Plataforma web para la identificación y el análisis de eventos en Twitter

Antonio Juárez-González[1], Griselda Velázquez-Villar[2], Esaú Villatoro-Tello[2], Gabriela Ramírez-de-la-Rosa[2]

[1] Universidad Politécnica de Tlaxcala,
Tecnologías de la Información,
Tlaxcala, México

[2] Universidad Autónoma Metropolitana Unidad Cuajimalpa,
Departamento de Tecnologías de la Información,
México, D.F.

antonio.juarez@uptlax.edu.mx, grisvillar@yahoo.com.mx,
{evillatoro,gramirez}@correo.cua.uam.mx

Resumen. Debido a la gran popularidad que han adquirido actualmente las redes sociales entre personas, empresas, figuras públicas etc., surge la necesidad de contar con métodos automáticos que faciliten la búsqueda, recuperación y análisis de grandes cantidades de información. Ante esto, el Analista de Reputación en Línea (ARL) enfrenta el reto de identificar temas relevantes alrededor de un evento, producto y/o figura pública; a partir de lo cual puede proponer diferentes estrategias para fortalecer y/o revertir tendencias. Por lo tanto, en este trabajo se propone y describe una herramienta web que tiene como objetivo principal apoyar en las tareas desempeñadas por un ARL. Las técnicas de visualización propuestas permiten identificar de manera inmediata la relevancia y el alcance de las opiniones generadas sobre un evento sucedido en Twitter.

Palabras clave: Agrupamiento, medidas de similitud, visualización de información.

1. Introducción

El surgimiento de las redes sociales en Internet han propiciado que un mayor número de personas tenga la posibilidad de publicar libremente opiniones y comentarios acerca de una gran variedad de temas sociales, culturales, deportivos, científicos e incluso opiniones sobre productos y servicios.

Gracias a la popularidad que han adquirido estas redes sociales, actualmente es de gran interés para muchas entidades conocer lo que se dice de ellas dentro de este mundo digital. Al mismo tiempo, a través de estos medios de comunicación, es posible tener un acercamiento con distintos usuarios, mismos que aprovechan esta interacción para dar a conocer de forma específica su opinión sobre determinados temas, productos o servicios. Una de las redes sociales más

utilizadas para este fin es Twitter que permite enviar mensajes cortos llamados tuits (*tweets* en Inglés), con una longitud máxima de 140 caracteres[1]. De acuerdo a Statisticbrain[2] a enero del 2014, el número total de usuarios activos en Twitter asciende a 645,750,000 a nivel mundial y el número de tuits al día es de 58 millones. Por su parte, México ocupa el séptimo lugar entre los países más tuiteros del mundo con casi 15 millones de usuarios.

Con la finalidad de aprovechar la gran cantidad de información obtenida de las interacciones (directas e indirectas) entre usuarios y empresas, estas últimas han creado la figura de un Analista de Reputación en Línea (ARL). El trabajo de este profesional pasa por tres fases: la primera fase consiste en el monitoreo, éste permite conocer en todo momento lo que se está publicando con relación a la empresa, producto o figura pública de interés. La segunda fase es la identificación de temas relevantes dentro de la comunidad de usuarios de Twitter, priorizando los mensajes con las implicaciones más importantes, negativas o positivas, hacia la entidad en cuestión. Finalmente, la tercera fase consiste en proponer estrategias de mercado que permitan revertir los efectos negativos previamente identificados o incluso fortalecer los aspectos positivos de la entidad en revisión.

Debido a que las publicaciones crecen de manera acelerada, el análisis manual de esta información resulta complicado y desgastante para el ARL. En consecuencia, surge la necesidad de contar con sistemas automáticos que permitan realizar este análisis de forma más sencilla y oportuna. Recientemente diversos grupos de investigación están interesados en esta problemática [1] y se han dado a la tarea de desarrollar sistemas enfocados al análisis de opiniones generadas en Twitter. En este contexto surge Replab[3] como un foro internacional en el cual se han propuestos y evaluado distintos sistemas automáticos enfocados en el análisis de la reputación en línea, específicamente de la información producida en Twitter.

