# Confianza y desconfianza en redes sociales: Detección de *trolls*

F. Javier Ortega, Fermín L. Cruz, José A. Troyano, Carlos G. Vallejo, Fernando Enríquez

*Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla*

**<{javierortega, fcruz, troyano, vallejo, fenros}@us.es>**

**Resumen:** Uno de los aspectos más importantes en la empresa 2.0 es la gestión de la reputación *on-line*. En este trabajo, nos enfrentamos a los comportamientos deshonestos que se pueden llevar a cabo en la web 2.0 para alterar la reputación de los usuarios (o entidades) de una red social. La principal novedad de este trabajo consiste en la habilidad de procesar una red *on-line* con relaciones positivas y negativas entre sus usuarios. En este trabajo proponemos un sistema de reputación de usuarios que tiene en consideración estas relaciones positivas y negativas. Este sistema se basa en un algoritmo de ranking sobre grafos que construirá una lista de los usuarios de la red social, ordenados según su reputación. Aparte de los resultados mostrados, esta técnica presenta algunas ideas novedosas que pueden ser adaptadas a otras tareas donde las relaciones positivas y negativas entre elementos sean relevantes. Algunas de estas aplicaciones también se discuten brevemente, destacando la aplicabilidad de nuestra investigación.

**Palabras clave:** Algoritmos basados en grafos, confianza y desconfianza, redes sociales.

**Autores**

**Francisco Javier Ortega Rodríguez** es Doctor en Informática, actualmente investigador en el departamento de Lenguajes y Sistemas Informáticos en la Universidad de Sevilla. Su investigación está enfocada principalmente al Análisis de Redes Sociales y el Procesamiento del Lenguaje Natural. Sus trabajos más recientes están relacionados con la minería de opiniones, el análisis de redes sociales, la detección de *spam* en la web y el cálculo de la reputación y fiabilidad de usuarios en redes sociales.

**Fermín L. Cruz-Mata** es Doctor en Ingeniería Informática y profesor Colaborador en la Universidad de Sevilla. Actualmente su investigación se centra en el análisis del sentimiento y la minería de opiniones, pero también ha trabajado en temas relacionados con el Procesamiento del Lenguaje Natural, como el etiquetado basado en grafos y el etiquetado de roles semánticos.

**José Antonio Troyano Jiménez** es Doctor en Informática y Profesor Titular del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla. Sus líneas de investigación están centradas en la aplicación de técnicas de aprendizaje automático al tratamiento de información no estructurada. Sus últimos trabajos tratan sobre el análisis de opiniones en textos generados por usuarios, análisis de reputación en redes sociales y detección de tendencias en Twitter.

**Carlos Antonio García Vallejo** es Doctor en Informática por la Universidad de Sevilla. Centra su investigación en algoritmos sobre grafos aplicados al Aprendizaje Automático y sus aplicaciones al Análisis de Redes Sociales. Sus últimos trabajos están relacionados con la minería de opiniones, la clasificación y la selección de instancias.

**Fernando Enríquez de Salamanca Ros** trabaja como profesor e investigador en la Universidad de Sevilla, en la que obtuvo su título de Doctor en Ingeniería Informática. Su línea de investigación está enfocada principalmente en la aplicación del aprendizaje automático a tareas de Procesamiento del Lenguaje Natural, y más concretamente en la combinación de diferentes algoritmos de clasificación para el análisis de textos.

## 1. Introducción

La Empresa 2.0 es un concepto que engloba todos los aspectos relacionados con un nuevo paradigma en la gestión empresarial que adopta algunas de las ideas fundamentales de la llamada Web 2.0, integrándolas en el modelo empresarial, de forma que se puedan aprovechar estos nuevos esquemas de comunicación y actuación, no sólo dentro de la propia empresa, sino también en el proceso comunicativo entre la empresa y sus clientes, los socios o con otras empresas. Dentro de este ámbito, resulta de gran interés el uso de las plataformas de redes sociales en el ámbito empresarial, que posibilita un contacto más directo entre la empresa y sus potenciales clientes.

