# Estudio para detectar usuarios influyentes de Twitter

Pablo Pellecchia[1], Leonardo Ibáñez[1], Maximiliano Colman[1], Martín Agüero[1]

[1] Universidad de Palermo, Facultad de Ingeniería, Grupo Turing
maguer@palermo.edu

**Resumen.** Por su masividad y trascendencia, las redes sociales han captado un gran interés tanto de usuarios como de analistas. Son un campo de estudio novedoso donde, internautas de todo el mundo, interactúan entre sí permitiendo un contacto directo con personalidades destacadas de la sociedad. Numerosos trabajos de investigación analizan el comportamiento de los usuarios, el contenido de los mensajes y el modo que se propaga una idea u opinión a través de estas redes. En el siguiente trabajo se presenta una nueva herramienta denominada TwitterDigger, desarrollada con el propósito de medir la influencia de los usuarios de la red social Twitter. El software analiza las características de los mensajes emitidos por un usuario y los compara con los emitidos por su red de seguidores. A partir de ciertos atributos clave y mediante el resultado de aplicar una fórmula denominada Índice de Propagación, el sistema determina el nivel de propagación de un usuario. Finalmente se presenta un caso de estudio realizado a partir de un conjunto de datos reales y se obtienen las conclusiones.

**Palabras Clave:** Redes Sociales, Twitter, Procesamiento de Texto, Grafos, Ingeniería de Software.

## 1 Introducción

La difusión boca-en-boca ha sido considerada como un importante mecanismo por el cual la información alcanza la masividad y ejerce una considerable influencia sobre la opinión pública [1]. Estudios recientes se focalizan en maximizar la difusión de un producto nuevo o información a través de individuos particulares, llamados "influyentes" quienes poseen una combinación de atributos personales tales como credibilidad, conocimiento o entusiasmo que les permiten influenciar con cierta facilidad a sus seguidores [2]. Este tipo de difusión es llevada a cabo por un número desproporcionadamente bajo de usuarios clave, seleccionados a partir de estudios empíricos que adolecen de dos dificultades principales: la difusión boca-en-boca es inobservable [3] y los datos sobre difusión son, en gran medida, sesgados hacia los eventos de difusión "exitosos" [4]. Por estas razones, el servicio de micro-blogging Twitter, presenta un ambiente ideal para el estudio del proceso de difusión [5], no sólo porque está integrada por una red de usuarios interconectados, sino porque además, los obliga a todos a comunicarse de la misma manera: a través de tweets a sus seguidores.

Si bien todo usuario que difunda una idea a través de sus seguidores puede considerarse que está influyendo sobre otro; son los expertos, periodistas, actores, músicos o políticos son los individuos capaces de lograr formación de opinión a gran

escala. Por otro lado, también existen muchos casos donde usuarios con una misma cantidad de seguidores no ejercen el mismo nivel de influencia.

En la actualidad, conocer el impacto de un individuo en Twitter, puede ser considerado información muy relevante para la realización de estudios sociales, campañas publicitarias y políticas. Contar con una herramienta que permita obtener una medida cuantitativa confiable de los usuarios clave, facilitaría el descubrimiento de aquellos usuarios de Twitter considerados recursos clave para la propagación de ideas u opiniones.

El presente artículo se organiza de la siguiente manera. En la sección 2 se explican otros trabajos del área. La sección 3 describe brevemente la arquitectura del prototipo desarrollado para el estudio. En la sección 4 se explica la fórmula propuesta para medir la influencia de los usuarios. La sección 5 presentan las características de los datos seleccionados para el análisis. La penúltima sección describe los resultados obtenidos y finalmente en la sección 7 se establecen conclusiones y trabajo futuro.

