S/\theta)$ o es ham $P(H/\theta)$.

Notamos las probabilidades de la siguiente manera:

$$P(W = w/S;\theta) = \theta_w|_S \quad \textbf{(10)}$$

$$P(W = w/H;\theta) = \theta_w|_H \quad \textbf{(11)}$$

$$P(S/\theta) = \theta_S \quad \textbf{(12)}$$

$$P(H/\theta) = \theta_H \quad \textbf{(13)}$$

El cálculo del parámetro θ es mediante el estimador de θ, el cual denotamos $\hat{\theta}$ y se calcula a partir de la recopilación del cuerpo del spam *S* y del ham *H*. El número de ocurrencias de cada palabra se contará por separado en todo el cuerpo del spam y del ham, y cuando se encuentra una palabra que aún no ha aparecido en otros spam, se añade al vocabulario $W = \{w_1, w_2, ...\}$.

La probabilidad de que cualquier palabra aparezca en el vocabulario $W = \{w_1, w_2, ...\}$ del spam, puede ser estimada utilizando la idea de la interpretación frecuentista, la cual afirma que la "probabilidad" es el número de veces que aparece la palabra en el spam dividido el número total de ocurrencias de todas las palabras en el spam.

Bajo este razonamiento, la probabilidad de aparición de alguna palabra en el vocabulario $W=\{w_1,w_2,...\}$ del spam, dada por **(10)**, puede ser estimada por:

$$\hat{\theta}_{w|S} = \frac{\textit{número de w en S}}{\textit{número de todas las palabras en S}} \quad \textbf{(14)}$$

Si llamamos $N_S(w)$ al *número de veces que aparece la palabra en el spam*, podemos reescribir **(14)** de la siguiente manera:

$$\hat{\theta}_{w|S} = \frac{N_S(w)}{\sum_{v\in W} N_S(v)} \quad \textbf{(15)}$$

Análogamente, la probabilidad de alguna palabra en el vocabulario del $W=\{w_1,w_2,...\}$ ham, dada por **(11)**, puede ser estimada por:

$$\hat{\theta}_{w|H} = \frac{\textit{número de w en H}}{\textit{número de todas las palabras en H}} \quad \textbf{(16)}$$

Si llamamos $N_H(w)$ al *número de veces que aparece la palabra en el ham*:

$$\hat{\theta}_{w|H} = \frac{N_H(w)}{\sum_{v\in W} N_H(v)} \quad \textbf{(17)}$$

La probabilidad de que un e-mail sea spam $\hat{\theta}_S$ es el *número de e- mails spam* $N_{\hat{\theta}}(S)$ *dividido el número total de e-mails* $(N(S)+N(H))$, es decir:

$$\hat{\theta}_S = \frac{N(S)}{N(S)+N(H)} \quad \textbf{(18)}$$

La ecuación **(18)** también puede ser utilizada para calcular la probabilidad $\hat{\theta}_H$ de que un e-mail sea ham, sustituyendo S por H de la siguiente manera:

$$\hat{\theta}_H = \frac{N(H)}{N(S)+N(H)} \quad \textbf{(19)}$$

A partir de **(16)**, **(17)**, **(18)** y **(19)**, reescribimos **(10)**:

$$P\left(S/D;\hat{\theta}\right) = \frac{\hat{\theta}_S \prod_{W\in D} \hat{\theta}_{W|S}}{\hat{\theta}_S \prod_{W\in D} \hat{\theta}_{W|S} + \hat{\theta}_H \prod_{W\in D} \hat{\theta}_{W|H}} \quad \textbf{(20)}$$

