Modelado, extracción y análisis de información del flujo de datos de Twitter

Miguel Ángel Domingo, Julià Minguillón
Universitat Oberta de Catalunya
madomingo@gmail.com, jminguillona@uoc.edu

Diciembre 2012

Resumen

En este artículo se propone el análisis de las interacciones entre usuarios de Twitter, tanto lo que se genera alrededor de un usuario concreto como el análisis de un *hashtag* dado durante un periodo de tiempo establecido. En este análisis se tendrán en cuenta los usuarios que intervienen, se detectará la estructura de menciones, retweets y replies que se generan, *hashtags* utilizados y que recursos se han compartido, ya sean imágenes, páginas web u otros recursos. Para evaluar el comportamiento de la herramienta desarrollada se han realizado tres experimentos de captura de unos *hashtags* concretos, que nos permitirán sacar conclusiones sobre el comportamiento del sistema, además de analizar la información ya estructurada del flujo de datos de Twitter.

Keywords
Twitter, Mysql, API Twitter, PHP, Data Mining, Open Data

1. Introducción

En la actualidad las redes sociales son una de las principales fuentes de comunicación entre usuarios de Internet además de ser un lugar de recolección de información ya que cualquier tipo de usuario puede interactuar con el sistema de maneras diferentes. Todas estas características las cumple con creces la red social Twitter, una red social creada a finales del 2006 con una evolución exponencial en el número de usuarios alrededor de 140 millones y sobre 340 millones de Tweets publicados por día. Todos estos usuarios con todos sus mensajes inundan el flujo de datos de esta red y producen una fuente inagotable de información. Esta información que fluye en Twitter se caracteriza por:

- Tener una estructura poco clara y muchas veces confusa.

- La información que fluye está en múltiples idiomas. Será necesario mantener bases de datos capaces de comprender diferentes idiomas para poder interpretar la información que fluye.

- Los límites que se establecen en los mensajes de Twitter, en la mayoría de casos, provocan que la ortografía no sea la más correcta y que se utilicen abreviaciones o palabras nuevas específicas de las redes sociales complicando su interpretación automática. En los artículos [4],[5] concluyen que las técnicas clásicas de detección de entidades o análisis lingüístico no son aplicables a los mensajes de Twitter por sus características especiales. Y muestran nuevas técnicas mucho más complejas como accesos a corpus de palabras extraídos de Wikipedia o incluso sistemas como Web-ngram[1] de Microsoft que mediante probabilidades es capaz de reconocer entidades.

- Palabras nuevas que son propias de la comunicación en Twitter. Será necesario evaluarlas según las características de cada una de ellas.

- Diferentes formas de acceso: página web, correo electrónico, SMS o desde aplicaciones externas.

- Twitter agrupa todos los estratos sociales y de diferentes niveles culturales representados en los Tweets que publican. En el artículo [10] resalta que Twitter aglutina una audiencia que se distribuye en todos los estratos sociales y que permite recolectar todo tipo de opiniones, por ejemplo desde un estudiante hasta el presidente de una multinacional. Con este tipo de distribución las muestras que se obtienen son representativas de todos los ámbitos.

En este proyecto se pretende capturar la información poco estructurada de la red social Twitter y estructurarla en una base de datos relacional. Con esta información ya estructurada será posible generar nueva información y exportarla para que pueda ser procesada por otra herramienta. Con todo esto, el interés del proyecto es alto ya que permitirá de una forma organizada capturar información que fluye sin un patrón claro y poderla interpretar de una forma mucho más concreta.

Este proyecto se sitúa dentro del proyecto *MAVSEL*[2] (Minería, Análisis y Visualización de datos basados en aspectos sociales en *E-learning*). El análisis de las interacciones de un grupo de usuarios de Twitter es una parte importante de este proyecto. La utilización de herramientas de software libre será vital para la realización de este proyecto, ya que se desarrollará sobre un sistema tipo *Gnu/Linux* y se aprovechará toda la versatilidad del mismo aprovechando las herramientas de Software libre que dispone, tanto herramientas de programación como de bases de datos.