## 2 Antecedentes

En el último tiempo han surgido numerosas investigaciones sobre la manera en que los usuarios se influencian en Twitter. Se puede citar el trabajo de Wu [6] donde a partir de un hecho de público se estudia en tiempo real la forma en que la información se propaga y como los usuarios se influencian entre sí. En Itakura [7] se analizó la estructura de tres tipos de grafos diferentes (retweets, menciones y respuestas) llegando a la conclusión que el grafo de menciones es sobre el que mejor se propaga la información. Dai y Den [8] realizaron estudios sobre SinaWeibo, versión china de Twitter, y luego establecieron un modelo probabilístico basado en la teoría de Bayes para medir la posibilidad de que un tweet sea retwiteado. Galuba et al [8] también utiliza probabilidades para pronosticar cuándo una URL va a ser compartida por otros usuarios. Kempe et al. [10] observa a una red social como un boca en boca y busca encontrar los usuarios más influyentes a partir de los cuales maximizar la propagación de una campaña publicitaria sobre una red social. Benvenuto et al. [12] agrega al estudio de la influencia de los usuarios el factor de la pasividad de los receptores en el momento de decidir o no retwitear un tweet. En ese estudio se presenta un algoritmo que toma en cuenta ambos factores concluyendo que la influencia de un usuario no está directamente ligada a la popularidad del mismo. Asimismo utiliza el número de menciones de un usuario, la cantidad de retweets y la cantidad de seguidores para establecer un nivel de influencia de un usuario. El resultado obtenido es que no necesariamente tener una gran cantidad de seguidores implica tener un alto nivel de influencia. Weng et al. [13] propone una medida denominada TwitterRank que mide la influencia de un usuario tomando como referencia la similitud entre los tópicos que publica un usuario respecto de los de sus seguidores y la manera en que los usuarios están interconectados entre sí. Otros trabajos agregan el análisis semántico de los tweets que publica un usuario para el estudio de la influencia. Chowdury et al. [14] utiliza nombres de productos y marcas y luego analiza la estructura y sentimiento de los tweets que nombran alguno de estos factores. En Go, Bhayani y Huang [15] seemplean algoritmos de aprendizaje automático para extraer el sentimiento de los

tweets y clasificarlos en positivos o negativos. Este artículo dio lugar a Sentiment 140 [16] que es una herramienta web para realizar el análisis de tweets en tiempo real. Pak y Paroubek [17] seleccionan un set de datos de tweets y lo utilizan para entrenar a un clasificador de sentimiento y opiniones que luego emplean sobre un conjunto de datos reales, las técnicas que se proponen pueden aplicarse a diferentes idiomas.

Actualmente también existen herramientas online que realizan análisis en tiempo real y presentan los resultados de forma visual. Podemos citar las siguientes herramientas: Twithacolic [18] y TrendsMap [19] que rankean a los usuarios y temas por zonas geográficas. Twitalyzer [20] que estudia el impacto de los tweets para medir la influencia que un usuario tiene sobre los seguidores. Kred [21] muestra los contenidos que más influyen sobre la comunidad a la que pertenece el usuario utilizando dos índices, uno que muestra su influencia sobre sus seguidores y el segundo su actividad en Twitter. RetweetRank [22], Hashtagify [23] y SocialMention [24] estudian la influencia basándose únicamente en los hashtags, retweets y menciones de un usuario. Se puede decir que son pocas las herramientas que combinan todos los indicadores de medición de usuarios de una red social (retweets, hashtag, favoritos, respuestas, menciones). Es por eso que se puede afirmar que existe una oportunidad para proponer un sistema que, además de medir la influencia con los indicadores propios de la red social, también evalúe propagación a través de un análisis de similitud por coeficiente de distancia. A continuación se describirá la arquitectura de la herramienta de análisis de usuarios de Twitter propuesta en este trabajo, que ha sido denominada TwitterDigger [25]. Asimismo también se explicarán y fundamentarán una serie de mejoras introducidas en esta última evolución del prototipo.

## 3 Arquitectura del prototipo

La arquitectura general de TwitterDigger se divide en 3 módulos (ver Fig. 1):

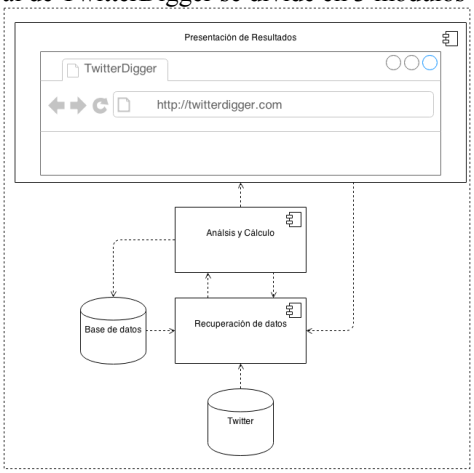


**Fig. 1** - Arquitectura general de TwitterDigger

Recuperación de Datos, Análisis y Cálculo, y como interfaz con el usuario el Módulo de Presentación de Resultados.