Otra tarea interesante que puede ser abordada gracias al uso de estas tecnologías es la búsqueda de expertos [18] en un sistema *on-line* mediante el análisis de las interacciones de los usuarios de la red y las valoraciones que pueden hacer sobre las aportaciones de otros usuarios.

Si bien este tipo de sistemas puede reportar muchos beneficios, también traen consigo algunos problemas que deben ser tenidos en cuenta. Uno de estos problemas es la gestión de la reputación *on-line*. Al igual que en su día ocurría (y sigue ocurriendo) en la web con las páginas de *spam*, las plataformas de redes sociales vienen sufriendo determinados "ataques" por parte de ciertos usuarios, encaminados a alterar el funcionamiento normal de estos sistemas, o incluso obtener algún tipo de beneficio de ellos, como por ejemplo un aumento de su popularidad, o para dañar la reputación de determinadas entidades de la red, ya sean personajes públicos, marcas, productos o empresas.

Los sistemas de Gestión de la Reputación *on-line* (TRS, de las siglas en inglés de *Trust and Reputation Systems*) son los encargados de tratar de evitar estas situaciones indeseadas.

Este problema no es nuevo, y existen diversas propuestas para tratar de abordarlo, por ejemplo la creación de un equipo de moderadores en foros *on-line* o redes sociales de noticias, donde a una serie de usuarios se les otorga la capacidad (y la responsabilidad) para expulsar a aquellos usuarios que ellos consideren malintencionados o "*trolls*". Estos mecanismos suelen ser efectivos en comunidades *on-line* pequeñas, ya que al aumentar el número de usuarios e interacciones en la red, es necesario aumentar el número de moderadores. Por lo tanto este sistema no es escalable a grandes comunidades *on-line*.

Otros sistemas de gestión de la reputación *on-line* tratan de evitar esta limitación delegando las tareas de moderación a todo el conjunto de usuarios de la red, de forma que los usuarios pueden valorar a otros usuarios o los contenidos que estos generan. La mayor ventaja de estos sistemas es su descentralización y, por tanto, su escalabilidad. Existen diversos sistemas para tratar esta información que proveen los usuarios de una red, propagando las opiniones o valoraciones de cada usuario siguiendo la topología de la red de forma que se calcule un valor de fiabilidad para cada usuario, según las

> "En este trabajo mostramos un sistema de gestión de la reputación basado en la propagación de las opiniones positivas y negativas de los usuarios de una red"

opiniones del resto. Por otro lado, la mayor desventaja de estos sistemas es que no existe una autoridad central que arbitre determinados comportamientos que pueden ser considerados ilícitos, como por ejemplo el hecho de que un grupo de usuarios puedan formar una coalición para valorar positiva o negativamente a otro(s).

En este trabajo mostramos un sistema de gestión de la reputación basado en la propagación de las opiniones positivas y negativas de los usuarios de una red, tratando de obtener un ranking de usuarios de forma que se promocione a aquellos usuarios más fiables según la comunidad, y se penalice a aquellos usuarios con una baja consideración dentro del sistema.

## 2. Trabajos relacionados

Como se resalta en [19], el uso de tecnologías provenientes de la Web 2.0 en el entorno empresarial puede reportar muchos beneficios. La adopción de plataformas sociales provee a la empresa de nuevas fuentes de información que pueden resultar decisivas, tanto en la relación de empresa a empresa, como de empresa a cliente. Y, por supuesto, la integración de estos sistemas dentro de los procesos de negocio de una empresa también hace posible un mayor conocimiento del capital humano dentro de la propia empresa, que puede ayudar a la hora de buscar el agente humano más apropiado para cada tarea, por ejemplo.

Tanto para esta tarea como para una adecuada gestión de la reputación *on-line* de una empresa, los sistemas de reputación en redes sociales cobran una gran relevancia en el ámbito de la Empresa 2.0. En este entorno de reputación *on-line*, un primer paso podría ser la gestión de la identidad digital, es decir, la unificación de los distintos perfiles u opiniones sobre una misma empresa en diversas plataformas *on-line* [5]. Una vez resuelto este primer paso, la etapa siguiente consistiría en tener una idea de la reputación *on-line* de la empresa.

