Revista Tecnocientífica URU
Universidad Rafael Urdaneta
Facultad de Ingeniería
No. 11 Julio – Diciembre 2016
Depósito Legal: PPI 201402ZU4464
ISSN: 2343-6360

Categorización de usuarios de Twitter

Francisco Luengo, Carla Morillo y Yaskelly Yedra
Departamentos de Matemática y Computación. Universidad del Zulia. Facultad Experimental de Ciencias.
Maracaibo, Estado Zulia, Venezuela.
fluengo@fec.luz.edu.ve, carlamorillo2@hotmail.com, yyedra@fec.luz.edu.ve

Recibido: 08-12-2015. Aceptado: 29-11-2016.

Resumen

Según estudios, se estima que el conjunto de todos los datos digitales creados, replicados y consumidos por año, pasará de 130 exabytes en 2005 a 40 zettabytes en 2020. Tan solo en la red social Twitter se publicó un promedio de 433 mil tuits por minuto durante el año 2014. Estas inmensas cantidades de datos han dado origen al surgimiento de técnicas y herramientas de análisis que han cobrado gran importancia en el ámbito de los negocios, aunque también encuentran aplicación en estudios tan diversos como los antropológicos, de seguridad nacional, seguimiento de enfermedades, entre otros. Este trabajo tiene como objetivo la implementación de técnicas para establecer categorías de usuarios de Twitter. Tales categorías quedan definidas por el comportamiento de dichos usuarios dentro de esta red social, lo cual provee un conocimiento valioso para estudios como los comentados. Adicionalmente, se presenta una breve descripción del estado del arte de las aplicaciones para análisis de tuits y sus usos, lo cual sirve de soporte a la presente investigación. El trabajo abarca conexión, captura, organización, y análisisde tuitsa través de técnicas de clasificación y agrupamiento.Los resultados muestran lo efectivo de las técnicas empleadas y su potencial para resultados mas ambiciosos.

Palabras claves: Twitter, clasificación, k-means

Categorizing Twitter users

Abstract

According to some researches, it is estimated that the set of all digital data created, replicated and consumed per year, will grow from 130 exabytes in 2005 to 40 zettabytes in 2020. Only in the social network Twitter were published an average of 433,000 tweets per minute during 2014. These vast amounts of data have led to the emergence of techniques and analysis tools which have been very important in the field of business, as well as in fields so diverse like anthropology, national security, monitoring of diseases, among others. This work aims to implement techniques to establish categories of Twitter users. Those categories are defined by the behavior of the users into the social network, which provides valuable information for researches like mentioned above. Additionally, a brief description of the state of the art about applications for analysis of tweets and their uses is presented, which serves to support this research. The work covers connection, capture, organization, and analysis of tweets through classification and clustering techniques. The results show how effective the techniques used are and the potential for more ambitious results.

Key words: Twitter, clasification, k-means

Introducción

El Big Data no se refiere sólo a grandes volúmenes de datos, también es el término que se emplea hoy en día para describir el conjunto de procesos, tecnologías y metodologías para tratamiento y análisis de enormes repositorios de datos, tan desproporcionadamente grandes que resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales; y así capturar el valor que los propios datos encierran. El desafío entonces esta en comprender y extraer información útil de estas fuentes de datos complejas y no estructuradas. (Eaton et al [6]) (Rui Han et al, 2015[7]).

Distintas variables son las que influyen para que la cantidad de información producida día a día a nivel mundial se multiplique constantemente. La penetración de Internet y de los dispositivos conectados es un claro ejemplo, al igual que el desarrollo de las tecnologías inalámbricas, los productos inteligentes y los negocios definidos por software. Estadísticas como las proporcionadas por Internet WorldStats (Internet World Stats [4]) y Global Web Index (Global Web Index [5]) nos dejan claro el gran poder de penetración que tiene internet dentro de las sociedades, y de igual modo las redes sociales, formando parte de nuestro estilo de vida y ofreciendo una fuente inagotable de información en cuyo contenido reposa latente un conocimiento que espera ser descubierto.