### 3.1 Módulo de Recuperación de Datos

Este módulo recupera de la base de datos de Twitter los datos de los usuarios y los mensajes emitidos en un lapso de tiempo. La salida de este módulo es un grafo de seguidores (SocialGraph) que es construido de manera dinámica, recuperando los datos de un usuario en función de lo requerido por el módulo de Análisis y Cálculo como se verá en la sección 3.2.

El SocialGraph representa las relaciones de seguidor-seguido. Para su armado se define un usuario (nodo raíz), que marca el punto desde donde comenzar la extracción de datos, y una cantidad de niveles de seguidores. Cada nodo del grafo está compuesto por los usuarios del seguidor así como los tweets de éstos.

TwitterDigger funciona en dos modalidades, online y offline, en la primera los datos se obtienen en tiempo real desde Twitter utilizando la API Twitter4J [26]. Twitter limita la cantidad de solicitudes que se pueden realizar a sus servidores por lo cual, esta herramienta conmuta entre conexiones autorizadas para evitar los límites establecidos. Cada vez que se recupera un usuario, se obtienen también sus seguidores y sus tweets. Una de las innovaciones introducidas en esta última versión es que los seguidores se recuperan de manera aleatoria y no según el histórico. De esta manera, no se sigue ningún criterio establecido por la API de Twitter4J y es posible encontrar sets de datos más heterogéneos para el análisis. Finalmente, los datos son almacenados en una base de datos relacional. En la recuperación offline, los datos son recuperados desde la base de datos donde fueron guardados previamente.

### 3.2 Módulo de Análisis y Cálculo

El módulo de Análisis y Cálculo utiliza los servicios del módulo de Recuperación de Datos de manera inteligente para recuperar los datos desde Twitter solamente cuando sea necesario. Este módulo utiliza el patrón de carga diferida (Lazy Loading) [27] para recuperar únicamente los seguidores de un usuario en los casos en que exista propagación, postergando la recuperación de datos sólo cuando es necesario. Cada vez que se realiza un llamado al módulo de Recuperación de Datos, se descubren nuevos nodos del SocialGraph ya que se obtienen los seguidores de un usuario y los mensajes de estos. Luego se procede a la búsqueda de coincidencias entre los tweets de los seguidores y los de un usuario. Donde existe coincidencia entre tweets, se determina que hubo propagación entre niveles. Posteriormente, si no se llegó a la cantidad de niveles máxima, se procede a recuperar los seguidores del siguiente usuario solamente en los casos en los que se detectaron coincidencias.

Para el análisis cruzado entre un nodo y los seguidores, se utiliza el algoritmo de búsqueda de amplitud [28] explorando horizontalmente los nodos seguidores. En el caso de existir ciclos, el algoritmo los elimina para no analizar un nodo más de una vez. Solamente se evalúan los tweets de los seguidores que se publicaron después del mensaje del nodo raíz. Las coincidencias que se buscan son: retweets, hashtags,

respuestas, favoritos y métrica. Finalmente, con todos los análisis ejecutados, se realiza el cálculo del Índice de Propagación (IP) que se describirá en la sección 4.

**Comparador por Métrica:** Dentro del módulo de Análisis y Cálculo, uno de los analizadores es el comparador por métrica, donde un coeficiente de similitud de texto define el grado de igualdad entre dos tweets.
En la Fig. 2 se muestra un estudio comparativo entre diferentes algoritmos de similitud de texto realizado para este proyecto. La prueba (O,D1 ordenado) compara mensajes originales con derivados, donde el resultado debería ser cercado a 1. En (O,D1 - random) se comparan mensajes diferentes, donde el valor esperado es 0. A partir de los resultados se ha decidido utilizar Sørensen-Dice [29] para medir similitud entre textos. En este algoritmo, se comparan bigramas de caracteres pertenecientes a cada palabra de cada tweet, buscando las coincidencias que existen entre ambos mensajes. El resultado de este proceso devuelve una medida de que tan parecidos son ambos tweets. En función del estudio realizado se estableció un umbral mínimo para la similitud de los mensajes: 0.65, esto significa que si la medida es menor a este valor, no se consideran como textos similares.

