Inferencia Gramatical para la Detección de Spam

Eduardo GROSCLAUDE

Departamento de Informática y Estadística
Universidad Nacional del Comahue
Buenos Aires 1400 - (8300) Neuquén, Argentina. FAX: (54)(0299)4490313
e-mail: oso@uncoma.edu.ar

Resumen

El *spam* representa un problema de mala utilización de recursos técnicos y una molestia para los usuarios de correo electrónico. Tomando este problema como aplicación práctica, se pretende mostrar, con su justificación teórica, decisiones de diseño de un posible sistema inteligente destinado a controlarlo. El trabajo que se describe está actualmente en curso en el marco de un proyecto de investigación de la Universidad Nacional del Comahue.

Palabras clave: Inteligencia Computacional, Machine Learning, Inferencia Gramatical.

Introducción

Los mensajes de correo electrónico no solicitados (*spam*) son una vía de comunicación comercial usada con creciente frecuencia para la promoción de bienes y servicios. El atractivo de las campañas publicitarias mediante *spam* reside, para los anunciantes, en el bajo costo de llegada. Quizás por esta razón, la focalización o *targeting* de estos mensajes es por lo común inexistente. El usuario final de un sistema de correo electrónico se ve financiando, a través de sus pagos por costos de comunicaciones, campañas de promoción de productos que no le interesan en absoluto. A esto se agrega el abuso de recursos tales como el ancho de banda de las redes, el espacio de almacenamiento en servidores, recursos lógicos de los sistemas de comunicación (que deben demorar el tráfico genuino) y el tiempo que los usuarios invierten en descartar los mensajes no deseados y reorganizar sus buzones. La actitud general de la comunidad usuaria de Internet es considerar al *spam* como una actividad éticamente criticable.

No existiendo a la fecha herramientas legales útiles para manejar este problema, se han efectuado varias aproximaciones técnicas. Estas consisten por lo general en la escritura manual, a cargo de un administrador o usuario, de un conjunto de reglas de filtrado, predicadas sobre el contenido o metadatos del mensaje [1]. Esta técnica estática tiene inconvenientes obvios para adaptarse a los cambios, y no es fácil para el usuario corriente producir conjuntos de reglas consistentes [2].

Este trabajo se propone investigar una nueva solución al problema del *spam* mediante técnicas inteligentes de inferencia gramatical.

Modelo de los individuos

La amplia práctica desarrollada en clasificación de textos por métodos inteligentes ha dado buenos resultados en el dominio especial de los mensajes de correo electrónico [2, 3, 4, 5]. Las soluciones presentadas se basan generalmente en la extracción y manipulación de características de vocabulario, considerado como la evidencia de una determinada semántica inmanente en los documentos. La popularidad de los métodos de aprendizaje bayesianos, simples y sin embargo robustos, ha extendido

el uso de la representación de individuos como conjuntos de palabras, sin orden, mapeadas a sus frecuencias observadas [3, 6, 7, 8].

Sin embargo, en este dominio se cuenta con evidencia adicional aportada por la estructura o morfología de los individuos. Los mensajes de *e-mail* tienen una estructura normatizada por una cantidad de especificaciones [9]. Por ejemplo, los diferentes items presentes en la cabecera de los mensajes indican emisor, destinatario, asunto y otros metadatos del mensaje, y los formatos de los medios no textuales que se adjuntan quedan especificados en forma legible dentro del cuerpo. Un experto humano que identifica *spam* utiliza señales presentes en la cabecera de los mensajes, tales como la indicación del asunto, direcciones remitentes fraguadas, envío en difusión a listas de interés (lo cual maximiza el alcance de un *spam*), o remitentes conocidos y "aceptados" (evidencia de *nospam*).

Por otro lado, es posible que los métodos clásicos de aprendizaje no capturen todas las características útiles. El experto humano, al abrir un mensaje y a primera vista, puede determinar si lo eliminará de su buzón por no interesante haciendo uso de muy pocas claves y aun sin leerlo, es decir, con prescindencia de su contenido. Uno puede, introspectivamente, preguntarse qué características son las que permiten tomar esta decisión subjetiva. Por ejemplo, la abundancia de imágenes adjuntas, la longitud del mensaje, la utilización de formatos de hipertexto, diseño visualmente impactante (técnicas mixtas, colores, tipos grandes), etc. Todas estas características escapan a la vista de un *learner* cuyo único interés es el vocabulario.

Agregando complejidad al modelo descriptivo de los individuos para reflejar la estructura, podrán utilizarse nuevas formas de inferencia automática para obtener una clase diferente de conocimiento.