Centrándonos en el cálculo de la reputación de entidades en redes sociales, hay un buen número de trabajos que estudian el tema, incluyendo los mayores desafíos que esta tarea presenta. En esta línea, los trabajos en [6][7][8][10] analizan algunas de las dificultades asociadas a esta tarea, como el sesgo que suele existir en las redes sociales hacia una mayoría de opiniones positivas o negativas, según la temática de la misma.

También resaltan la ausencia en muchos casos de incentivos que animen a los usuarios del sistema a enriquecer la información contenida en el mismo mediante sus valoraciones u opiniones. Esta ausencia de información es uno de los mayores problemas de estos sistemas. Otros trabajos estudian la naturaleza de las relaciones en determinadas redes sociales, que suele influir enormemente en las estrategias para el cálculo de la reputación. Por ejemplo en [2] se estudia el fenómeno conocido como "relaciones nepóticas", consistente en que un usuario da una opinión positiva a aquellos que han opinado positivamente sobre él.

Por otro lado, existen en la bibliografía ejemplos de algoritmos pensados para calcular la reputación de los elementos de una red, teniendo en cuenta diversos aspectos del sistema. Podemos destacar algunos de los más relevantes, como por ejemplo EigenTrust [10], TidalTrust [3] o PowerTrust [19]; todos ellos tratan al cálculo de la reputación de los usuarios teniendo en cuenta las opiniones y relaciones establecidas en la red social, la mayoría de una forma directa, es decir, que la puntuación de cada usuario variará según las valoraciones de los usuarios con los que tenga contacto directo en la red.

Otros trabajos están más enfocados al estudio de la propagación de la información a través de una red, analizando de qué forma se pueden transferir las opiniones positivas y negativas de unos usuarios sobre otros a través de la red, de forma que estas opiniones sirvan para obtener valoraciones globales de los usuarios del sistema.

En concreto, en [4] presentan el concepto de transitividad de la desconfianza, que representa las diferentes formas en las que las opiniones negativas se propagan a través de una red. En el trabajo mencionado las resumen en tres: transitividad multiplicativa, que se podría enunciar como "*El enemigo de mi enemigo es mi amigo*"; luego tenemos la transitividad aditiva, o "*No confiar en alguien en quien no confía alguien en quien tú no confías*"; y por último estaría la transitividad neutral, que consiste en "*No tener en cuenta de las opiniones de alguien en quien no se confía*". Estos tres modelos se pueden representar gráficamente como se muestra en la **figura 1.**