Los datos se capturan mediante la API de Twitter. Posteriormente se darán pautas de acceso a la API de Twitter para poder descargar los Tweets. La API

[1] http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx
[2] http://www.ieru.org/projects/mavsel/index.html

de Twitter esta en continuo crecimiento en una constante evolución dotando a esta de cambios, restricciones[3] y mejoras acorde con la evolución de las redes sociales.

En este proyecto nos centraremos en el seguimiento de un evento o congreso, el cual genera mucha información alrededor del mismo, tanto información sobre el evento como recursos compartidos y la aparición de nuevos usuarios que se unen al evento ampliando los horizontes del mismo. Los principales productores de información son los conferenciantes, en inglés *Keynote Speakers* en las conferencias, la propia organización, los asistentes al evento y en muchos casos personas que no están presentes en el mismo pero generan mucha información que fluye por diferentes caminos y con motivaciones diferentes por ejemplo estudiantes o interesados en el tema. Mediante las capturas realizadas de los eventos analizaremos el uso de Twitter como herramienta de transmisión de información y las conclusiones que se pueden sacar a partir de la información estructurada. En [3] los autores realizan entrevistas en el ámbito académico a estudiantes, profesores, investigadores, doctores... y concluyen que la principal motivación para publicar y compartir contenido es *"to share knowledge/ study/ work about their field expertise"*; en general indica *"always"* publica material académico y *"sometimes"* material no académico, lo que nos empuja a pensar que Twitter es una muy buena herramienta para crear un resumen de un evento o más concretamente una conferencia. En [2] se muestra como Twitter puede ser un excelente *backchannel* para seguir una conferencia; además identifica, de una forma no automática los diferentes roles que publican en Twitter. Es una interesante fuente de partida para la investigación que se ha llevado a cabo ya que estructura claramente lo que se pretende extraer de una conferencia, aunque de una forma no automática.

Este artículo está organizado tal y como se describe a continuación. En la sección 2 se dan las ideas básicas del desarrollo del tema de investigación explicando la red social Twitter, los dos modos de acceso, *stream* y *search*, a la API de Twitter para capturar el flujo de datos. En la sección 3 se describe el funcionamiento de la herramienta desarrollada: como se capturan los datos del flujo de datos de Twitter, diseño de la base de datos que da soporte a la herramienta, extracción de los datos que contiene un Tweet, actualización y geolocalización de los usuarios y por último extracción de la información ya estructurada. En la sección 4 se extraen todas las estadísticas sobre los experimentos realizados, permitiéndolos comparar y sacar sacar conclusiones sobre el comportamiento de los eventos analizados desde Twitter. En la sección 5 se dan las conclusiones sobre el proyecto y los resultados del mismo. En esta misma sección se proponen trabajos futuros a partir de la herramienta desarrollada. Por último, se incluye la bibliografía que se ha consultado para realizar este proyecto.

[3] https://dev.twitter.com/blog/changes-coming-to-twitter-api

2. Desarrollo del tema de investigación

2.1. Twitter

Twitter[4] es un servicio microblogging creado en 2006[5] y que ha experimentado un crecimiento imparable durante estos años. Como se comentó en el apartado anterior Twitter se caracteriza por dos conceptos básicos: los usuarios y los Tweets. A continuación se muestran la principales características:

Los usuarios son el principal valor de Twitter y se caracteriza por tener la capacidad de mandar mensajes, Tweets, tanto privados (DM) como públicos visibles en el Time Line (TL) del usuario. Los usuarios tienen Amigos, grupo de usuarios que sigue el usuario, y de Followers, grupo de usuarios que siguen al usuario.

Los mensajes que mandan los usuarios son los Tweets, se caracterizan principalmente por tener un tamaño máximo de 140 caracteres. Existen caracteres especiales que se interpretan por el sistema:

- Mención *@nombre_usuario*: Si se incluye en un Tweet permite referenciar a otro usuario.