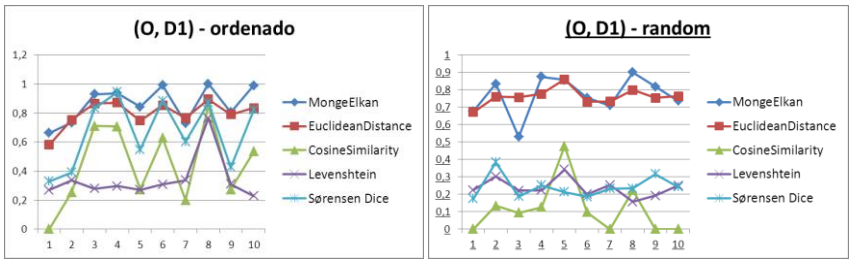


**Fig. 2 -** Análisis de algoritmos de similitud

Luego de realizado el análisis, se procede al cálculo del Índice de Propagación que se detallará en la sección 4. Dicho índice es una medida de la capacidad de un usuario de propagar una idea hacia sus seguidores, estableciendo influencia sobre los futuros tweets que publiquen.

### 3.3 Módulo de representación de datos

El grafo SocialGraph, se muestra al usuario con una representación gráfica desde el navegador web (ver Fig. 3). En color gris se grafica al usuario raíz con su correspondiente Índice de Propagación, en rojo los usuarios donde no se encontró propagación y en verde aquellos en los que si hubo. Al ubicar el puntero sobre alguno de los nodos, aparece un menú contextual con información adicional sobre ese nodo. La conexión entre la GUI y el back-end se realiza con tecnología Servlet y JSP. También se incorporaron recursos de los proyectos Cytoscape [30], JQuery [31] y JSTL [32] para renderizar el grafo.

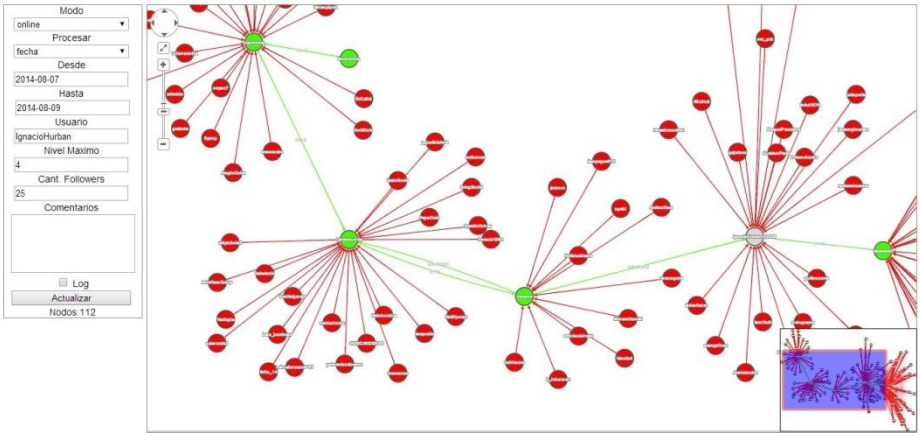


**Fig. 3 -** Interfaz gráfica de usuario de TwitterDigger. Análisis: 4 niveles a 25 usuarios por nivel

Finalmente y a modo de resumen de sección, en la Fig. 4 se presenta un diagrama de secuencia donde se describe la interrelación entre módulos durante una solicitud de análisis.