Hasta el momento, los diferentes grupos que han participado en Replab, se han enfocado en desarrollar métodos automáticos para tareas como: *i)* selección de tuits relevantes para una entidad[4] [4], *ii)* identificación de implicaciones negativas, positivas o neutrales hacia una entidad [4], *iii)* agrupamiento de opiniones por temática similar [4,3], y *iv)* la identificación de líderes de opinión dentro de una comunidad [8]. Sin embrago, el problema de cómo mostrar el resultado del análisis automático a un ARL de forma que se le facilite la toma de decisiones, ha sido poco explorado. Es por esto, que surge la necesidad de desarrollar sistemas que aprovechen los resultados de estos métodos automáticos y permitan generar representaciones visuales.

A partir de lo anterior, dentro de este trabajo se explora una alternativa de visualización de los resultados producidos por sistemas de análisis de contenido en Twitter, en particular sistemas desarrollados en el marco de la competencia de RepLab. El sistema propuesto se enfoca específicamente en la visualización

[1] http://about.twitter.com/
[2] http://www.statisticbrain.com/twitter-statistics/
[3] http://www.limosine-project.eu/events/
[4] Por *entidad* nos referimos al nombre de una figura pública y/o el de una organización.

de resultados del agrupamiento de opiniones por temática similar y la forma en cómo éstos se relacionan entre sí.

El resto del artículo se encuentra organizado de la siguiente manera. La sección 2 describe algunas de las plataformas existentes más cercanas a los objetivos del presente trabajo. La sección 3 describe detalladamente la arquitectura del sistema desarrollado. Posteriormente, la sección 4 muestra el funcionamiento y principales características de la plataforma web. Finalmente en la sección 5 se mencionan las principales conclusiones e ideas de trabajo futuro derivadas del presente proyecto.

2. Trabajo relacionado

Actualmente, existen disponibles en Internet variadas aplicaciones enfocadas al agrupamiento, análisis y visualización de información. Estas aplicaciones las podemos dividir en dos grandes categorías: *i)* herramientas especializadas en el agrupamiento y clasificación de grandes volúmenes de datos; y *ii)* herramientas no especializadas, que permiten realizar análisis cualitativo de datos a usuarios estándar.

Por un lado, entre las herramientas especializadas vale la pena mencionar que tienden a emplear un lenguaje muy técnico y como consecuencia se dificulta la interpretación de los resultados para usuarios no especialistas. Como ejemplo de éste tipo de herramientas podemos mencionar a Cluto [7] y Weka [6], herramientas multi-plataforma que tienen implementados gran variedad de métodos automáticos para el análisis de datos. Por otro lado, en la segunda categoría están las herramientas que se dedican al análisis de datos que se publican exclusivamente en redes sociales, y que buscan proporcionar a usuarios, expertos e inexpertos, con los elementos suficientes para realizar un análisis fácil e intuitivo de los resultados proporcionados por sus métodos de identificación de temáticas, polaridad, etc. Como ejemplo de tales herramientas podemos mencionar a Spot[5], AnaliticPro[6] y Socialmention*[7]; las cuales proporcionan al usuario variados esquemas de visualización de datos, los cuales buscan resaltar ciertos indicadores que permiten al analista evaluar y determinar la reputación que tiene un producto o tema en particular dentro de una comunidad específica.
Spot. Es una aplicación que permite la visualización interactiva de lo que se está publicando en Twitter en tiempo real. La idea principal es mostrar rápidamente las opiniones que se generan sobre un tema en particular. La forma de presentar los tuits, es a través de grupos contenidos en burbujas, las cuales se organizan y visualizan de diferentes maneras para resaltar distintos tipos de información sobre el tema de interés. Al realizar la búsqueda del tema, solo se recuperaran los últimos 200 tuits para la visualización. Hay que tomar en cuenta que los resultados de búsqueda de Twitter sólo se remontan alrededor de una semana.

[5] `http://neoformix.com/2012/IntroducingSpot.html`
[6] `http://www.analiticpro.cl/caracteristicas.php`
[7] `http://www.socialmention.com`

Por lo que la búsqueda y el análisis sólo se puede realizar sobre un conjunto muy limitado de tuits.