Este trabajo se enfoca en la red social Twitter (HaewoonKwak et al [8]), y en las técnicas para capturar, organizar y analizar la información que fluye por ella, con el propósito de poder clasificar grupos de sus usuarios según el comportamiento que estos exhiban dentro de la misma.

Fundamentos Teóricos

Siempre que hagamos una búsqueda en internet, enviemos un email, usemos un teléfono móvil, actualicemos una red social, usemos una tarjeta de crédito, activemos el GPS, juguemos en-línea o hagamos la compra en el supermercado dejamos detrás de nosotros una montaña de datos, huellas digitales y registros que ofrecen una información valiosa y cuyo estudio es del interés de muchos.

En un trabajo realizado por Martin Hilberty Priscila López [1], los investigadores estimaron que en el período comprendido entre 1986 al 2007, la humanidad había generado 2×1021bytes en comunicaciones, y almacenado 2.9×1020 bytes. De acuerdo con Hilberty y López [1], el 75% de la información almacenada en el mundo aún estaba en formato analógico en el año 2000, en su mayoría en forma de cintas de vídeo, y para el año 2007 el 94% de información del mundo ya era digital.

Por otro lado, en su 6to estudio sobre el Universo Digital presentado por IDC y patrocinado por EMC Corporation, llamado "THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and BiggestGrowth in theFar East" (EMC Corporation [2]), IDC nos dice que la mayoría de la información en el universo digital, 68% en 2012, fue creada y consumida por los propios usuarios (al ver televisión digital, interactuando en redes sociales, enviando fotografías y video entre dispositivos o internet, etc) y de la cual, solamente una pequeña fracción ha sido explorada con propósitos de análisis.

Para darnos una mejor idea nos referimos a los resultados obtenidos por la aplicación Qmee [3], la cual reporta que durante cada minuto del año 2014 se realizaron en promedio 293 mil actualizaciones de estatus en Facebook, se subieron 67 mil fotos a Instagram, visto más de 5 millones de videos en YouTube, y publicado 433 mil tuits.

De lo anterior puede decirse que las redes sociales, como plataforma para las publicaciones digitales, representan por si mismas un mini universo digital con los mismos retos del Big Data (capturar, descubrir, y analizar), y han sido los investigadores en el campo de las Tecnologías de la Información (TI) uno de los primeros en reconocer su riqueza; de allí surge el término Inteligencia de Negocios (Business Intelligence), refiriéndose a la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios y crear ventajas competitivas. (Stefano Rizzi [9]) (MihaelaMuntean [10])

Hoy en día, esas ventajas están enfocadas hacia el SEO (de sus siglas en inglés, Search Engine

Optimization, referidos al conjunto de técnicas para el posicionamiento dentro de buscadores), o mas ampliamente, hacia el SEM (del inglés, Search Engine Marketing, refiriéndose a las técnicas para promover las páginas web). Dando origen al surgimiento de lo que se denomina SMA (Social Media Analytics), es decir, el conjunto de herramientas destinadas a conocer las opiniones, sentimientos, tendencias y preferencias de los usuarios de las redes sociales. (Ribarsky et al[11]) (ChristophBoden [12]) (Kent, P [14]).

Si bien estas herramientas de análisis son mas conocidas por su importancia en el ámbito de los negocios, sus técnicas también son aplicadas a un amplio espectro de problemas que incluyen estudios antropológicos, asuntos de seguridad nacional, análisis de propagación de enfermedades, entre muchos otros (Dror Y et al [13]) (Moore, M.T [15]) Haibin Liu [16]) (Liu and Dongwon [17]).

En este sentido, son muchas las aplicaciones que han surgido, específicamente orientadas a localizar o definir perfiles de individuos o entidades, dentro de las redes sociales, así como su impacto, influencia y reputación online, es decir, su marca digital. Tres ejemplos de algunas de ellas son:

True Social Metrics [18], permite medir comentarios por post, shares por post, contenido favorito, tasa de conversación, tasa de amplificación, tasa de aplausos y valor económico de las principales redes sociales como Twitter, Facebook, Google+, YouTube, LinkedIn, Instagram, etc.