EL FILTRO DE PAUL GRAHAM

Graham (2003) elaboró el artículo "A Plan for Spam", que ayudó a popularizar los filtros Bayesianos contra el spam. Este artículo describe las técnicas de filtrado de spam utilizados en un cliente de e-mail. Comienza con el cuerpo de un mail spam y de uno no spam. Por el

momento cada uno contenía 4000 mensajes. Escanea el texto completo, incluyendo encabezados html y javascript de cada mensaje. Considera caracteres alfanuméricos, guiones, apóstrofes, y el signo de dólar como parte de un token (o palabra), y todo lo demás como una marca de separación. Ignora los tokens que son todos dígitos, y también ignora comentarios html. Cuenta el número de veces que aparece cada token en cada cuerpo, guardándolas en dos diccionarios. En uno, al que se llama "Spam", guarda los tokens del cuerpo del correo spam. En el otro directorio llamado "Ham", guarda los tokens del cuerpo de correos legítimos. Esta etapa finaliza con dos listas grandes, una para cada cuerpo, recopilando las palabras con un número de ocurrencias.

En una tercera etapa, se crea la tercera lista, esta vez calculando la probabilidad de que un e-mail es spam dado que contiene la palabra. El cálculo de esta probabilidad es algo diferente a la de Bayes proporcionada por la ecuación **(10)**, ya que emplea pesos para evitar falsos positivos. Como queremos tener tan pocos falsos positivos como sea posible, multiplicamos por 2 la cantidad de apariciones de las palabras en el correo legítimo. De esta forma damos más peso a las palabras que aparecen en el correo legítimo, lo cual ayuda a distinguir entre palabras que ocasionalmente ocurren en e-mails legítimos y palabras que casi nunca ocurren. Sólo considera palabras que ocurren más de cinco veces en total (por el momento, a causa de duplicar, ocurriendo tres veces en los e-mails ham sería suficiente). Y luego se pregunta qué probabilidad asignarle a las palabras que ocurren en un cuerpo (spam o ham) pero no en el otro. Entonces la probabilidad de que el e-mail es spam dado que contiene la palabra W es:

$$\hat{\theta}_{S|W} = P\left(S / W; \hat{\theta}\right) = \frac{\hat{\theta}_{W|S}}{\hat{\theta}_{W|S} + 2\hat{\theta}_{W|H}} \qquad \textbf{(21)}$$

Por ejemplo, si la palabra "madam" aparece en 99 de 3000 spam y en 1 de 6000 ham, la probabilidad de ser spam está dada será:

$$P\left(S / W = "Madam"; \hat{\theta}\right) = \frac{\hat{\theta}_{W|S}}{\hat{\theta}_{W|S} + 2\hat{\theta}_{W|H}} = \frac{\frac{99}{3000}}{\frac{99}{3000} + 2\frac{1}{6000}} = 0.99$$

Se eligió 0.01 y 0.99, como límites de las probabilidades, a fin de evitar errores de estimación: $0.01 \le P\left(S / W; \hat{\theta}\right) \le 0.99$. Se supone, a partir de la recopilación de los 4000 e-mails, que $\hat{\theta}_S = \hat{\theta}_H = 0.5$. Puede haber lugar para mejorar esto, pero como el cuerpo crece, tales mejoras sucederán de cualquier manera automáticamente.

Cuando llega el nuevo e-mail, lo explora dentro de las palabras, y elige las quince más interesantes, donde "interesante" significa aquella palabra cuya probabilidad de ser spam es muy lejana a 0.5. A continuación Graham calcula la probabilidad combinada de las quince palabras. Una pregunta que surge es qué probabilidad asignarle a una palabra que nunca ocurrió en la tabla de probabilidades de palabras. Graham descubrió, por prueba y error, que 0.4 es un buen número para utilizar.

Un Ejemplo para ilustrar el procedimiento

Mientras Graham escribía su artículo le llegó un spam publicitario (http://lib.store.yahoo.net/lib/paulgraham/spam2.txt), el cual analizó.