AnaliticPro. Aplicación que permite procesar grandes volúmenes de información producida en distintas redes sociales. Se pueden realizar mediciones con criterios personalizados, además de esto incorpora técnicas que permiten identificar el sentido (*i.e.,* positivo, negativo o neutral) de los comentarios, se pueden relacionar y construir frases para conocer la opinión generalizada dentro de una comunidad respecto a uno o varios temas. Una de las principales desventajas de esta herramienta es que para la construcción de sus informes se basa en técnicas semi-automáticas, es decir, hay informes que son generados con ayuda de expertos. Por tal motivo, para poder explotar al 100 % la infraestructura ofrecida por AnaliticPro se requiere del pago de licencias y/o servicios de análisis de reputación.

Socialmention*. Es una aplicación que monitorea y analiza la información que se está generando en distintas redes sociales de Internet en tiempo real. Además, permite seguir y medir fácilmente lo que se opina sobre alguna persona, empresa, producto, etc. Las búsquedas se realizan en más de 80 medios sociales incluyendo las más visitadas como son: Twitter, Facebook, friendFeed, YouTube, Digg, Google, etc.. A diferencia de las herramientas previas, Socialmention* propone diferentes medidas que facilitan la interpretación de los datos analizados, por ejemplo: *fuerza, sentimientos, pasión y alcance.* Intuitivamente, estas medidas proporcionan al usuario una idea de la importancia y el alcance del tema en revisión.

En general, las herramientas mencionadas anteriormente proponen distintos métodos para el análisis y visualización la información producida en redes sociales respecto a una entidad específica. Principalmente se han enfocado en identificar la polaridad de los comentarios, el origen de los mismos (*i.e.,* red social, dispositivo, tipo de usuario), y las posibles tendencias. Al contrario de las herramientas analizadas, el trabajo desarrollado en este artículo busca proporcionar herramientas que faciliten la identificación de temáticas relevantes y al mismo tiempo la relación que éstas pueden tener con otros aspectos secundarios sucedidos al rededor del mismo evento, aspecto que no es considerado por ninguna de las herramientas revisadas. De esta forma, la herramienta propuesta permitirá al ARL identificar de manera inmediata tanto los temas y sub-temas que suceden al rededor de un evento, así como la relevancia de los mismos. En las siguientes secciones se describe en más detalle el sistema desarrollado.

3. Sistema propuesto

El sistema propuesto se compone de tres grandes módulos, de los cuales el primero se encarga de la búsqueda y recuperación de tuits, posteriormente se hace un proceso de agrupamiento, el cual puede ser temático o no-temático, y finalmente se produce una salida gráfica la cual es mostrada al usuario. En las sub-secciones siguientes describimos los componentes principales de cada uno de estos módulos.

3.1. Recuperación de tuits

La recuperación de los tuits comienza con una consulta,en donde se especifica el tema a buscar. La consulta consiste en una cadena de caracteres que pueden ser uno o varios términos (*i.e.,* consultas compuestas), y el número de tuits que se desean recuperar. Para poder realizar este proceso se utilizó la librería twitter4j[8], la cual permite la conexión con la plataforma de Twitter. Es importante mencionar que para poder usar esta librería es necesario realizar un registro en la página de desarrolladores de Twitter. Este proceso permite la obtención de las llaves de acceso que permiten a sistemas automáticos hacer uso de la información que se genera en esta red social.

Así entonces, si la conexión a Twitter es exitosa, se recuperan los tuits, se almacenan para procesos posteriores, y además se muestran en la interfaz gráfica del sistema. Si en determinado momento sucede algún error de conexión, el usuario es notificado. Es importante mencionar que por medio de almacenar los tuits en una base de datos se permite a los usuarios acceder a su historial de búsquedas, lo cual es una opción importante para los ARL, pues permite analizar tendencias y/o hacer comparaciones de forma inmediata. Otro aspecto importante del proceso de recuperación de tuits es que está diseñado para obtener la mayor cantidad de meta-datos relacionados a cada tuit descargado, por ejemplo: el nombre de usuario, imágenes compartidas, información del perfil de usuario, fecha, hora, plataforma, idioma, etc.

