Sistema Adaptativo con Etiquetado Inteligente para la Clasificación de Correo Spam

José Ramón Méndez Reboredo

Director: Dr. Florentino Fernández Riverola
Departamento de Informática
Escuela Superior de Ingeniería Informática
Universidad de Vigo
moncho.mendez@uvigo.es

Resumen

A continuación se presenta un resumen de la tesis doctoral cuyo título coincide con el de este trabajo. En ésta se presenta un modelo híbrido de Inteligencia Artificial para el problema de detección y filtrado de correo *spam*. Esta tesis fue leída el 21 de septiembre de 2006, obteniendo la calificación de Sobresaliente cum Laude.

Palabras clave: IBR, clasificación, *spam*, GTC, PWV, EIRN, instancias, indexación, SAEICS, SPAMHUNTING.

1. Introducción

En este trabajo se presenta un sistema híbrido de Inteligencia Artificial capaz de detectar y filtrar mensajes *spam*. Debido a la naturaleza volátil del concepto *spam* (*concept drift*), resulta importante contar con herramientas capaces de adquirir dinámicamente conocimiento sobre el dominio, descartando aquel que, con el paso del tiempo, se vuelve obsoleto. Por otro lado, teniendo en cuenta que el concepto *spam* es disjunto, los modelos de filtrado deben incorporar técnicas de selección y representación disjunta del conocimiento.

En este trabajo se ha demostrado que, en el ámbito del filtrado de correo *spam*, un modelo híbrido que cumpla las características indicadas anteriormente puede resultar más eficaz, que la utilización de otras técnicas de Inteligencia Artificial. Teniendo en cuenta estas características y, con el objetivo de llevar a cabo el desarrollo de un filtro de mensajes *spam*, se ha mostrado conveniente el empleo de una metodología para la construcción de un sistema CBR (*Case-Based Reasoning*, Sistema de Razonamiento Basado en Casos) [Aamodt&Plaza 94].

Concretamente, el modelo propuesto se basa en la utilización de un sistema IBR (*Instante-Based Reasoning*, Sistema de Razonamiento Basado en Instancias) que incorpora una estructura de indexación de mensajes, una estrategia de votación, un mecanismo para el cálculo de la calidad de las soluciones generadas y una técnica para la identificación y eliminación del conocimiento irrelevante, con el objetivo de obtener un alto nivel de precisión y permitir una adaptación rápida ante cambios que se produzcan en el entorno.

2. Contribuciones

Los trabajos de investigación han culminado con la creación de SPAMHUNTING, un filtro capaz de analizar correos en base a los términos más relevantes que componen cada mensaje y que incorpora las características indicadas con anterioridad. La detección de términos relevantes de un correo se fundamenta en la combinación de

información sobre la relevancia del término dentro del mensaje y en el contexto semántico presente en el momento de recepción del mensaje.

Los resultados de esta investigación reflejan la importancia de la actualización continua del conocimiento como instrumento útil para evitar los efectos derivados del *concept drift*. Este fenómeno, que puede ser entendido como la pérdida de representatividad de un término sobre una categoría, resulta difícil de detectar sin recurrir al uso de estrategias de aprendizaje continuo. La eliminación de los términos que pierden representatividad en una categoría (afectados por el *concept drift*), así como la adquisición de otros que pueden aportar nuevo conocimiento, permite un mejor funcionamiento de los filtros de mensajes *spam*.

Por otro lado, un esquema de representación disjunto de los mensajes es la pieza clave que permite el empleo de mecanismos para descartar el conocimiento sobre términos que dejan de ser relevantes en el problema. Este enfoque permite incorporar dinámicamente nuevos términos relevantes, y realizar una selección de características acorde con el contexto semántico establecido por el marco temporal de recepción de los mensajes.

