# Motores de Búsqueda Web
# Tarea Tema 4: Adversarial IR

**José Alberto Benítez Andrades**

**71454586A**

**Motores de Búsqueda Web**

**Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web**

**UNED**

**20/02/2011**

20 de febrero de 2011

# Tarea Tema 4: Adversarial IR

## Enunciado del ejercicio

El tema al que los grandes buscadores dedican más esfuerzos en la actualidad es a combatir el "spamdexing", es decir, el uso de técnicas para manipular los resultados de los buscadores. Se conoce como "adversarial IR" el campo, dentro de la recuperación de información, que trata sobre cómo combatir el spamdexing, es decir, cómo crear algoritmos de ranking que sean robustos a los intentos de manipulación.

Echa un vistazo a los proceedings de la última edición del AIRWeb (congreso sobre adversarial IR) y comenta cuáles son las principales líneas de trabajo en este tema.

La URL del congreso es: http://airweb.cse.lehigh.edu/

## Resolución

**Looking into the Past to Better Classify Web Spam** — *Na Dai, Brian D. Davison and Xiaoguang Qi*

Técnicas de spamming web destinadas a alcanzar una clasificación inmerecida en los resultados de búsqueda. La investigación ha sido realizada ampliamente en la identificación de spam tal y neutralizar su influencia. Sin embargo, los trabajos existentes de detección de spam solamente tienen en cuenta la información actual. Se argumenta que la información histórica de páginas web también puede ser importante en la clasificación de spam. En este trabajo, utilizaron sus autores funciones de contenido de las versiones históricas de páginas web para mejorar la clasificación de spam. Utilizaron técnicas de aprendizaje supervisadas para combinar clasificadores basados en el contenido de la página actual con clasificadores basados en las características temporales. Los experimentos llevaron a que su enfoque mejora la clasificación de spam mejorando el rendimiento medio en un 30% en comparación con un clasificador de referencia, que sólo tiene en cuenta el contenido de la página actual.

**A Study of Link Farm Distribution and Evolution Using a Time Series of Web Snapshots** — *Young-joo Chung, Masashi Toyoda and Masaru Kitsuregawa*

En este trabajo se estudia la estructura global de spam basada en el enlace y su evolución, que sería de gran ayuda para el desarrollo de herramientas de análisis robustas y de investigación de spam Web como una actividad social en el espacio cibernético. En primer lugar, el uso de componentes fuertemente conectados (SCC) para separar la descomposición en muchas granjas de enlace de la mayor SCC, por lo que se llama el núcleo. Se muestra que

las explotaciones más densas en el núcleo de enlace se puede extraer por el nodo de filtrado y la aplicación recursiva de descomposición SCC hasta la médula. Sorprendentemente, podemos encontrar nuevas granjas de enlaces grandes en cada iteración y esta tendencia se mantiene hasta por lo menos 10 iteraciones. Además, sus autores midieron la "spamicidad" de tales granjas de enlaces. A continuación, la evolución de las granjas de enlaces se examinaron más de dos años. Los resultados muestran que casi todas las granjas de enlaces grandes ya no crecen, mientras que algunos de ellos se reducen, y muchas granjas de enlaces grandes se crean en un año.

**Web Spam Filtering in Internet Archives** — *Miklós Erdélyi, András A. Benczúr, Julien Masanes and Dávid Siklósi*

Mientras que el spam Web está dirigido por el alto valor comercial de los motores de búsqueda, se puede observar el deterioro de la calidad de los archivos web y el desperdicio de recursos como efecto secundario. Hasta ahora las tecnologías de filtrado Web de spam se usan muy poco en los archivos Web, pero tiene un gran futuro, como se indica en una encuesta con respuestas de más de 20 instituciones en todo el mundo. Estos archivos suelen operar en un nivel modesto de presupuesto que prohíbe la operación de spam independiente de filtrado Web, pero los esfuerzos de colaboración podrían conducir a una solución de alta calidad para ellos. En este artículo se ilustran las necesidades del filtrado de spam, las oportunidades y los bloqueadores de los archivos de Internet a través de análisis de varias instantáneas de rastreo y la dificultad de migrar a través de diferentes modelos de filtros se arrastra a través del ejemplo de los 13. Instantáneas del Reino Unido realizado por UbiCrawler que incluyen webspam-UK2006 y webspam- UK2007. Análisis de contenido.