Nótese que un paso previo al almacenamiento de los tuits en la base de datos es un módulo de *preprocesamiento*. Este paso es común en muchas tareas de procesamiento de lenguaje natural, y tiene como principal objetivo eliminar información que se considera sin carga temática. Para este caso se consideraron las siguientes operaciones de preprocesamiento: el texto es llevado a minúsculas, se eliminan símbolos de puntuación, se eliminan las URLs y se quitan palabras funcionales[9]. Finalmente, los tuits originales y pre-procesados quedan almacenados en la base de datos, listos para ser procesados por el módulo de agrupamiento.

3.2. Agrupamiento temático

Un primer paso necesario para realizar a tarea de agrupamiento temático es el *indexado* de los documentos a analizar, actividad que denota hacer el mapeo de un documento d_j en una forma compacta de su contenido. La representación más comúnmente utilizada es un vector con términos[10] ponderados como entradas, concepto tomado del modelo de espacio vectorial usado en recuperación de información [2]. Es decir, un texto d_j es representado como el vector $\overrightarrow{d}_j = \langle w_{kj}, \ldots, w_{|\tau|j} \rangle$, donde τ es el *diccionario*, *i.e.*, el conjunto de términos que ocurren al menos una vez en algún documento, mientras que w_{kj} representa la importancia del término t_k dentro del contenido del documento d_j.

[8] `http://twitter4j.org/en/index.html`

[9] También conocidas como palabras vacías o *stopwords* en Inglés

[10] Entiéndase por términos ya sea palabras y/o n-gramas de palabras o caracteres.

Esta técnica, conocida como bolsa de palabras (BOW[11]), es la forma tradicionalmente utilizada para representar los documentos [5]. Dentro de la herramienta desarrollada se consideraron sólo palabras simples como los elementos del vector. El peso w_{kj} puede ser calculado utilizando diferentes enfoques, el más simple de estos enfoques es el *booleano* que consiste en asignar un valor de 1 al término si éste aparece en el documento, y 0 en caso contrario. Agregado a éste, otros esquemas de pesado muy comunes son el conocido como frecuencia (*TF*) y frecuencia relativa (*TF-IDF*) [9]. Es conveniente mencionar que la herramienta desarrollada incluye estos tres esquemas de pesado.

Una vez que se tiene una representación apropiada de los documentos (*i.e.,* tuits) podemos proceder con el proceso de agrupamiento. Los grupos deben cumplir una serie de propiedades, *e.g.,* los documentos pertenecientes al mismo grupo deben ser muy similares, mientras que al mismo tiempo los documentos que pertenecen a grupos distintos deben ser tan diferentes como sea posible; a estas propiedades se les conoce como *homogeneidad* y *heterogeneidad* respectivamente. En general, para lograr aproximarse a dichas propiedades, en necesario determinar similitudes entre los objetos a partir de los valores de sus atributos; para nuestro caso se utilizó la medida del coseno.

En el sistema descrito en este artículo se trabajó con dos distintas técnicas de agrupamiento, específicamente se implementó un algoritmo de partición (*k-means*) y un algoritmo jerárquico (*Hierarchical Clustering*).

Por un lado, los algoritmos de partición agrupan los elementos entorno a elementos centrales llamados *centroides*. El algoritmo de *k-means* es un método iterativo que tiene como parámetro importante el valor de k (el número de grupos a formar), a pesar de lo cual es un algoritmo muy efectivo [10]. Por otro lado, los algoritmos jerárquicos se caracterizan por generar una estructura de árbol, llamada dendograma, en la que cada nivel del árbol es un posible agrupamiento de los objetos de la colección. El método de *Hierarchical Clustering* es un algoritmo jerárquico de tipo aglomerativo, es decir parte de las hojas del árbol, considerando a cada elemento como un grupo. Posteriormente y de forma iterativa va uniendo elementos en grupos más cercanos hasta que todos los documentos se encuentran dentro de un grupo [10].

Es importante mencionar que en el sistema desarrollado se incorporaron las implementaciones hechas en Weka [6] de los algoritmos de agrupamiento descritos con sus configuraciones por defecto.