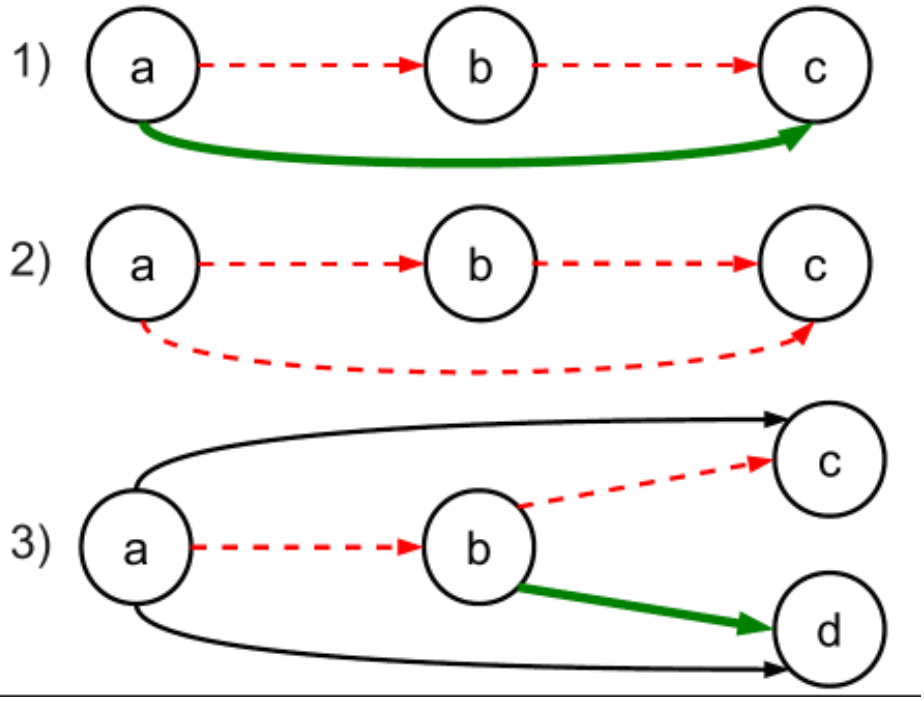


**Figura 1.** Modelos de transitividad de la desconfianza: (1) Desconfianza transitiva. (2) Desconfianza aditiva. (3) Desconfianza neutral. Las líneas discontínuas se corresponden con opiniones negativas, mientras que las líneas gruesas con opiniones positivas y las líneas delgadas son opiniones indefinidas.

> "PolarityTrust está basado en ideas similares a las de PageRank, el conocido algoritmo utilizado en un principio para ordenar las páginas web en los resultados de búsquedas de Google"

En [13] se muestra el procesado de la reputación de los usuarios de una red social real extraída de la web de noticias Slashdot.org, llegando a la conclusión de que la red formada por los usuarios del sistema y sus opiniones positivas y negativas tienen una estructura multiplicativa. En el citado trabajo presentan un método para el cálculo de la reputación en esta red, si bien no tienen en cuenta la transitividad de las opiniones negativas en sus cálculos.

Por último, otros trabajos estudian los problemas que los sistemas de cálculo de la reputación *on-line* deben resolver, dado el entorno cambiante y dinámico en el que se desenvuelven. En [10] estudian lo que ellos llaman "modelos de ataques" que usuarios malintencionados pueden ejecutar contra un sistema de reputación *on-line*, tratando de aprovecharse del mecanismo para obtener algún tipo de beneficio.

Estos modelos clasifican las posibles acciones de los usuarios malintencionados en cinco tipos:

- **Modelo A:** Existen usuarios con una baja reputación en el sistema, de forma aislada. Sería el modelo normal de comportamiento de usuarios en una red social.
- **Modelo B:** Los usuarios malintencionados forman una coalición para ganar una alta reputación dando buenas opiniones sobre los demás miembros de la coalición.
- **Modelo C:** Los usuarios malintencionados pueden comportarse de manera correcta esporádicamente ante algunos usuarios con el objetivo de ganar buena reputación.
- **Modelo D:** Se introduce un nuevo tipo de usuario malintencionado, que se encarga de ganar una buena reputación, para luego opinar positivamente de otros usuarios malintencionados, de forma que la reputación de éstos se vea incrementada.
- **Modelo E:** Los usuarios malintencionados no sólo tratan de ganar una alta reputación, sino que también intentan bajar la reputación de los demás.

En nuestra propuesta trataremos de evitar en la medida de lo posible los efectos de estos comportamientos en una red social.

## 3. Nuestra propuesta

En esta sección presentaremos nuestra propuesta de un sistema de cálculo de la reputación de usuarios en redes sociales: PolarityTrust [14][15]. Básicamente se trata de un sistema basado en la propagación de información a través de la red formada por los usuarios de un sistema y las relaciones entre ellos, que pueden ser positivas o negativas dependiendo de la semántica de la red.

Utilizando la topología de la red es posible calcular la confiabilidad de un usuario dentro del sistema. En nuestro caso utilizamos dos valores para ello, $PT^+$ y $PT^-$: el primer valor representará la buena reputación de ese usuario en la red, y análogamente, el otro valor representa la mala reputación del usuario en la red.

La combinación de ambas puntuaciones sirve para calcular la reputación global del usuario. El objetivo de nuestro sistema de cálculo de la reputación *on-line* es evitar que usuarios malintencionados se aprovechen del sistema para ganar de forma ilícita una muy buena reputación, o que provoquen un aumento de la mala reputación a otros usuarios.

### 3.1. Modelo principal

PolarityTrust está basado en ideas similares a las de PageRank [16], el conocido algoritmo utilizado en un principio para ordenar las páginas web en los resultados de búsquedas de Google.