Sentimentalytics [19], es una herramienta que funciona en tiempo real y multiidioma, muestra el sentimiento que desprende una publicación o tweets, además analiza menciones, mensajes temas y usuarios.

Semrush [20], es una de las herramientas más importante para el análisis de la competencia. Con ella se puede analizar el SEO y SEM de empresas de la competencia.

Todas estas herramientas se apoyan en técnicas de minería de datos; sin embargo, la tendencia actual es combinarlas con otras técnicas como las proporcionadas por la minería de textos, para mejorar la calidad de los resultados. Algunas herramientas que emplean estas técnicas son:

WordStat [21], se utiliza para realizar análisis de competencia en sitios web, análisis de sentimientos y contenido en preguntas abiertas. Permite realizar categorización, clasificación mediante Naïve-Bayes o el algoritmo de vecindad (con palabras o conceptos), análisis de correspondencia y relacionar textos no estructurados con datos estructurados (previamente conocidos) entre otras utilidades.

Attensity [22], análisis y minería de texto se utiliza para analizar la información y la inteligencia colectiva en redes sociales y foros.

Las herramientas de análisis de la diversidad léxica, minería de opiniones, análisis de los sentimientos o mapas de conexiones semánticas entre usuarios son solo algunos de los desarrollos que han proliferado dentro de las técnicas de búsqueda de información en las redes sociales. Son herramientas y técnicas imprescindibles en el época de Big Data, si queremos tener la más mínima posibilidad de ser capaces procesar/analizar los altos volúmenes de información actual.

Parte Experimental

Esta investigación abarcó cuatro fases: conexión, captura, organización y análisis. Para lograr las tres primeras se desarrolló una aplicación que permite la conexión con Twitter, capturar tuits de usuarios determinados y estructurarlos en una base de datos para facilitar su posterior análisis. Para la última fase se implementaron en la aplicación técnicas de clasificación, agrupamiento, y probabilísticas, que facilitaron la interpretación de los resultados.

Resultados y Discusión

La Aplicación

Para esta investigación se desarrolló una aplicación, llamadaTuitCLUS; con el propósito de cubrir las tres primeras fases del proyecto y facilitar la cuarta.

TuitCLUS fue desarrollado en el lenguaje de programación Java bajo el entorno de desarrollo de Eclipse Juno, empleando la librería Twitter4J[25] para el acceso, autenticación, autorización y uso de los datos proporcionados por la red social Twitter. Para la organización y almacenamiento de los tuits se utilizó el manejador de base de datos MySQL.

Si bien, Twitter4J no es una librería oficial para la API (AplicationProgramming Interface) de esta red red social, ella puede ser integrada a cualquier aplicación de Java junto con el servicio de Twitter. Esta librería está disponible para los desarrolladores con las siguientes características (Twitter4J [25]): librería funcional para la plataforma Java versión 5 o posterior, brinda soporte a las plataformas de Android y Google App Engine, no posee dependencia(es decir, no requiere de algún archivo, aplicación o herramienta adicional), posee soporte integrado con OAuth, es compatible en un 100% con la API de Twitter v1.1, e incluye software de JSON.org para analizar respuestas en formato JSON de la API de Twitter.

Todos los pasos llevados a cabo por TuitCLUS se resume en la siguiente figura.

Figura 1. Secuencia de procesamiento deTuitCLUS

A continuación se describen cada una de las fases implementadas.

Conexión

La plataforma de Twitter provee a los desarrolladores e investigadores una amplia documentación sobre las diversas APIs, herramientas, kits de desarrollo, soluciones y muchísima información referente a esta red social (Twitter Documentation [23]).