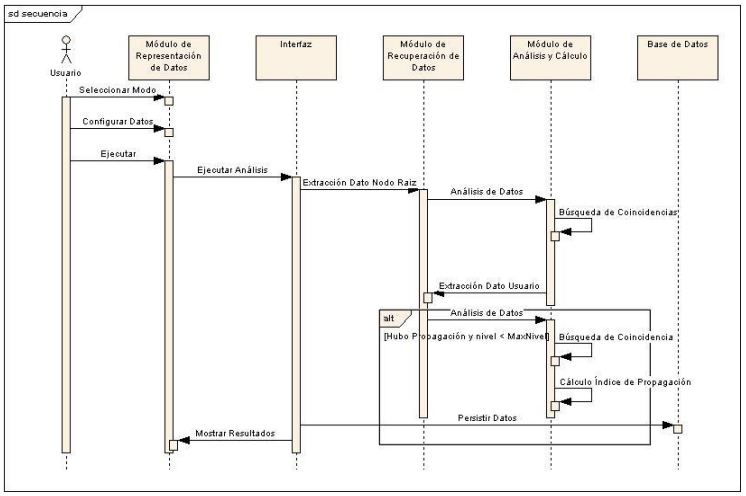


**Fig. 4** - Diagrama de secuencia principal

## 4 Índice de Propagación

Se define al Índice de Propagación como una medida capaz de cuantificar la capacidad de un usuario de propagar una idea u opinión hacia sus seguidores. De esta

manera, entre más alto sea el índice, mayores son las chances de que los seguidores de un usuario compartan sus ideas.

Para calcular este índice, se toma la información obtenida en la etapa de análisis y comparación de tweets realizada por el Módulo de Análisis y Cálculo, y en función del tipo de coincidencias se asigna un peso o factor para cada ocurrencia (ver Tabla 1). El factor marca un grado de importancia a cada tipo de coincidencia por el tipo de relación que establece entre seguidor – seguido. Si un usuario hace retweet a otro, se está involucrando con el contenido del mensaje y al mismo tiempo señala la fuente desde donde comparte la información. Por el otro lado, si se comparte un hashtag se comparten las palabras clave del tema pero en la mayoría de los casos no se marca la fuente. En el caso de las respuestas se ha decidido que estas tengan un peso menor que el del hashtag ya que si bien se marca explícitamente al usuario al que se contesta, una respuesta no implica que haya influenciado a un seguidor. También se establece un peso para las demás coincidencias: métrica y favorito.

**Tabla 1.** Factor por tipo de coincidencia.

| Tipo de coincidencia | Factor |
|---|---|
| Retweet | 0.9 |
| Métrica | 0.6 |
| Hashtag | 0.5 |
| Respuesta | 0.4 |
| Favorito | 0.1 |

A continuación la ecuación (1) que conforma al Índice de Propagación. Se define a $s$ como la cantidad de nodos en el SocialGraph, a $m$ la cantidad de tweets que emite un usuario:

$$IP = \left(\frac{RTprop + MEprop + HTprop + RPprop + FVprop}{m}\right) * 100 \ (\mathbf{1})$$

Donde *RTprop, MEprop, HTprop, RPprop, FVprop* es la proporción de veces donde usuario logró propagar una idea hacia sus seguidores y *cantRT, cantME, cantHT, cantRP, cantFV* muestran cuantas veces se encontró un tipo de coincidencia:

$$XXprop = \frac{cantXX * fXX}{s} (\mathbf{2})$$

## 5 Características de los datos

Para la ejecución del siguiente caso, se han seleccionado usuarios de Twitter que pueden ser considerados formadores de opinión y que comparten las siguientes características: cantidad de seguidores, frecuencia en las publicaciones y vigencia en los medios de comunicación masivos (ver Tabla 2):

**Tabla 2.** Datos de usuarios de Twitter seleccionados.

| Nombre usuario | Perfil | Cant. Seguidores |
|---|---|---|
| @andykusnetzoff | Periodista | 1440032 |
| @rpettinato | Presentador | 1370273 |
| @fantinofantino | Presentador | 1282937 |
| @JPVarsky | Periodista | 968016 |
| @vh590 | Periodista | 719027 |

## 6 Resultados

En la Fig. 5 se observan los resultados de las pruebas utilizando TwitterDigger. Se analizó 1 nivel y de ese nivel se recuperaron 300 seguidores en forma aleatoria (no en orden cronológico de alta como seguidor). Los resultados muestran que los usuarios @vh590 y @andykusnetzoff han obtenido un mayor IP respecto de los demás. El usuario @JPVarsky tiene IP=0 porque no se han encontrado coincidencias con sus seguidores. A partir de las pruebas realizadas se ha notado que las coincidencias que se encuentran entre un usuario y sus seguidores varían por el momento en que se realiza el estudio. Si un usuario publica un mensaje con un contenido destacado, ya sea por ser considerado controversial o muy positivo, aumentan las probabilidades de que su IP sea mayor. Por tal motivo y para evitar el sesgo, el análisis se realiza sobre los últimos 200 tweets.