Las técnicas clásicas de filtrado determinan la clase de los mensajes en base a los términos que contienen. Este funcionamiento resulta poco adecuado en situaciones de ataque a filtros de correos *spam* mediante técnicas estadísticas o palabras dispersas. Por otro lado, al basar la clasificación en la presencia de términos, resulta más difícil identificar mensajes legítimos que contienen algunos términos frecuentes en mensajes *spam*. En el modelo propuesto, la determinación de la categoría de un mensaje se lleva a cabo en base a los correos más similares a uno dado.

Durante este trabajo de investigación, se ha creado un corpus multilingüe que combina de una forma equilibrada mensajes *spam* y legítimos. Este corpus contiene más de 20000 mensajes recopilados durante los años 2004, 2005 y 2006 por varios miembros de una comunidad universitaria.

A modo de síntesis, se puede afirmar que este trabajo ha aportado un conjunto de indicaciones guía, de vital importancia para la construcción de filtros de mensajes *spam*, junto con el desarrollo de un modelo que incorpora y saca partido de estas características.

3. Estructura

El trabajo realizado se presenta organizado en siete capítulos y tres apéndices. En el primer capítulo se define, en líneas generales, el problema que se pretende solucionar y se establece la hipótesis defendida.

En el capítulo dos se propone el problema a resolver y sus características inherentes, subrayando el interés que suscita el descubrimiento de técnicas eficaces que permitan remediarlo. Se presenta, además, una profusa descripción sobre la complejidad inherente al problema planteado, las fuentes de información disponibles y su formato así como los tratamientos empleados para preprocesar estos datos.

El capítulo tres se centra en el estudio de los mecanismos actuales de filtrado de correo electrónico que representan el estado del arte en este campo. El estudio contempla un compendio de métodos colaborativos (basados en la cooperación de usuarios) y centrados en el contenido (que se fundamentan en el estudio de los términos que componen los mensajes).

El capítulo cuatro se centra en el estudio de los sistemas IBR y su posible aplicación como metodología de desarrollo para la construcción del sistema híbrido propuesto. Se presentan los sistemas de razonamiento basados en casos como marco en el que se encuadran los sistemas IBR, exponiendo una posible clasificación de este tipo de modelos. Se estudia su ciclo de operación y se identifican las tecnologías de IA compatibles con cada etapa. Se realiza un estudio a nivel teórico de las distintas posibilidades a la hora de integrar mecanismos de IA, presentando las distintas clasificaciones existentes. Finalmente, se justifica la adecuación de los sistemas IBR como metodología para la construcción de un sistema híbrido de clasificación y se presenta una recopilación de los trabajos más relevantes desarrollados en este campo.

En el capítulo cinco se presenta el modelo híbrido de clasificación propuesto, como combinación del sistema IBR que utiliza una red EIRN (*Enhanced Instance Retrieval Network*, Red de Recuperación de Instancias Mejorada), un sistema de votación PWV (*Porportional Weigthed Voting*, Voto de Peso Proporcional) y un mecanismo para el mantenimiento del conocimiento GTC (*Garbagge Term Recolection,* Recolección de Términos Basura). Se describe en detalle la representación de

una instancia para el problema planteado y el funcionamiento general del modelo. A continuación se detallan, para cada etapa del sistema IBR, las técnicas empleadas y su integración dentro del sistema completo.

El capítulo seis detalla los resultados obtenidos con el modelo final propuesto y su comparación con las distintas técnicas analizadas para la resolución del problema planteado. El análisis realizado contempla 4 escenarios diferentes. En primer lugar, se analizan los resultados obtenidos mediante una evaluación estática de los modelos considerados. Posteriormente, se presenta un análisis de la eficacia de los modelos que incorporan la capacidad de aprender continuamente. En tercer lugar, se lleva a cabo una comparación de la capacidad de adaptación de los modelos evaluados ante cambios bruscos del entorno y, finalmente, se presenta una evaluación de los efectos del uso de diferentes idiomas para cada uno de los sistemas analizados.