- Retweet *RT @nombre_usuario Tweet_usuario*: Si se encuentra algo con esta estructura significa que el usuario ha Retwitteado el mensaje de nombre_usuario y el Tweet Tweet_usuario.

- *Hashtag #palabra* permiten unirse a tendencias, agrupar Tweets sobre un evento o noticia. Cualquier usuario puede unirse a una conversación de un *hashtag* dado.

- Enlaces a recursos: Muchos Tweets disponen de enlaces a recursos externos ya sean páginas web, documentos, fotos o otro tipo de recurso. A causa del límite establecido de 140 caracteres es necesario utilizar servicios para acortar la longitud de la web ya que si no sería imposible su inclusión en los Tweets.

2.2. Evaluación de la API de Twitter

La primera decisión que toma un desarrollador de una herramienta para el acceso al flujo de información de Twitter es la selección de un tipo de acceso a la API; a continuación se muestran las diferencias existentes entre los dos accesos más típicos para capturar los Tweets alrededor de un *hashtag* dado: *Stream vs Search*. La página web 140dev[6] hace un extraordinario recorrido entre los dos modos de acceso. Sus principales características son:

[4] https://twitter.com/

[5] http://www.flickr.com/photos/jackdorsey/182613360/

[6] http://140dev.com/twitter-api-programming-tutorials/aggregating-tweets-search-api-vs-streaming-api/

2.2.1. Stream

La captura de datos mediante Stream se caracteriza por ser una conexión continua. Durante el tiempo que transcurre una captura de un evento, determinado por el usuario de la herramienta, se establece una conexión continua con los servidores de datos de Twitter, recibiendo directamente cada uno de los Tweets que cumplen con los criterios de búsqueda, en nuestro caso un *hashtag* determinado. Esto implica que sólo se capturarán Tweets que estén en el rango de captura, en el tiempo que dure la captura, y nunca Tweets anteriores. Se dice que es una conexión online sin capacidad de recuperar Tweets pasados.

Las consultas pueden ser todo lo complejas que sea necesario y pueden contener varios criterios de búsqueda, por ejemplo combinando *hashtags* con acento y otros sin acento, permitiendo así ampliar el rango de captura.

Otra característica muy importante es que todos los Tweets que se descargan mediante esta técnica facilitan el acceso a los datos, ya que dentro de los Tweets se dispone de información agregada del usuario que realizó el Tweet, además de otras entidades. Esto reduce el coste de análisis, ya que en las entidades está todo estructurado para su consulta y su almacenamiento. Las entidades son campos que disponen de información sobre el Tweet: recursos compartidos, menciones, hashtags...

Para el acceso a este tipo de consultas es necesario usar una cuenta de usuario simple, no es necesario el uso de *Oauth*. No obstante, está previsto que esto cambie ya que el resto de consultas de la API de Twitter usan este protocolo para autentificar al usuario ya que es mucho más seguro.

2.2.2. Search

La captura de datos mediante Search es mucho más sencilla, ya que no es necesario tener constantemente activo el sistema de captura de los Tweets, dado que es posible capturar éstos cuando el evento está en desarrollo ,o lo más habitual, cuando ha finalizado el mismo. No obstante, esta afirmación no es totalmente cierta ya que no es necesario establecer una captura continua pero si cada cierto tiempo, ya que en principio se establece un límite máximo de búsqueda de 1500 Tweets en bloques de 100 y paginados en 15 páginas. Por lo tanto, si el evento es muy popular es posible que si no existe una tasa de refresco apropiada se pierdan Tweets, ya que hay que controlar el tiempo entre diferentes conexiones. Otro problema habitual de la captura mediante search es que dependiendo de la carga del sistema los Tweets almacenados sobre un tema serán más o menos antiguos e incluso ordenados por relevancia de los mismos, con lo cual debe haber un control sobre la captura de los mismos. De una forma general se entregan antes los Tweets más relevantes que incluso los más nuevos.