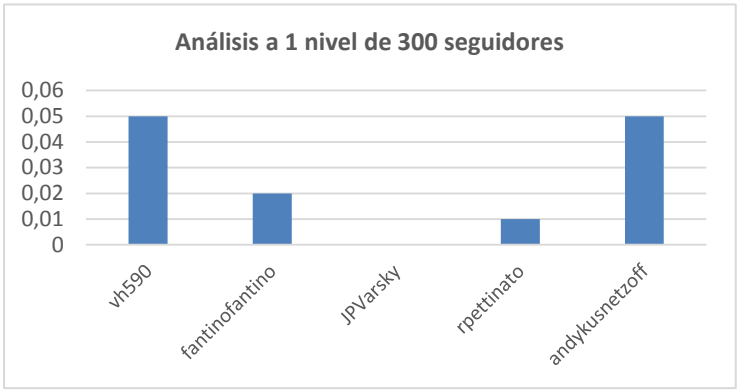


**Fig. 5 -** Resultado de estudio comparativo de usuarios

En la Fig. 6 se detalla el tipo de propagación de los usuarios @vh590 y @andykusnetzoff. Se puede apreciar que las mayores coincidencias han sido por hashtag en ambos casos. También se observa que, si bien en el segundo gráfico la cantidad de hashtags es menor que la del primero, la cantidad de coincidencias por métrica es mayor, por lo tanto y dado el factor asignado a ese tipo de coincidencia, se incrementa el IP de ese usuario.

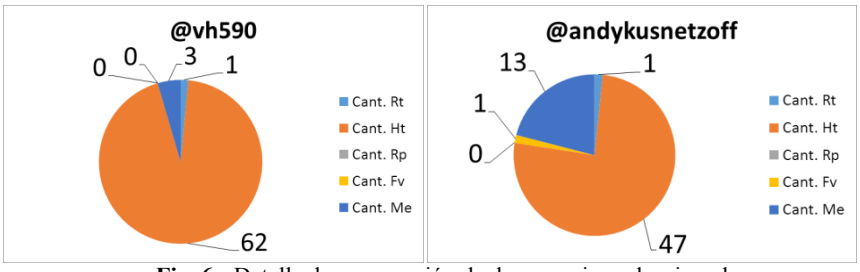


**Fig. 6 -** Detalle de propagación de dos usuarios seleccionados

## 7 Conclusiones y Trabajo Futuro

El análisis de la influencia de usuarios de redes sociales y la manera en que se propaga la información, se presenta como un área de estudio en notable crecimiento. La información obtenida a partir de la interacción entre sus usuarios ya es considerada una herramienta de decisión para empresas, campañas publicitarios, políticas o en las ciencias sociales.

En este trabajo se ha presentado una breve descripción de las principales investigaciones realizadas hasta el momento así como también un conjunto de herramientas online. En función del estudio del arte realizado, se propone a la herramienta TwitterDigger como una alternativa capaz de ampliar el modo de medir propagación en Twitter a través del empleo de un coeficiente de similitud entre textos. Asimismo se propone cuantificar la capacidad de influir a sus seguidores a través de un Índice de Propagación.

En la siguiente fase se proyecta optimizar el procedimiento utilizado para recuperar datos de Twitter así como también incrementar la velocidad de procesamiento de los mensajes. Está previsto incorporar una vista "línea de tiempo" para graficar la evolución del IP de un usuario.