En nuestra propuesta hemos extendido sus capacidades de forma que podamos procesar grafos con aristas positivas y negativas, además de posibilitar la inclusión de diferentes esquemas de propagación, como los vistos en la **sección 2**, a la hora de ejecutar el algoritmo de paso aleatorio, de tal forma que podemos determinar de qué manera las opiniones (positivas o negativas) de los usuarios van a influenciar las puntuaciones de otros usuarios.

Así pues, la fórmula que caracteriza la reputación de un usuario $v_i$ de una red social, según nuestra propuesta es la que se muestra en la **figura 2**.

Donde $p_{ij}$ es el valor de la opinión del usuario $v_i$ sobre el usuario $v_j$. $In^+(v_i)$ y $In^-(v_i)$ representan el conjunto de usuarios con opiniones positivas y negativas sobre $v_i$, respectivamente. $Out(v_j)$ es el conjunto de usuarios sobre los que opina $v_j$. Por último, d es el factor aleatorio del algoritmo de PageRank, que determina la probabilidad de saltar a un nodo de la red que no esté directamente relacionado con el actual; y los valores $e^+_i$ y $e^-_i$ pueden entenderse en este caso como la probabilidad, a priori, de que el nodo $v_i$ tenga una buena o mala reputación, respectivamente. En nuestra propuesta, $e^+_i = 1$ sólo si el usuario $v_i$ es completamente fiable.

Por lo tanto, un requerimiento de PolarityTrust es la existencia de un conjunto inicial de usuarios fiables. En el entorno de las redes *on-line*, esta es una suposición factible, ya que podemos tomar como totalmente fiables al conjunto de usuarios moderadores, o a los administradores del sistema *on-line*, de forma que mediante este algoritmo sus opiniones se propagan con una mayor fuerza a lo largo de la red. Como veremos en la sección de evaluación, este conjunto puede ser bastante pequeño en relación al tamaño total de la red.

$$PT^+(v_i) = (1-d)e_i^+ + d(+\sum_{j \in In^+(v_i)} \frac{p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PT^+(v_j) + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PT^-(v_j))$$

$$PT^-(v_i) = (1-d)e_i^- + d(\sum_{j \in In^+(v_i)} \frac{p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PT^-(v_j) + \sum_{j \in In^-(v_i)} \frac{-p_{ji}}{\sum_{k \in Out(v_j)} |p_{jk}|} PT^+(v_j))$$

**Figura 2.** Fórmula propuesta para caracterizar la reputación de un usuario en una red social.

> "A diferencia de otras propuestas, puede verse en la fórmula que en nuestro sistema las opiniones positivas y las negativas afectan tanto a la puntuación positiva como a la negativa de un usuario"

Como puede comprobarse, PolarityTrust adopta por defecto un esquema multiplicativo de transitividad a la hora de propagar las opiniones negativas. Además, y a diferencia de otras propuestas, puede verse en la fórmula que en nuestro sistema las opiniones positivas y las negativas afectan tanto a la puntuación positiva como a la negativa de un usuario. De esta forma tratamos de minimizar algunos de los efectos negativos que los ataques vistos anteriormente pueden provocar en la reputación de los usuarios de una red social.

### 3.2. Modelos adicionales

Una vez expuesto el modelo básico de PolarityTrust, pasaremos a explicar brevemente las características de los modelos adicionales [15] desarrollados para evitar en la medida de lo posible los efectos perniciosos de las acciones llevadas a cabo por usuarios malintencionados en el sistema.

En primer lugar, hemos dotado al modelo básico de una forma de evitar la propagación al resto de nodos de la red de las opiniones negativas de los usuarios malintencionados. De esta forma se intentan atajar los ataques del tipo E (ver **sección 2**), mediante los cuales los usuarios malintencionados tratan de dañar la reputación de otros usuarios. Este modelo lo llamamos "modelo de propagación no-negativa".

El otro gran problema con el que tenemos que lidiar en estos sistemas es el llamado "opiniones deshonestas", que consiste en que algunos usuarios pueden dar de forma intencionada opiniones negativas sobre usuarios buenos, o viceversa. Para evitar este problema hemos añadido el modelo de "acción-reacción" de PolarityTrust, que se basa en penalizar a aquellos usuarios que realizan este tipo de acciones. Asumiendo que el otorgar una opinión positiva a usuarios malintencionados (o viceversa) también puede deberse a un error o incluso a ser objeto de un ataque de tipo C, nuestro modelo será flexible a la hora de aplicar dichas penalizaciones. De esta forma, los usuarios serán penalizados de forma proporcional al número de veces que han cometido este tipo de acciones.