El acceso viene delimitado por la cantidad de veces que se consume información. Es posible que en momentos de mucha carga se penalize a las IP's, con mayor consumo, con la reducción en la velocidad. Para realizar las consultas es necesario disponer de autentificación *Oauth*[7] para el acceso a la API Twitter

[7] https://dev.twitter.com/docs/auth/oauth

search.

Uno de los grandes problemas que tiene Search es que no aporta valor añadido a los Tweets capturados ya que la información de los usuarios es escasa y además no facilita el acceso a entidades, lo cual penaliza el rendimiento de la herramienta.

3. Descripción de la herramienta

Se ha seleccionado la tecnología de Stream para acceder a Twitter y poder capturar los Tweets del flujo de datos. Como se ha podido ver en la sección 2.2 para casos muy puntuales puede ser una buena idea el uso de Search, pero para un uso más intensivo es mucho más productivo el uso de Stream ya que se reduce el trabajo de la herramienta. Para este menester se han usado las librerías que ofrece el proyecto 140dev[8] en su página web que hacen uso de la clase *Phirehose*[9], que permite el acceso al flujo de datos de Twitter mediante API Stream. Esta consulta es sencilla y no realiza otro tipo de proceso que pueda sobrecargar el sistema. A la hora de configurar la herramienta hay que indicar un usuario válido de Twitter, la base de datos donde se van a almacenar las capturas y por último el término de búsqueda en el flujo de datos de los Tweets. Hay que indicar que el sistema tiene que estar continuamente disponible para no perder ningún Tweet, ya que esta diseñado únicamente para capturar los Tweets de una forma online. En [7], [8] se marcan las pautas de como capturar de una forma sencilla el flujo de Twitter y almacenarlo en una BBDD. En este proyecto se han tenido en cuenta muchas de las pautas que indican.

La herramienta se divide en diferentes componentes o procesos que la hacen muy modular. A lo largo de este capítulo se describirá el funcionamiento de la herramienta, las posibilidades que tiene además de los resultados que se obtienen a la hora de comparar diferentes eventos. En la Figura 1 se puede ver la estructura general de la herramienta desarrollada.

3.1. Captura de Datos

Este desarrollo implementa un consumidor de datos, una extensión de la clase *Phirehose*, que se ejecuta de una forma continua capturando Tweets que coincidan con alguno de los parámetros pasados. Concretamente, para este proyecto se han buscado tres hashtags #catdades, #OSC2012 y #maratóTV3 que nos han permitido seguir la evolución en Twitter de dos conferencias y un evento muy popular en Catalunya. La herramienta *get_tweets* almacena estos datos en la tabla *json_cache*, almacenando hora de captura, el Tweet con toda su información codificada con json, identificador del Tweet e identificador en la BBDD. Es interesante indicar, que hay veces que la captura de Tweets no es totalmente correcta ya que en situaciones con la carga de sistema alta es posible que se

[8] http://140dev.com/free-twitter-api-source-code-library/twitter-database-server/get-tweets-php/

[9] https://github.com/fennb/phirehose/wiki/Introduction

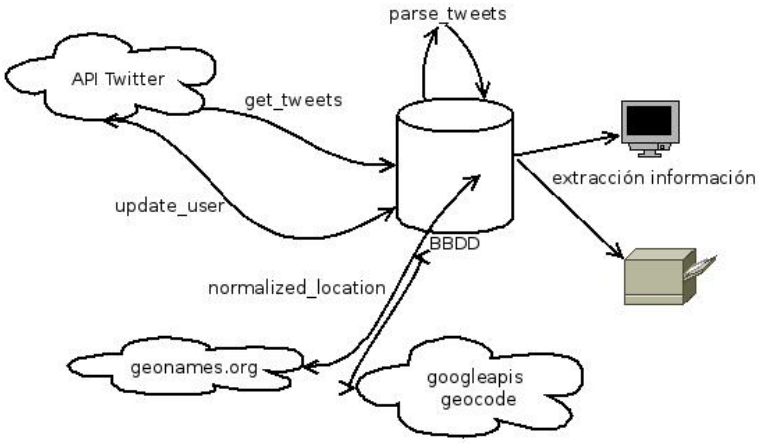


Figura 1: Estructura general de la herramienta propuesta.

repitan identificadores de Tweet, mediante la inclusión de un identificador en la BBDD se corrige posibles duplicidades. La función, que se verá en el apartado 3.3, comprobará que los Tweets que vaya almacenando en la tabla Tweets no se repitan y si se repiten se descarta el último en llegar.