## 8 Referencias

1. Katz, E., Lazarsfeld, P. F.: Personal influence; the part played by people in the flow of mass communications. Free Press, Glencoe, Ill." (1955)
2. Keller, E., Berry, J.: The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy. Free Press, New York, NY (2003)
3. Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using internet chain-letter data. En Proceedings of the National Academy of Sciences, 105(12):4633. (2008)
4. Denrell, J.: Vicarious learning, undersampling of failure, and the myths of management. Organization Science, 14(3):227–243 (2003)
5. Bakshy, E., Hofman, J.: Everyone's an Influencer: Quantifying Influence on Twitter. En Proceedings of the fourth ACM international conference on Web search and data mining. (2011)
6. Wu, F., Ye S.: Measuring message propagation and social influence on Twitter.com. En Lecture Notes in Computer Science, vol. 6430, Springer (2010)

7. Itakura, K., Noboru, S.: Using Twitter's Mentions for Efficient Emergency Message Propagation En Proceedings of the 2013 International Conference on Availability, Reliability and Security (ARES '13). (2013)
8. Dai, Y., Deng, D.: How Your Friends Influence You: Quantifying Pairwise Influences on Twitter. En Proceedings of the 2012 International Conference on Cloud and Service Computing, CSC '12. (2012)
9. Aberer, K., Chakraborty, D., Galuba, W.,Despotovic, Z., Kellerer. W.: Outtweeting theTwitterers - Predicting Information Cascades. En Microblogs 3rd Workshop on Online Social Networks,WOSN. (2010)
10. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the Spread of Influence through a Social Network. En Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03. (2003)
11. Asur, S., Galuba, W., Huberman, B., Romero, D.: Influence and passivity in social media. En Proceedings of the 20th international conference companion on World Wide Web, WWW '11. (2011)
12. Benvenuto, F., Cha, M., Gummadi, K., Haddadi, H.,: Measuring User Influence in Twitter: The Million Follower Fallacy. En Fourth International AAAI Conference on Weblogs and Social Media. (2010)
13. Jiang, J., Lim, E., Weng,J.: Twitterrank: Finding Topic-Sensitive Influential Twitterers. En ACM International Conference on Web Search and Data Mining. (2010)
14. Chowdury, A., Jansen, B., Sobel, K., Zhang, M.: Twitter power: Tweets as electronic word of mouth. En Journal of the American Society for Information Science and Technology. (2009)
15. Bhayani, R., Go Huang, L.: Twitter Sentiment Classification using Distant Supervision. Stanford University. (2009)
16. Sentiment140.Disponible en: http://www.sentiment140.com/. Accedido el: 25/07/2014
17. Pak, A., Paroubek, P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining. En Proceedings of the Seventh International Conference on Language Resources and Evaluation. (2010)
18. Twitaholic. Disponible en: http://twitaholic.com/. Accedido el: 25/07/2014
19. TrendsMap. Disponible en: http://trendsmap.com/. Accedido el: 25/07/2014
20. Twitalyzer. Disponible en: http://www.twitalyzer.com/5/index.asp. Accedido el: 25/07/2014
21. Kred. Disponible en: http://kred.com/. Accedido el: 25/07/2014
22. RetweetRank. Disponible en: http://www.retweetrank.com/. Accedido el: 25/07/2014
23. Hashtagify. Disponible en: http://hashtagify.me/ Accedido el: 25/07/2014
24. SocialMention. Disponible en: http://www.socialmention.com/. Accedido el: 25/07/2014
25. Ibáñez L., Pellecchia, P., Agüero, M.: Software para Medir la Capacidad de Propagación de Usuarios de Twitter. 43 JAIIO. (2014)
26. Twitter4j.Disponible en: http://twitter4j.org/en/index.html. Accedido el: 25/07/2014
27. Fowler, M.: Patterns of enterprise application architecture (1st edition). Addison-Wesley Professional. (2002)
28. Cormen, T. H., Leiserson, C. E., Riverst, R. L., Stein, C.: Introduction to algorithms (2nd edition). MIT Press and McGraw-Hill. (2001)
29. Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. En BiologiskeSkrifter, 5 (4): 1 - 34. I kommission hos E. Munksgaard. (1948)
30. Cytoscape.js. Disponible en: http://cytoscape.github.io/cytoscape.js/. Accedido el: 25/07/2014
31. JQuery.Disponible en: http://jquery.com/. Accedido el: 25/07/2014
32. Jstl. Disponible en: https://jstl.java.net/. Accedido el: 25/07/2014