## 4. Evaluación

En esta sección detallaremos los experimentos llevados a cabo para validar nuestra propuesta, comenzando con la presentación de los conjuntos de datos utilizados para ello, los sistemas con los que hemos comparado PolarityTrust, las métricas utilizadas en la evaluación y por último mostrando los resultados obtenidos.

### 4.1. Datos de evaluación

Para testear el comportamiento de PolarityTrust y de otros sistemas para el cálculo de la reputación de usuarios en una red social, hemos utilizado un conjunto de datos que han sido usados en otros trabajos [13], extraído de una red social real, Slashdot.org, consistente en los usuarios de esa red social y las relaciones (positivas o negativas) existentes entre ellos.

En cuanto a las dimensiones del conjunto de datos para la evaluación, en total cuenta con unos 71.500 usuarios, con más de 510.000 relaciones entre ellos, de las cuales alrededor de un 24% son negativas.

Una de las ventajas de utilizar estos datos es la existencia en la red social Slashdot de una lista de usuarios malintencionados, mantenida por los moderadores, con lo que facilita la evaluación de estos sistemas. En total, la muestra obtenida cuenta con 96 *trolls* conocidos. Por lo tanto, parte de la evaluación consistirá en identificar correctamente a estos usuarios, haciendo que aparezcan en las últimas posiciones del ranking de reputación de usuarios.

Adicionalmente a este conjunto de datos, le hemos añadido una serie de usuarios malintencionados, modelando los ataques vistos en la **sección 2**, de forma que podamos evaluar la robustez de todos los sistemas analizados.

### 4.2. Otros sistemas

Para realizar una comparativa con nuestro sistema, hemos elegido tres sistemas de cálculo de la reputación *on-line* de distintas características.

En primer lugar, hemos implementado un modelo básico basado en la experiencia directa de los usuarios. Se trata de una métrica a la que hemos llamado "*Fans minus Foes*" (amigos menos enemigos), consistente simplemente en restar el número de opiniones positivas y negativas sobre un usuario.

Por otro lado, hemos realizado una implementación del algoritmo de EigenTrust [10], que tiene en cuenta la estructura de la red ejecutando un algoritmo de paso aleatorio simplificado, pero basa su cálculo en una métrica parecida a la anterior.

Por último, compararemos nuestra propuesta con NegativeRanking [13], un algoritmo de paso aleatorio que tiene en cuenta en su ejecución las opiniones positivas y negativas de la red de forma separada, para luego calcular una puntuación de reputación para cada nodo.

### 4.3. Métrica

La forma de evaluar los resultados obtenidos por cada técnica consiste en construir un ranking de usuarios, según las puntuaciones calculadas con cada algoritmo, y comparar estos rankings con el ranking ideal (aquel cuyas últimas posiciones están ocupadas por los usuarios considerados malintencio-nados, o *trolls*). Esta métrica se denomina *Normalized Discounted Cumulative Gain* (nDCG), y sigue la siguiente fórmula:

$$nDCG_p = \frac{relevance_1 + \sum_{i=2}^{p} \frac{relevance_i}{log_2 i}}{IDCG_p}$$

donde *p* es una posición dada del ranking. El valor de *relevance* será 1 si el usuario que ocupa esa posición es bueno, y 0 en caso de ser un usuario malintencionado. Finalmente, *IDCG* es el valor del nDCG ideal, es decir, el valor del ranking perfecto. Cuanto mayor sea el valor de nDCG, mejor será el resultado obtenido por la técnica analizada.

### 4.4. Resultados

En esta sección mostraremos los resultados obtenidos por cada técnica, de forma que podamos evaluar el comportamiento de cada una de ellas en las situaciones que nos hemos planteado.