Gramáticas probabilísticas

Un modelo de aprendizaje completamente general, que contiene tanto información de lingüística como de estructura de los individuos, es el de las gramáticas probabilísticas [10, 11]. En este modelo se considera a los individuos como cadenas generadas por una gramática ideal, y se busca inducir, en base a los individuos, un conjunto de reglas, afectadas por un determinado grado de incertidumbre, que aproximen la gramática ideal. En un momento posterior, nuevos individuos son sometidos a un analizador basado en dicha gramática que emite una clasificación con un nivel de confianza explícito.

El trabajo que se describe tiene como objetivo explorar los fundamentos que influenciarían el diseño de un sistema de inferencia gramatical con varias restricciones interesantes, algunas propias del problema de la detección del *spam*. A continuación se comentan las principales consideraciones de diseño que anticipamos.

Considerando	***Problemática***	***Herramientas teóricas***
Espacio de búsqueda	El análisis informal de los individuos debe sugerir una clase de gramáticas que los modele adecuadamente para los propósitos de la aplicación. Factores de complejidad orientan la elección a una clase determinada de gramáticas. Existen resultados sobre aproximación de una clase por ejemplares de otra. Se cuenta con ejemplos positivos y negativos, lo que según resultados de la teoría permite enfrentar adecuadamente el problema.	***Teoría chomskiana. Concepto de bias inductivo [3]. Conceptos de identificación en el límite [14] y aprendizaje PAC [8]. Clases de lenguajes aprendibles [12, 13, 14, 15, 16]. Resultados sobre aproximación de LC por gramáticas regulares [17, 22].***
Mecanismos de inferencia	Proponer y seleccionar mecanismos de generalización basados en estructura. Existen	***Algoritmos de inferencia de autómatas [10, 12, 13, 18, 19,***

Considerando	*Problemática*	*Herramientas teóricas*
	diversos métodos de inferencia completa de gramáticas regulares.	***20]. Teoría Inferencial del Aprendizaje [21]. Aprendizaje bayesiano [3, 10].***
Representación de individuos	Proponer y seleccionar mecanismos de generalización basados en vocabulario. Preparación de las experiencias de entrenamiento.	Teoría de la Información y principio de Mínima Longitud de Descripción (MDL) [18]. Principios de trabajo del campo de Information Retrieval. Clustering. Técnicas de análisis estadístico multivariado.
El problema del falso positivo	El problema exhibe muy poca tolerancia hacia los errores de falsos positivos. No es aceptable para un usuario que el sistema filtre equivocadamente un mensaje que era de su interés. Se impone una consideración de costos en las metas de aprendizaje.	Aprendizaje sensitivo a costos.

Comentarios finales

Se pretende explorar, desde la teoría pero con una motivación práctica, un enfoque relativamente poco transitado de la clasificación de textos. El alcance del trabajo en curso, tal cual está planteado, no involucra necesariamente una implementación. Es la intención, sin embargo, continuar en etapas posteriores con aspectos de desarrollo de una solución según lo aprendido en esta primera etapa.

Referencias

1.Lindberg G., *RFC 2505 Anti-Spam recommendations for SMTP MTAs*
2.Rennie J., ifile: *An Application of Machine Learning to E-Mail Filtering*
3.Mitchell T., *Machine Learning*
4.Sahami M., Dumais S., Heckerman D., Horvitz E., A *Bayesian Approach to Filtering Junk E-Mail*
5.de Vel O., *Mining E-Mail Authorship*
6.McCallum A., Nigam K., *A Comparison of Event Models for Naive Bayes Text Classification*
7.Koller D., Sahami M., *Hierarchically classifying documents using very few words*
8.Michalewicz M., *A Postgraduate Course on Data Analysis and Artificial Intelligence*
9.Freed N., Borenstein N., *RFC 2045 MIME: Format of Internet Message Bodies*
10.Stolcke A., Omohundro S., *Inducing Probabilistic Grammars by Bayesian Model Merging*
11.Stolcke A., *Bayesian Learning of Probabilistic Language Models*, thesis
12.Lee L., *Learning of Context-Free Languages: A Survey of the Literature*
13.Dupont P., *Incremental Regular Inference*
14.Dupont P., Miclet L., *Inférence grammaticale regulière: fondements théoriques et principaux algorithmes*
15.Firoiu L. et al., *Learning Regular Languages from Positive Evidence*
16.Coste F., A*pprentissage d'Automates Classifieurs en Inférence Grammaticale*, thèse
17.Pereira F., Wright R., *Finite-State Approximation of Phrase-Structure Grammars*, en *Finite State Language Processing*, edit. E. Roche and Y. Schabes
18.Grünwald P., *A Minimum Description Length Approach to Grammar Inference*
19.Dupont P., *Grammatical Inference: formal and heuristic methods*
20.Luzeaux D., *A universal approach to positive grammar inference*
21.Michalski R., *A Theory and Methodology of Inductive Learning*
22.Abney S., McAllester D., Pereira F., *Relating Probabilistic Grammars and Automata*