**Web Spam Identification Through Language Model Analysis** — *Juan Martinez-Romo and Lourdes Araujo*

En este trabajo se aplica un enfoque de modelo de la lengua a diferentes fuentes de información extraída de una página Web, con el fin de proporcionar indicadores de alta calidad en la detección de Spam Web. Dos páginas unidas por un enlace que debe ser relacionado con el tópico, a pesar de que se tratara de una relación contextual débil. Por este motivo, sus autores han analizado las diferentes fuentes de información de una página Web que pertenece al contexto de una relación y han aplicado la divergencia de Kullback-Leibler de ellos para caracterizar la relación entre dos páginas enlazadas. Por otra parte, se combinan algunas de estas fuentes de información con el fin de obtener modelos de lenguaje más rico. Dada la diferente naturaleza de los enlaces internos y externos, en nuestro estudio también se distinguen estos tipos de vínculos conseguir una mejora significativa en las tareas de clasificación. El resultado es un sistema que mejora la detección de spam en la web de dos grandes conjuntos de datos y el público como webspam-UK2006 y webspam UK2007.

**An Empirical Study on Selective Sampling in Active Learning for Splog Detection** — *Taichi Katayama, Takehito Utsuro, Yuuki Sato, Takayuki Yoshinaka, Yasuhide Kawada and Tomohiro Fukuhara*

Este trabajo estudia cómo reducir la cantidad de supervisión humana para la identificación de splogs / blogs auténticos en el contexto de la continua actualización de los datos splog establece año tras año. A raíz de los trabajos previos en el aprendizaje activo, en contra de la tarea de splog / detección blog auténtico, este trabajo examina empíricamente varias estrategias para el muestreo selectivo en el aprendizaje activo de máquinas de soporte vectorial (SVM). Como una medida de confianza del MSV aprendizaje, cuentan con la distancia del hiperplano que separa a cada instancia de prueba, que han sido bien estudiados en el aprendizaje activo para la clasificación de texto. A diferencia de los resultados de la aplicación de aprendizaje activo a las tareas de clasificación de textos, en la tarea de splog / detección auténtico blog de este trabajo, no es el caso de que la adición de al menos las muestras de confianza en un mejor rendimiento.

**Linked Latent Dirichlet Allocation in Web Spam Filtering** — *István Bíró, Dávid Siklósi, Jácint Szabó and András Benczúr*

Latent Dirichlet allocation (LDA) (Blei, Ng, Jordan 2003) es un modelo estadístico completamente generativo del lenguaje sobre el contenido y los temas de un corpus de documentos. En este artículo se aplicará una extensión de LDA para la clasificación web, correo no deseado. La vinculación técnica LDA lleva también la vinculación en cuenta: los temas se propagan a lo largo de enlaces de tal manera que el documento vinculado influye directamente en las palabras en el documento de vinculación. El modelo de deducir LDA se puede aplicar para la clasificación como la reducción de dimensionalidad de manera similar a la indexación semántica latente. Probamos vinculados LDA en el spam Web-UK2007 corpus. Mediante el uso de clasificador BayesNet, en términos de las AUC de la clasificación, conseguimos 3% de mejoría sobre plano LDA con BayesNet, y 8% con respecto a las características de enlace público con C4.5. La adición de este método para un diario de probabilidades de combinación basada en la estrecha relación y los resultados de referencia el contenido de los clasificadores en una mejora del 3% en el AUC. Nuestro método incluso mejora ligeramente en el mejor Web Spam Challenge 2008 resultado.Social Spam

**Social Spam Detection** — *Benjamin Markines, Ciro Cattuto and Filippo Menczer*

La popularidad de los sitios bookmarking sociales los ha convertido en un blanco perfecto para los spammers. Muchos de estos sistemas requieren un tiempo de un administrador y la energía para filtrar de forma manual o eliminar el spam. Aquí hablan de las motivaciones de spam social, y presentan un estudio de detección automática de los spammers en un sistema de etiquetado social. Se identifican y analizan seis características diferentes que se ocupan de diversas propiedades de spam sociales, encontrando que cada una de estas características proporciona una señal de ayuda a los spammers discriminación de los usuarios legítimos. Estas características se utilizan en varios algoritmos de aprendizaje máquina para la clasificación, alcanzando más del 98% de precisión en la detección de los

spammers social con 2% positivos falsos. Estos prometedores resultados proporcionan una nueva base para los futuros esfuerzos sobre el spam social. Hacemos nuestra base de datos a disposición del público a la comunidad científica.