3.3. Visualización

Para la representación visual de los resultados de agrupamiento temático y no-temático, se hace uso de la librería D3js[12], específicamente del tipo de gráfica denominada Bubble Chart.

[11] Bag Of Words por sus siglas en Inglés.

[12] D3JS (Data-Driven Documents) es una librería JavaScript que permite manipular y visualizar distintos tipos de datos (`http://d3js.org`).

Cada burbuja de la gráfica representa un grupo, resultado del agrupamiento temático realizado en la etapa anterior. Además, el tamaño de la burbujas representa, hasta cierto punto, la importancia del sub-grupo identificado. Así entonces, dentro de las burbujas que representan a los diferentes sub-grupos se congregan los tuits que corresponden a ese grupo como una serie de burbujas más pequeña. Una de las ventajas de la visualización propuesta es que es posible, mediante el posicionamiento del curso sobre un tuit (*i.e.,* las burbujas más pequeñas), ver el contenido de éste. Adicionalmente, mediante esta gráfica es posible ver los n términos más representativos el grupo, *i.e.*, los términos más frecuentes en los tuits del grupo en cuestión.

Como se mencionó en secciones anteriores, una de las ventajas de nuestro sistema es que permite al ARL, además de identificar los diferentes temas y sub-temas que suceden alrededor de un evento dado, muestra las relaciones temáticas entre los distintos sub-temas. Para lograr esto hacemos los siguiente: *1)* se identifican los conceptos más representativos de cada sub-tema, *2)* para cada par de sub-grupos se buscan los conceptos contenidos en la intersección, y *3)* finalmente, los conceptos compartidos entre cada par de sub-grupos son mostrados al usuario en forma textual. Con esto, en analista puede identificar rápidamente conceptos (palabras) clave que la comunidad de usuarios está empleando para referiste al evento de interés.

Adicionalmente, el sistema desarrollado también permite generar gráficas con información extraída de los meta-datos de los tuits que se han recuperado. Particularmente, es posible construir gráficas agrupando los tuits por plataforma empleada para leer y escribir en Twitter, *e.g.*, Android, IOS, web, etc; por el número de favoritos, número de retuits. En conjunto, esta información resulta de utilidad para el ARL debido a que le permite identificar la relevancia y el alcance del evento que este siendo estudiado.

4. La plataforma en funcionamiento

Para ilustrar el módulo de visualización implementada en la plataforma propuesta, se realizó una búsqueda de tuits sobre el tema *Ayotzinapa*[13]. La consulta recuperó 3,000 tuits, número de tuits que se especificó mediante las opciones de la plataforma web. Con el objetivo de mostrar algunas de las características de la visualización, la Figura 1 muestra dos distintos resultados después de hacer un agrupamiento por temática similar. En la imagen de la izquierda se puede observar que el resultado del agrupamiento generó dos grupos, mientras que la imagen de la derecha se muestra una salida que resultó en tres sub-temas. Es importante recordar que uno de los parámetros requeridos por la plataforma es el valor de k, el cual indica la cantidad de sub-temas que queremos identificar. Intuitivamente, entre mayor sea el valor de k estaremos exigiendo mayor nivel de especialidad en los sub-temas generados, mientras que un valor muy pequeño permite mayor generalidad en los sub-temas.

[13] La consulta se realizó el día 26 de septiembre a las 12:27 horas, el día del aniversario de las desapariciones de 43 normalistas en el estado de Guerrero, México.