El primer experimento, cuyos resultados se muestran en la **tabla 1**, aplica cada técnica al conjunto de datos de Slashdot, de forma que se intenten identificar los usuarios malintencionados que ya están en la red. Para ello tomamos como usuarios totalmente fiables al creador de Slashdot.org (su nombre de usuario en el sistema es *CmdrTaco*) y a aquellos usuarios a los que *CmdrTaco* apunta en su lista de amigos (en total 6 usuarios fiables).

Adicionalmente, hemos incluido el porcentaje de usuarios malintencionados que han obtenido posiciones en el ranking más altas

> "Una de las ventajas de utilizar estos datos es la existencia en la red social Slashdot de una lista de usuarios malintencionados, mantenida por los moderadores, con lo que facilita la evaluación de estos sistemas"

| Métricas | ET | FmF | NR | PT |
|---|---|---|---|---|
| Error % | 0,99 | 0,901 | 0,881 | 0,861 |
| nDCG | 0,31 | 0,46 | 0,479 | 0,588 |

**Tabla 1.** Porcentaje de errores y nDCG para cada técnica probada. ET = *EigenTrust*; FmF=*Fans minus Foes*; NR=*Negative Ranking*; PT=*PolarityTrust*.

| Threats | ET | FmF | NR | PT |
|---|---|---|---|---|
| A | 0,31 | 0,46 | 0,477 | 0,588 |
| A,B | 0,308 | 0,46 | 0,477 | 0,588 |
| A,B,C | 0,311 | 0,46 | 0,484 | 0,588 |
| A,B,C,D | 0,37 | 0,476 | 0,501 | 0,586 |
| A,B,C,D,E | 0,37 | 0,475 | 0,496 | 0,588 |

**Tabla 2.** nDCG obtenido por cada técnica para cada ataque propuesto.

que las que les correspondería en el caso ideal. Como puede verse, en este primer experimento los resultados de PolarityTrust son prometedores, demostrando ya una mejora con respecto al resto de sistemas.

En segundo lugar repetimos el experimento, esta vez añadiendo a los datos originales una serie de ataques, efectuados por los 96 *trolls* identificados en la red, de forma que traten de ganar buena reputación aprovechándose del sistema. Los resultados pueden verse en la **tabla 2**.

Podemos ver cómo los sistemas a priori más simples se ven bastante afectados por estos ataques, mientras que PolarityTrust mantiene su buen rendimiento.

| Ataques | PT |
|---|---|
| A | 0,846 |
| A,B | 0,846 |
| A,B,C | 0,846 |
| A,B,C,D | 0,782 |
| A,B,C,D,E | 0,781 |

**Tabla 3.** nDCG para PolarityTrust, utilizando información sobre usuarios malintencionados conocidos.

Por último, en la **tabla 3** mostramos los resultados obtenidos por PolarityTrust añadiéndole información sobre una serie de usuarios malintencionados conocidos. La información de estos usuarios se consigna en el parámetro $e^-$ de la fórmula vista en la **sección 3**. En estos experimentos hemos incluido en el algoritmo información sobre 5 usuarios de entre los que están anotados como *trolls* conocidos.

Resulta evidente y notable la mejora al utilizar información sobre algunos usuarios malintencionados conocidos dentro del algoritmo. Esto da una idea el margen de mejora que tiene nuestro sistema de ser utilizado a lo largo del tiempo en una red social.

## 5. Conclusiones

La principal contribución de este trabajo de investigación es el desarrollo de PolarityTrust, un algoritmo de ranking sobre grafos con la capacidad de procesar una red con opiniones positivas y negativas entre sus usuarios, y además de penalizar a usuarios malintencionados que tratan de aprovecharse del sistema de reputación. Estas características hacen que PolarityTrust presente una gran flexibilidad, siendo capaz de ejecutar distintos tipos de propagación de información sobre una red, dependiendo de la naturaleza de las relaciones entre los usuarios de la misma.

Una tarea a abordar en un trabajo futuro será la búsqueda de expertos dentro de una corporación. En esta tarea puede ser interesante aplicar los esquemas de PolarityTrust vistos en este trabajo pero utilizando esquemas de propagación aditivos, de forma que las opiniones de los usuarios sirvan para reforzar o debilitar la imagen de un usuario como experto en determinado tema.

## Agradecimientos

Esta investigación ha sido parcialmente financiada por los proyectos: TIN2012-38536-C03-02 y TIN2011-14726-E.