Todas las APIs de Twitter están basadas en el protocolo HTTP. Esto quiere decir que cualquier software que las use envía una serie de mensajes (solicitudes) estructurados a los servidores de Twitter. Sin embargo, para que dichas solicitudes puedan ser efectivamente procesadas por los servidores de Twitter, éstas deben proveer la siguiente información: qué aplicación hace la solicitud, nombre del usuario que hace la solicitud, si el usuario ha otorgado permiso a la aplicación para realizar la solicitud, y si la solicitud fue alterada por un tercero mientras estuvo en tránsito.Para permitir a las aplicaciones proporcionar esta información, Twitter se basa en el protocolo de autenticación OAuth [24].

Para el desarrollo deTuitCLUS fue necesario registrarlaaplicación en la dirección web apps.twitter.com/app/newy asociarla a una cuenta Twiter creada para tal fin. Bajo esta modalidad, Twitter provee a los desarrolladores las credenciales y claves de acceso para establecer conexiones a los recursos de sus APIs.Estas credenciales son secretas y sólo debe conocerlas el desarrollador de la aplicación, ya que

estas identifican las solicitudes que recibe la API de Twitter con la aplicación que las está haciendo.

Una vez obtenidos los códigos de autorización y autenticación (ConsumerKey, ConsumerSecret, AccessToken, y el AccessTokenSecret), se establecieron como una constante en la aplicación, haciendo uso de la clase "ConfigurationBuilder" definida por la librería Twitter4J, tal como se muestra a continuación:

```
private finalConfigurationBuildercb;
cb = new ConfigurationBuilder();
cb.setOAuthConsumerKey("***Aqui el ConsumerKey***");
cb.setOAuthConsumerSecret("***Aqui el ConsumerSecret***");
cb.setOAuthAccessToken("***Aqui el AccessToken***");
cb.setOAuthAccessTokenSecret("***Aqui el AccessTokenSecret***");
```

Con esta información de autenticación, ya se esta listo para ejecutar consultas a los servidores de Twitter.

Captura

El objetivo ahora es tener acceso a la información pública de usuarios de Twitter, para lo cual esta red social proporciona básicamente dos APIs: REST API y Streaming API.

La REST (RespresentationalState Transfer) API proporciona acceso programático aleer timeline, twitear y seguir a usuarios. LaStreaming API le da a los desarrolladores una baja latencia de acceso a la corriente mundial de los datos de los tuits de Twitter.

Una misma aplicación puede emplear ambas APIs de requerilas; sin embargo, para los propósitos de esta investigación solo se requirió los servicios de la REST API, los cuales fueron solicitados a través de los métodos proporcionados por Twitter4J.

Ahora bien, lo primero es agregar aquellas cuentas de las que se desea descargar información, y para ello, lo principal fue llevar a cabo la validación de dichas cuentas, conocer si existen o no, si son privadas o públicas. El siguiente código proporciona dicha información:

```
String cuentaTwitter;
User usuario=twitter.showUser(cuentaTwitter);
usuario.isProtected();
usuario.getStatus();
```

Una vez verificada la cuenta, se descarga su timeline; es decir, los tuits que haya publicado, fucionando las interfaces Twitter y Status de Twitter4J:

```
List<Status>statuses;
statuses = twitter.getUserTimeline(cuentaTwitter);
```

Organización

Ya teniendo el timeline en statuses, los métodos de la clase Status dan acceso a los distintos componentes de un tuit:ID del usuario asignado por Twitter, nombre de usuario, descripción del usuario, la imagen de usuario, cantidad de seguidores, cantidad de amigos o seguidos, los hashtags utilizados, idioma, localización, el texto del tuit, cantidad de retweets del tuit, cantidad de favoritos del tuit, si se

realizaron menciones en el tuit, si es un retweet o no, si es una cuenta verificada o no.

La Figura 2 muestra la estructura de la base de datos creada para relacionar toda la información de los tuits de los usuarios.