**Tag Spam Creates Large Non-Giant Connected Components** — *Nicolas Neubauer, Robert Wetzker and Klaus Obermayer*

Los creadores de spam en los sistemas de marcadores sociales tratan de imitar el comportamiento de marcadores de usuarios reales para ganar la atención de otros usuarios o de los motores de búsqueda. Varios métodos han sido propuestos para la detección de spam tales, incluidas las características de dominio específico (como términos URL) o similitud de los usuarios a los spammers previamente identificados. Sin embargo, como se muestra en su trabajo anterior, es posible identificar una gran parte de los usuarios de correo no deseado basándose en las características puramente estructurales. El hipergrafo conexión documentos, los usuarios, y las etiquetas se pueden descomponer en componentes conectados, y grandes hay, pero los componentes no-gigante resultó ser habitado casi en su totalidad por los usuarios de correo no deseado en el conjunto de datos examinados. En este sentido, se prueba hasta qué punto la descomposición de la hipergrafo completa es realmente necesaria, examinar la estructura de componentes del usuario inducida / usuarios y documentos / gráficos de etiquetas. Mientras que la conectividad del usuario gráfico de etiquetas no ayuda en la clasificación de los spammers, la conectividad del usuario / documento gráfico ya está muy informativo. No obstante, se puede aumentar con la información de conectividad de la hipergrafo. En la opinión de los autores, la detección de spam basado en las características estructurales, como la que aquí se propone, requiere de complejas estrategias de adaptación de los spammers y pueden complementar otros métodos, la detección más tradicionales.

**Spam Research Collections**

**Nullification Test Collections for Web Spam and SEO** — *Timothy Jones, David Hawking, Ramesh Sankaranarayana and Nick Craswell*

La investigación en el área de recuperación de información contradictoria ha sido facilitada por la disponibilidad de los UK-2006/UK-2007 colecciones, que abarcan datos de rastreo, el gráfico de enlace, y las etiquetas de correo no deseado. Sin embargo, la investigación de anular el efecto negativo del spam u optimización excesiva motor de búsqueda (SEO) en el ranking de las páginas que no son spam no ha tenido el apoyo de estos recursos. Tampoco es el estudio de las técnicas de encubrimiento o de spam. Por último, el carácter de dominio restringido de un rastreo del Reino Unido. Significa que solo una parte de los icebergs granja de enlaces pueden ser visibles en estos rastreos. Se introduce la anulación, término que se define como "la prevención de las páginas es un problema que afecta de manera negativa los resultados de la búsqueda". Muestran algunas diferencias importantes entre las propiedades de la corriente. Uk-restringida se arrastra y los reportados previamente para la Web en su conjunto. Identificamos la necesidad de una colección IR contradictorio que no es de dominio restringido y que es apoyado por un conjunto de conjuntos de consultas apropiadas y (con optimismo) los datos de comportamiento de los usuarios. El rastreo sin

restricciones de millones de páginas se están llevando a cabo por el CMU (web09-bst) y se utilizarán para evaluar como posible base para una colección nueva prueba de AIR. Se discuten los pros y los contras de su escala, y la posibilidad de agregar recursos como las listas de consulta para mejorar la utilidad de la colección para la investigación de AIR.

**Web Spam Challenge Proposal for Filtering in Archives**— *András A. Benczúr, Miklós Erdélyi, Julien Masanes and Dávid Siklósi*

En este artículo proponen sus autores nuevas tareas para un posible desafío futuro de Web Spam motivado por las necesidades de la comunidad archivística. La comunidad archivística Web consta de varias instituciones relativamente pequeñas que operan de forma independiente y, posiblemente, en diferentes dominios de nivel superior (TLD). Cada uno de ellos puede tener un gran conjunto histórico de rastreos. Eficaz filtrado por lo tanto, sería necesario (1) un mayor uso de las series temporales de las instantáneas de dominio y la colaboración (2) mediante la transferencia de modelos a través de diferentes dominios de nivel superior. Correspondientes tareas desafío por lo tanto, podría incluir la distribución de datos de instantáneas de rastreo para la generación de función, así como la clasificación de nuevo sin etiqueta se arrastra de los TLD diferentes mismo o incluso..

## Referencias

1. http://airweb.cse.lehigh.edu/