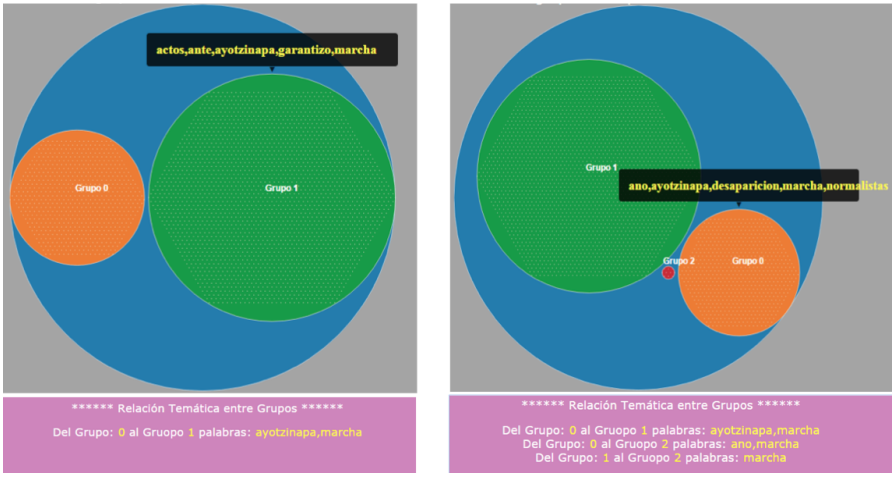

Fig. 1. Visualización del agrupamiento temático para 3000 tuits del tema Ayotzinapa. En la izquierda se muestra el resultado de generar dos grupos, mientras que en la derecha se muestra el resultado de la generación de tres grupos.

Como puede verse en la Figura 1, es posible obtener el conjunto de palabras más representativas de cada grupo, las cuales indican, hasta cierto punto, el contenido semántico de cada sub-grupo. En el caso de la imagen a la izquierda, el Grupo 1 se puede describir con las palabras *actos*, *ante*, *Ayotzinapa*, *garantizo*, y *marcha*, en contraste con el contenido semántico del Grupo 0 que se puede definir por las palabras *año*, *Ayotzinapa*, *desaparición*, *marcha*, *normalistas*. Ante este ejemplo, un ARL podría discernir que mientras todos los tuits de ambos grupos hablan sobre la marcha que se realizó sobre el caso Ayotzinapa, un subgrupo hace referencia al aniversario de la desaparición de normalistas en Ayotzinapa, mientras que el otro subgrupo hace mención sobre las garantías que se prometieron para los actos realizados en el contexto de la marcha. Este tipo de información podría fácilmente corroborase al posicionar el cursor sobre un tuit particular y ver su contenido.

Por otro lado en la imagen de la derecha de la Figura 1, el tercer grupo (Grupo 2) hace mención de los términos descriptivos: *año*, *marcha*, *México*, *normal*, *tragedia*; mientras que el grupo es mucho menor en relación a los grupos 0 y 1, es claro el sub-tema que éste describe, generalizando el problema a nivel país y describiendo el evento como una tragedia.

Agregado a lo anterior, en la Figura 1 también se pueden ver, de manera muy simple, las palabras que los grupos comparten. Por ejemplo, en el caso del agrupamiento de la imagen en la izquierda, dado que sólo existen dos grupos, los términos comunes entre ellos son *Ayotzinapa* y *marcha*. Con esta información sintetizada, el ARL podría tener un panorama general del tema en análisis.

Como se mencionó en la Sección 3.3, además de mostrar información sobre el agrupamiento temático, también es posible ver gráficas de los metadatos de

los tuits. En la Figura 2 se pueden observar las gráficas generadas por tipo de plataforma usada para enviar el tuit, por número de retuits y por número de favoritos (de izquierda a derecha en la Figura 2). De la gráfica que agrupa los tuits por plataforma de publicación podemos ver que de los 3000 tuits recuperados, 928 fueron enviados desde la aplicación de Twitter para Android; de la gráfica que agrupa los tuits por el número de retuits que éstos tienen, es posible ver que unos de los tuits mas retuiteados (436 veces) es un tuit que contiene las palabras temáticas "No dejarse engañar por la telenovela de PGR-Televisa, pide sobreviviente de Ayotzinapa"[14]. Finalmente, la gráfica que visualiza los tuits agrupados por favoritos muestra que el tuit marcado más veces como favorito es el video de la postura de la figura pública Fher (integrante del grupo musical Maná).