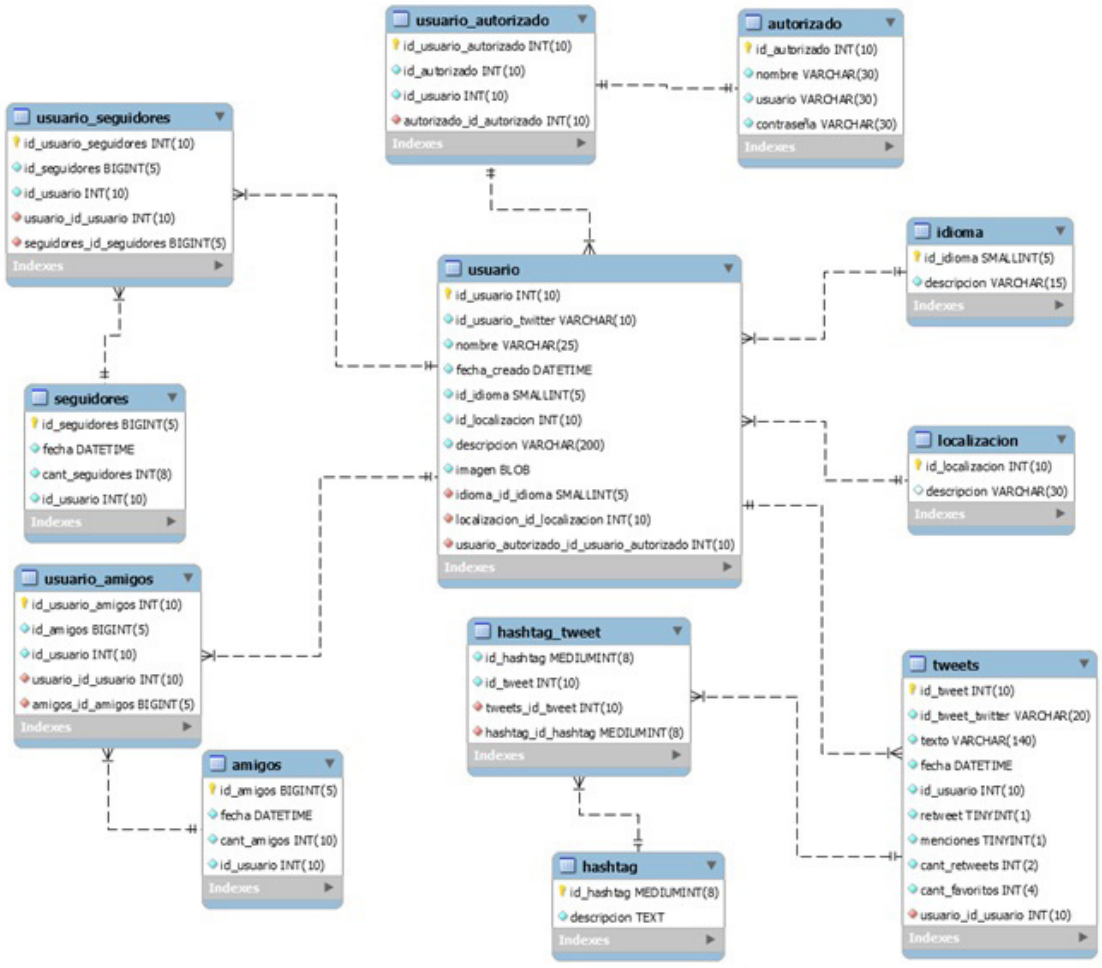

Figura 2. Base de Datos de TuitCLUS

Análisis

Finalmente, para establecer categorías de usuarios se consiguen similitudes entre la cuentas empleando técnicas de clasificación k-means (Berkhin [26]) y probabilísticas naive-bayes (Frank et al [27]), de manera que cada grupo formado represente una categoría.

Para aplicar el k-means se definió un vector característico asociado a cada cuenta. Los componentes de dicho vector corresponden a cuantificadores asociados al comportamiento de la cuenta, tales como: total de tuits, cantidad de retuits que hace y que le hacen, frecuencia y promedio de tuiteo, frecuencia y promedio con la que es retuiteado y con la que retuitea, cantidad de seguidos y seguidores, tasas de crecimiento de seguidores y seguidos, nivel de impacto de los tuits, tiempo de creación de la cuenta, si la cuenta es certificada o no, etc. De estos atributos se construye el vector característico para cada cuenta y se generan los centroides que definirán los grupos o categorías finales.