Por último, el capítulo siete presenta las conclusiones extraídas de la aplicación del modelo desarrollado al problema planteado. Se detallan los objetivos alcanzados, la aplicabilidad del modelo presentado y las repercusiones directas de la investigación llevada a cabo. Finalmente, se especifican las líneas de trabajo futuro a desarrollar, utilizando como punto de partida la investigación presentada en el trabajo.

4. Resumen y Conclusiones

El modelo desarrollado está basado en un sistema IBR, que encapsula una estructura de recuperación de instancias (EIRN), una técnica de reutilización basada en votación (PWV), un esquema de cálculo del grado de validez para cada solución generada y un mecanismo de identificación de términos basura (GTC). Cada una de estas técnicas se utiliza en una fase diferente del ciclo de razonamiento clásico presentado por [Aamodt&Plaza 94], con el fin de recuperar situaciones pasadas, adaptarlas al problema actual (generando una clasificación inicial), elaborar una respuesta revisada definitiva y, finalmente, retener el conocimiento obtenido a partir de la clasificación realizada.

En el sistema propuesto, cada correo electrónico se representa mediante un descriptor. Éste está formado por una serie de valores numéricos y textuales que se obtienen de forma automática a partir del contenido del mensaje objetivo y de los mensajes recibidos por el sistema. Para cada descriptor, el sistema debe llevar a cabo un ciclo completo de razonamiento. A este descriptor de problema, generado cada vez que el sistema recibe un nuevo mensaje, se le denomina instancia-problema.

Para la generación de cada instancia-problema, se lleva a cabo una etapa de selección de aquellos términos más relevantes del mensaje recibido con el fin de identificar los términos que mejor representan su contenido semántico. En esta etapa se tienen en cuenta cuestiones de representatividad de los términos tanto en el contexto temporal de la recepción del mensaje (*concept drift*), como en el marco semántico establecido por el propio correo electrónico (contenido del mensaje). Debido a la naturaleza del proceso de selección de términos relevantes y el carácter disjunto de los conceptos *spam* y legítimo, los vocablos escogidos para dos instancias-problema distintas y su número no tienen por qué coincidir.

Cada vez que se recibe un correo electrónico se presenta una nueva instancia-problema al sistema. La red EIRN identifica, mediante una proyección, las instancias (situaciones anteriores) almacenadas en el sistema IBR más similares al problema actual. Las instancias recuperadas se caracterizan porque al menos uno de sus términos relevantes es también relevante en la instancia-problema a clasificar. Debido a la estructura de la red EIRN utilizada como mecanismo de indexación y cálculo de similitud, el número de instancias inicialmente seleccionadas es variable y depende del número de términos relevantes presentes en la instancia-problema así como de la densidad de instancias asociadas a estos términos. Dado que el nivel de similitud de algunas de las instancias recuperadas con el problema objetivo puede resultar muy bajo, durante la etapa de recuperación se seleccionan sólo aquellos que comparten un mayor número de características relevantes con la instancia-problema.

Durante la fase de adaptación, las instancias recuperadas en la fase anterior se emplean para determinar la clase del nuevo mensaje. De esta forma, esta etapa comienza con la realización de una ordenación previa de las instancias según su grado de similitud estructural con el problema objetivo. La medida de similitud propuesta se centra en el análisis de la importancia de los términos que cada instancia recuperada y la instancia-problema tienen en común. Finalmente, se aplica el esquema de votación PWV que otorga mayor peso al voto

emitido por las instancias recuperadas más similares al nuevo mensaje a clasificar.

Durante las fases de recuperación y adaptación el sistema genera una solución con la mayor precisión posible. En este sentido, la fase de revisión se concibe con el objetivo de determinar la calidad de solución final obtenida y no como una forma de mejorar la exactitud de una respuesta que ha sido generada empleando de forma rigurosa todo el conocimiento disponible sobre el dominio del problema. Concretamente, el mecanismo de revisión implementado analiza en que medida las instancias recuperadas pueden representar a la instancia-problema.