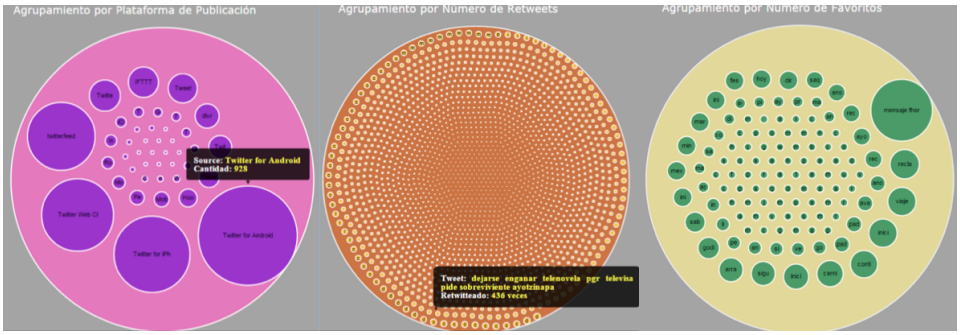

Fig. 2. Visualización por metadatos de una colección de tuits. De izquierda a derecha: agrupamiento por plataforma de publicación, agrupamiento por número de retuits, y agrupamiento por número de favoritos.

5. Conclusiones y trabajo futuro

En este artículo se describió el trabajo realizado para la construcción de una herramienta web diseñada para apoyar en las actividades desempeñadas por un ARL. Específicamente, la herramienta propuesta permite hacer la identificación automática de temas y sub-temas (y la relación temática entre éstos) sucedidos alrededor de un evento ocurrido en Twitter. Una de las ventajas ofrecidas por la aplicación desarrollada es que gracias a su propuesta de visualización de resultados, un analista puede, de manera sencilla e inmediata, identificar la relevancia y el alcance de las opiniones expresadas entorno a un evento de interés.

El uso de técnicas tradicionales de agrupamiento nos permitió definir una estrategia para lograr la identificación de sub-temas dentro de un conjunto de tuits. En particular, se emplearon dos tipos de algoritmos de agrupamiento, de

[14] `https://twitter.com/Revolucion3_0/status/645042141064925184`

partición y jerárquico, los cuales han mostrado ser efectivos en diversas tareas de agrupamiento de documentos. Como forma de representación de los textos se empleó la técnica conocida como bolsa de palabras (BOW), así como varios esquemas de pesado. En general, los métodos y técnicas empleadas son métodos muy intuitivos y en consecuencia fáciles de entender. Sin embargo, es necesario que el ARL conozca el significado de los parámetros que requieren estos métodos para poder hacer un uso eficiente de la herramienta desarrollada.

Como trabajo futuro se pretende adaptar la herramienta de manera que proporcione mayores facilidades a los usuarios para hacer compliación de corpus de tuiter.

Agradecimientos. Los autores agradecen a CONACyT por el apoyo otorgado a través del programa de redes temáticas (Red Temática de Tecnologías del Lenguaje, proyecto no. 260178). Adicionalmente, los autores agradecen también a UPTlax, UAM-C y SNI-CONACyT por todas las facilidades proporcionadas.

Referencias

1. Amigó, E., de Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2014: Author profiling and reputation dimensions for online reputation management. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings. pp. 307–322 (2014)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrival. Addison Wesley (1999)
3. Berrocal, J.L.A., Figuerola, C.G., Ángel Zazo Rodríguez: Reina at replab2013 topic detection task: Community detection. In: Proceedings of the Fourth International Conference of the CLEF initiative (2013)
4. Cossu, J.V., Bigot, B., Bonnefoy, L., Morchid, M., Bost, X., Senay, G., Dufour, R., Bouvier, V., Torres-Moreno, J.M., El-Beze, M.: Lia@replab 2013. In: Proceedings of the Fourth International Conference of the CLEF initiative (2013)
5. F., S.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009)
7. Karypis, G.: Cluto a clustering toolkit. Tech. Rep. Technical Report 02-017, Dept. of Computer Science, University of Minnesota (2002)
8. Ramírez-de-la Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H., Sánchez-Sánchez, C.: Towards automatic detection of user influence in twitter by means of stylistic and behavioral features. In: Gelbukh, A., Espinoza, F., Galicia-Haro, S. (eds.) Human-Inspired Computing and Its Applications, Lecture Notes in Computer Science, vol. 8856, pp. 245–256. Springer International Publishing (2014)
9. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24(5), 513–523 (Aug 1988)
10. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: In KDD Workshop on Text Mining (2000)