La siguiente sección describe en mas detalle los resultados mencionados.

Pruebas y Resultados

Para el momento de las pruebas se contó con alrededor de 1.061.457tuits correspondientes a 108 cuentas de usuarios, descargados en el período dejulio aoctubre de 2015.

Inicialmente se realizaron varias pruebas utilizando solo dos atributos asociados a las cuentas de usuarios, para definir sus vectores característicos. La Tabla 1 muestra los resultados de aplicar el k-meanspara dos características a la vez, sobre 50 cuentas. Para cada par de características se buscó establecer 5, 4, 3 y 2 grupos de usuarios, respectivamente.

Al usar solo dos características es más fácil notar la relación entre las cuentas que terminan asociadas acada grupo formado, además de poder visualizar los centroides finales en relación a la distribución espacial de los vectores característicos (ver Fig. 3). Esta información ayuda a definir comportamientos claves para categorizar usuarios, los cuales pueden traducirse en condiciones específicas que los definen.

Tabla 1. Cantidad total de cuentas asociadas a cada grupo según características.

	Prom. de Tuiteo y %Tuits Propios	Total Tuits y % Tuits Propios	%Retuits y % Tuits Propios	Tasa Seguidores y TasaSeguidos
5 Grupos	[22, 2, 13, 9, 4]	[29, 10, 5, 3, 3]	[6, 19, 5, 20, 0]	[41, 1, 1, 6, 1]
4 Grupos	[23, 2, 23, 2]	[9, 25, 6, 10]	[14, 14, 11, 11]	[5, 5, 38, 2]
3 Grupos	[3, 38, 9]	[29, 2, 19]	[13, 25, 12]	[10, 37, 3]
2 Grupos	[23, 27]	[27,23]	[12,38]	[45, 5]

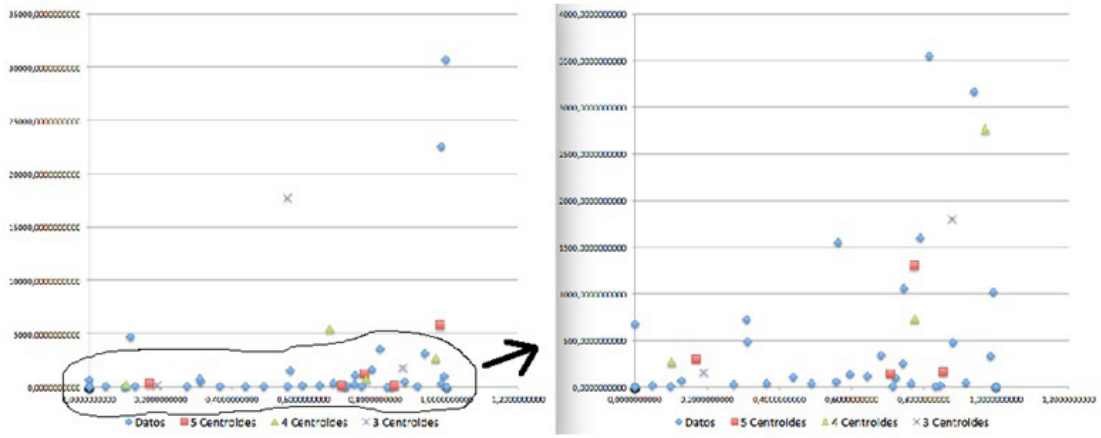

Figura 3. Distribución de los vectores característicos y de los centroides finales, considerando el porcentaje de tuits propios y el total de tuits hechos.