El proceso de retención o aprendizaje, se lleva a cabo tras haber determinado la clase del problema objetivo. Durante esta fase se crea una nueva instancia añadiendo a la instancia-problema la clasificación generada por el sistema, y se incorpora a la memoria del sistema IBR para su uso posterior. El aprendizaje del sistema se complementa con la ejecución periódica de un proceso de edición de la base de conocimiento denominado GTC, que permite eliminar las características e instancias indexadas en la red EIRN que han dejado de ser relevantes para clasificar nuevos mensajes.

La justificación de la hipótesis defendida en este trabajo se hace de forma experimental, empleando tres corpus de prueba diferentes y más de 30000 mensajes. Los resultados obtenidos a partir de los experimentos realizados con el modelo desarrollado, se comparan con los generados mediante la utilización de distintas técnicas clásicas de Inteligencia Artificial, lo que permite la realización de un análisis cuantitativo y cualitativo de la eficacia y la eficiencia del sistema propuesto.

A la vista de los resultados obtenidos, se concluye que la aplicación del modelo híbrido defendido es de especial interés en el ámbito del problema estudiado. En este sentido, el modelo desarrollado genera resultados adecuados y es capaz de adaptarse de forma dinámica a las características cambiantes del entorno. Finalmente, se aporta un conjunto de ideas clave para afrontar la construcción de sistemas de filtrado de correo *spam* que derivan del trabajo realizado.

Los resultados de esta investigación han sido divulgados en varios congresos y revistas internacionales [Méndez et al. 06a, Méndez et al. 06b, Riverola et al. 07a, Riverola et al. 07b]. Próximamente se publicarán, empleando estos medios, distintos aspectos y ventajas relevantes de la propuesta realizada que todavía no se han dado a conocer.

Agradecimientos

Este trabajo se ha realizado gracias a la inestimable ayuda de varios profesores y alumnos de la Universidad de Vigo que han cedido gran cantidad de correos electrónicos clasificados para la realización de las pruebas. Por otro lado, el trabajo desarrollado ha sido financiado por el proyecto titulado *SAEICS: Sistema Adaptativo con Etiquetado Inteligente para el Correo Spam* dentro del programa de ayudas a grupos de investigación emergentes de la Universidad de Vigo.

Referencias

[Aamodt&Plaza 94] Aamodt, A., y Plaza, E. 'Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches'. AI Communications, 7(1):39-59. (1994).

[Méndez et al. 06a] Méndez, J. R., Fdez-Riverola, F., Iglesias, E. L., Díaz, F. y Corchado, J. M. 'Tracking Concept Drift at Feature Selection Stage in SpamHunting: an Anti-Spam Instance-Based Reasoning System'. En Proceedings of the 8th European Conference on Case-Based Reasoning, ECCBR-06 (pp. 504-518). (2006).

[Méndez et al. 06b] Méndez, J. R., Fdez-Riverola, F., Díaz, F., Iglesias, E. L. y Corchado, J. M. 'A Comparative Performance Study of Feature Selection Methods for the Anti-Spam Filtering Domain'. En Proceedings of the 6th Industrial Conference on Data Mining, ICDM-2006 (pp. 106-120). (2006).

[Riverola et al. 07a] Riverola, F, Iglesias, E. L., Díaz, F., Méndez, J. R. y Corchado, J. M. 'SpamHunting: An Instance-Based Reasoning System for Spam Labelling and Filtering'. Decision Support Systems. (In press).

[Riverola et al. 07b] Fdez-Riverola, F., Iglesias, E. L., Díaz, F., Méndez, J. R. y Corchado, J. M. 'Applying Lazy Learning Algorithms to Tackle Concept Drift in Spam Filtering'. Expert Systems With Applications. 33(1). (In press).