Las probabilidades calculadas de las quince palabras más "interesantes" a partir de **(21)** son:

Palabra	Probabilidad
madam	*0.99*
promotion	*0.99*
republic	*0.99*

shortest	*0.047225013*
mandatory	*0.047225013*
standardization	*0.07347802*
sorry	*0.08221981*
supported	*0.09019077*
people's	*0.09019077*
enter	*0.9075001*
quality	*0.8921298*
organization	*0.12454646*
investment	*0.8568143*
very	*0.14758544*
valuable	*0.82347786*

Tabla 1. Probabilidad de algunas palabras del e-mail

Si combinamos todas las probabilidades tenemos:

$$P\left(S/D;\hat{\theta}\right)=\frac{\prod_{W\in D}\hat{\theta}_{S}\big|_{W}}{\prod_{W\in D}\hat{\theta}_{S}\big|_{W}+\prod_{W\in D}\hat{\theta}_{H}\big|_{W}}=$$

$$=\frac{(0.99)(0.99)(0.99)(0.047225013)...(0.8568143)(0.14758544)(0.82347786)}{(0.99)(0.99)(0.99)...(0.14758544)(0.82347786)+(1-0.99)(1-0.99)(1-0.99)...(1-0.14758544)(1-0.82347786)}=0.9027 \quad \textbf{(22)}$$

Luego, $P\left(S/D;\hat{\theta}\right)=0.9027$, es decir, la probabilidad de que el e-mail que contiene estas 15 palabras sea spam es 0.9027.

CONCLUSIONES

El filtro Bayesiano ofrece múltiples ventajas que lo hacen mejor que otros métodos de detección de spam. Tiene en cuenta la totalidad del e-mail (reconoce palabras clave que identifican el spam, pero también reconoce palabras que denotan si es ham). Está constantemente "auto-adaptándose" porque evoluciona y se adapta a nuevas técnicas spam. Por ejemplo, cuando los spammers comenzaron a utilizar "f-r-e-e" en lugar de "free" consiguieron eludir los análisis de palabras hasta que "f-r-e-e" fue incluido en la base de datos de las palabras. Es difícil de burlar, a diferencia de un filtro en base a listas negras y blancas.
Actualmente se están intentando otros métodos basados en pares de palabras, o aún en triples, más que en palabras individuales, llamados "Filtros Markovianos". En estos casos se utiliza la teoría de cadenas de Markov, que está teniendo difusión también en áreas de lingüística computacional.
El problema del filtrado del e-mail por medio de técnicas estadísticas y probabilistas es un tema de investigación actual y una estrategia motivadora para introducir este temario de la currícula en los cursos universitarios que lo incluyen.

BIBLIOGRAFÍA

- BERNARDO, J. M. 1981. Bioestadística. Una perspectiva Bayesiana. Barcelona: Vicens-Vives S.A.
- CHUNG, K.L. 2001. A course in probability theory. San Diego, CA: Academic Press.
- DURRETT, R. 1995. Probability theory with examples. 2nd Edition. Belmont, CA: Duxbury, Press.
- FISH, J. 2009. Bayesian Markovian spam filtering. Disponible: http://maths.dur.ac.uk/Ug/projects/library/CM3/000509947r.pdf
- GRAHAM, P. 2003. A Plan for Spam. Disponible: http://lib.store.yahoo.net/lib/paulgraham/spam2.txt

- KONG, J. S., REZAEI, B. A., SARSHAR, N., ROYCHOWDHURY, W. P, y BOYKIN, P. O. 2006. Collaborative spam filtering using e-mail networks. Computer, 39(8), 67-73. UC Los Angeles.
- MUKHOPADHYAY, N. 2000. Probability and Statistical Inference. New York: Marcel Dekker, Inc.
- O'CONNOR, B. 2007. Markovian spam filtering. Disponible: http://maths.dur.ac.uk/Ug/projects/library/CM3/000424248t.pdf
- RINCÓN, L. 2007. Curso intermedio de Probabilidad. México DF. Disponible: http://www.matematicas.unam.mx/lars/libros/cip.pdf
- Wikipedia. Bayesian spam filtering.Wikipedia. Disponible: http://en.wikipedia.org/wiki/Bayesian_spam_filtering