En otra prueba, se aplicó el k-means con el propósito de definir cinco (5) grupos entre 106 cuentas de usuario, para lo cual se consideraron los siguientes atributos para el vector característico: cantidad de retuits recibidos, frecuencia de tuiteo, tasa de seguidores, tasa de seguidos, y porcentaje de retuits que hace. La Tabla 2 muestra los resultados de aplicar el método desde dos grupos de centroides distintos; en el primero los centroides están distribuidos uniformemente en el espacio de cuentas, mientras que para el segundo grupo se empleó la base canónica (utilizando el máximo valor que alcanza cada atributo para formar su respectivo vector) para el espacio 5-dimensional.

Tabla 2. Distribución de cuentas resultante para dos tipos distintos de centroides iniciales.

Centroidesiniciales uniformemente distribuidos	[Grupo 1]--> cantidad de cuentas asociadas= 23 [Grupo 2]--> cantidad de cuentas asociadas= 45 [Grupo 3]--> cantidad de cuentas asociadas= 18 [Grupo 4]--> cantidad de cuentas asociadas= 10 [Grupo 5]--> cantidad de cuentas asociadas= 4
Centroidesiniciales formando base canónica	[Grupo 1]--> cantidad de cuentas asociadas= 14 [Grupo 2]--> cantidad de cuentas asociadas= 13 [Grupo 3]--> cantidad de cuentas asociadas= 14 [Grupo 4]--> cantidad de cuentas asociadas= 6 [Grupo 5]--> cantidad de cuentas asociadas= 53

Si bien la distribución de cuentas por grupo es parecida, en lo que respecta a cantidades, no resulto de esa manera en cuanto a las cuentas de usuario en sí. Las cuentas pertenecientes a los grupos derivados de los resultados originados por los centroides iniciales canónicos fueron mas parecidas entre ellas; o sea, quecompartieronmas características. De hecho, como resultado de estos centroides iniciales se formaron grupos entre los que destacaban cuentas dedicadas a la transmisión de noticias, de cuentas inactivas, o de cuentas activas en tuiteo, muy retuiteadas y de rápido crecimiento.

Estudios previos han establecido lo sensible delk-meansa la solución inicial (ver [28]), lo cual era de esperarse por la forma de trabajar del método, y mas aún por la forma de dispersión de las cuentas en el espacio, lo que explica las diferencias en la distribución de cuentas en ambos resultados.

Para dar mayor confiabilidad a los resultados, se determinó un promedio del grado de pertenencia de cada cuenta a el grupo al que fue asignado con el naive-bayes, lo que confirmó el resultado alcanzado por los centroides iniciales canónicos como el más efectivo en la identificación de cuentas con comportamientos similares.

Conclusiones

La presencia cada vez mayor de las redes sociales y las muchas formas en las que permiten expresar nuestras opiniones, gustos y sentimientos, aumenta la atención sobre ellas como fuente de información para una amplia variedad de investigaciones.

Uno de los estudios mas comunes realizados para Twitter es la determinación de tendencias; de qué habla la gente. Pero también existe interés por clasificar las cuentas de usuarios según su comportamiento dentro de la red, nivel de influencia, alcance, si son autenticas o no, entre otras características; para lo cual se suelen establecercriterios basados en un par de características de las cuentas.

El k-meanses una técnica sencilla de implementar, y tomando las consideraciones adecuadas resulta efectiva, no solo en la clasificación de usuarios de Twitter según criterios establecidos, sino también en la determinación de las características mas relevantes que permiten categorizar sus comportamientos dentro de la red. Entre los aspectos a considerar destaca el hecho de que el método es sensible a la configuración inicial de los centroides; sin embargo, el uso de una base canónica arrojó buenos resultados. Además, combinando con otras técnicas como naive-bayes, o alguna métrica, se puede cuantificar la calidad de la solución obtenida, proporcionando una medida de confiabilidad.

Por otra parte, las técnicas de minería de texto aumenta su importancia en el estudio de este tipo de redes sociales, donde análisis de sentimiento, análisis semántico, etc, son herramientas cada vez mas empleadas; sin embargo, sus resultados son posibles de cuantificar e incluir como atributos de vectores característicos que son con los que trabaja el método.