## Referencias


**[1] N. Agarwal, H. Liu, L. Tang, P. S. Yu.** "Identifying the influential bloggers in a community". *Proc. ACM WSDM Conf*., pp. 207¨C218, 2008.

**[2] D. Gayo-Avello.** "*Nepotistic Relationships in Twitter and their Impact on Rank Prestige Algorithms*". Universidad de Oviedo, 2010. <http://arxiv.org/ftp/arxiv/papers/1004/1004.0816.pdf>.

**[3] J. A. Golbeck.** "Computing and applying trust in web-based social networks". University of Maryland at College Park, College Park, MD, USA, 2005.

**[4] R. Guha, R. Kumar, P. Raghavan, A. Tomkins.** "Propagation of trust and distrust". *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 403-412.

**[5] B. Jennings, A. Finkelstein.** "Digital Identity and Reputation in the Context of a Bounded Social Ecosystem". *Business Process Management Workshops, vol. 17*, D. Ardagna, M. Mecella y J. Yang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 687-697.

**[6] Jordi Sabater i Mir.** "*Trust and reputation for agent societies*". Universitat Autónoma de Barcelona.

**[7] A. Jøsang.** "Trust Management in Online Communities". *New forms of collaborative production and innovation on the Internet - interdisciplinary perspectives*, no. May, V. Wittke and H. Hanekop, Eds. SOFI Goettingen, University Press Goettingen, 2011, pp. 5-7.

**[8] A. Jøsang, R. Ismail, C. Boyd.** "A Survey of Trust and Reputation Systems for Online Service Provision". *Decis. Support Syst., vol. 43, no. 2*, pp. 618-644, 2007.

**[9] Y. Kakizawa.** "In-house use of web 2.0: Enterprise 2.0". *NEC Tech. J., vol. 2, no. 2*, pp. 46-49, 2007.

**[10] S. D. Kamvar, M. T. Schlosser, H. Garcia-Molina.** "The EigenTrust Algorithm for Reputation Management in P2P Networks". *Proceedings of the Twelfth International World Wide Web Conference*, 2003, pp. 640-651.

**[11] R. Kerr, R. Cohen.** "Smart Cheaters Do Prosper: Defeating Trust and Reputation Systems". *Proceedings of the 8th International Joint Conference on Autonomous Agents and Multiagent Systems*, 2009.

**[12] J. M. Kleinberg.** "Authoritative Sources in a Hyperlinked Environment". *J. ACM, vol. 46*, pp. 668-677, 1999.

**[13] J. Kunegis, A. Lommatzsch, C. Bauckhage.** "The Slashdot Zoo: Mining a Social Network with Negative Edges". *18th International World Wide Web Conference, 2009*, p. 741.

**[14] F. Javier Ortega.** "Detection of dishonest behaviors in on-line networks using graph-based ranking techniques". *AI Communications, vol. 26, no. 3*, pp. 327-329, 2013.

**[15] F. Javier Ortega, J. A. Troyano, F. L. Cruz, C. G. Vallejo, F. Enríquez.** "Propagation of trust and distrust for the detection of trolls in a social network". *Computer Networks. Vol. 56, no. 12*, pp. 2884-2895, agosto 2012.

**[16] L. Page, S. Brin, R. Motwani, T. Winograd.** "*The PageRank Citation Ranking: Bringing Order to the Web*", 1999.

**[17] C. A. Yeung, M. G. Noll, N. Gibbins, C. Meinel, N. Shadbolt.** "SPEAR: SPamming-resistant Expertise Analysis and Ranking in Collaborative Tagging Systems". *Comput. Intell., vol. 27, no. 3*, pp. 458.

**[18] J. Zhang, J. Tang, J. Li.** "Expert Finding in a Social Network". *Advances in Databases Concepts, Systems and Applications*, pp. 1066-1069, 2007.

**[19] R. Zhou, K. Hwang.** "PowerTrust: A Robust and Scalable Reputation System for Trusted Peer-to-Peer Computing". *IEEE Trans. Parallel Distrib. Syst., vol. 18, no. 4*, pp. 460-473, abril 2007.