Referencias Bibliográficas

1. Martin Hilbert, Priscila López. "The World's Technological Capacity to Store, Communicate, and Compute Information". Science.Vol. 332 no. 6025 pp. 60-65.Abril 2011.
2. EMC Corporation. http://www.emc.com/leadership/digital-universe/index.htm. "THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East". Diciembre 2012.
3. Qmee. "Lo que sucede en línea en 60 segundos".http://blog.qmee.com/online-in-60-seconds-infographic-a-year-later/
4. Internet World Stats. http://www.internetworldstats.com/
5. Global Web Index. https://www.globalwebindex.net/
6. Eaton, C., Deroos, D., Deutsch, T., Lapis, G., Zikopoulos, P. "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming data." McGraw-Hill. 2012.
7. Rui Han, Zhen Jia, WanlingGao, XinhuiTian, and Lei Wang. "Benchmarking Big Data Systems: State-of-the-Art and Future Directions".TECHNICAL REPORT. ICT, ACS. 4 Jan 2015.
8. HaewoonKwak, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?".Proceedings of the 19th International Conference on World Wide Web.Pp 591-600. 2010.
9. Stefano Rizzi. "New Frontiers in Business Intelligence: Distribution and Personalization". ADBIS 2010, LNCS 6295, pp. 23–30, 2010.
10. MihaelaMuntean. "Theory and Practice in Business Intelligence". MPRA Paper No. 41359, posted 16. September 2012.
11. William Ribarsky, Xiaoyu Wang, and Wenwen Dou."Social Media Analytics for Competitive Advantage".EuroVis Workshop on Visual Analytics (2013), pp. 1–5. 2013.
12. ChristophBoden , Volker Markl , Tu Berlin , Marcel Karnstedt , and Miriam Fernandez. "Large-Scale Social-Media Analytics on Stratosphere". International World Wide Web Conference Committee (IW3C2), pp 257-260. 2013
13. Dror Y. Kenett , Fred Morstatter , H. Eugene Stanley , Huan Liu. "Discovering Social Events through Online Attention".PLoS ONE 9(7). 2014.
14. Kent, P. "Viewpoint: Big data and big analytics means better business", BBC News, Oct 2012, http://www.bbc.co.uk/news/business-19969588
15. Moore, M.T. "Twitter index tracks sentiment on Obama, Romney", USA Today, Jan 2012, http://usatoday30.usatoday.com/news/politics/story/2012-08-01/twitter-political-index/56649678/1
16. Haibin Liu. "EXPLOITING USER-GENERATED DATA FOR KNOWLEDGE DISCOVERY AND RECOMMENDATION".Dissertation in Information Sciences and Technology for the Degree of Doctor of Philosophy.The Pennsylvania State University. August 2014.
17. Haibin Liu and Dongwon Lee."Quantifying Political Legitimacy from Twitter".LNCS 8393, pp. 108–115, 2014.
18. True Social Metrics. https://www.truesocialmetrics.com
19. Sentimentalytics . https://sentimentalytics.com/
20. SEMrush. Competitive Data. http://www.semrush.com/

21. WordStat. http://provalisresearch.com/
22. Attensity. http://www.attensity.com
23. Twitter Documentation. https://dev.twitter.com/overview/documentation
24. OAuth. http://oauth.net
25. Twitter4J. http://twitter4j.org/en/index.html
26. P. Berkhin. "A Survey of Clustering Data Mining Techniques".Chapter of Grouping Multidimensional Data, pp 25-71.Springer Berlin Heidelberg. 2006.
27. Eibe Frank, Mark Hall, and Bernhard Pfahringer. "Locally Weighted Naive Bayes". Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp 249-256. 2003.
28. Villagra, A., Guzmán, A., Pandolfi, D. y Leguizamón, G. "Análisis de medidas no-supervisadas de calidad en clusters obtenidos por K-means y ParticleSwarmOptimization". Trabajo presentado en el Congreso de Inteligencia Computacional Aplicada (CICA), realizado en Buenos Aires del 23 al 24 de julio de 2009.