3.2. Base de datos: Mysql

Antes de comentar el desarrollo de la herramienta, hay que indicar que para el correcto funcionamiento de ésta se ha desarrollado una BBDD en el gestor Mysql que nos permite tener almacenados todos los Tweets, además de todas las entidades que facilitan una posterior consulta. En la Figura 2 se pueden ver las tablas. A continuación se irán describiendo.

El sistema de captura del flujo de datos de Twitter almacena todos los Tweets que se capturan en la tabla *json_cache*. En esta tabla se almacenan los Tweets tal y como se capturan en formato Json[10]. Sería posible separar directamente todos los campos, aunque esto produce sobrecarga en el sistema e incluso puede producir la perdida de Tweets.

En esta BBDD se han incluido tres corpus de palabras frecuentes en tres idiomas: catalán, inglés y castellano (*RAE*). El sistema está preparado para incluir otros corpus de idiomas sin ningún esfuerzo. Además es posible ampliarlos para detectar un mayor número de palabras, para aumentar la estrategia de búsqueda de las palabras más repetidas que se comentará en posteriormente.

De una forma offline, para cada uno de los Tweets almacenados en la tabla *json_cache* se extrae el Tweet y todos sus campos y se almacena en cada una de las tablas. A continuación se describe brevemente cada una de ellas:

- *tweets*: Almacena los campos básicos que hacen referencia a un Tweet: texto, usuario que lo ha creado y su posicionamiento global si existe.

[10] http://www.json.org/

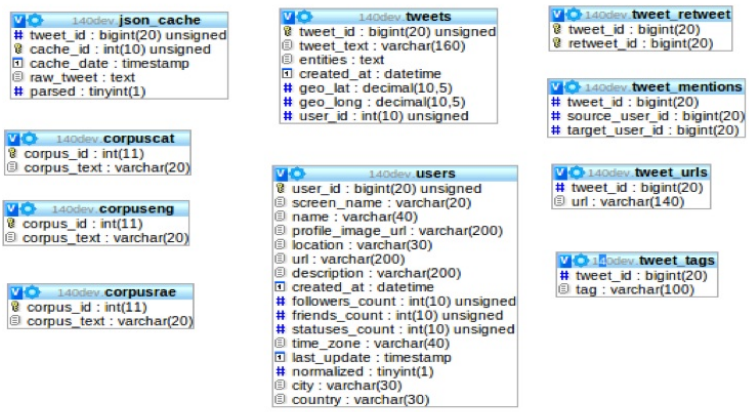


Figura 2: Estructura de las tablas de la BBDD.

- *users*: Se almacenan los campos más básico que caracterizan a un usuario: el nombre, cantidad de amigos, cantidad de followers, cantidad de Tweets, localización y la localización normalizada, además de los campos que se pueden ver en la Figura 2.

- *tweet_retweet*: Esta tabla expresa un comportamiento muy habitual en Twitter, el retweet (RT), que permite difundir un Tweet A1 creado por un usuario U1 por otro usuario U2, creando un nuevo Tweet A2. Esta tabla guarda el identificador del Tweet origen y el identificador del Tweet que se ha generado en el RT.

- *tweet_mentions*: En cada uno de los Tweets es posible que existan menciones, @, a otros usuarios. En esta tabla se asocian el usuario que menciona con los mencionados, pudiendo a posterioridad mostrar las relaciones que se establecen entre los usuarios en una captura específica.

- *tweet_urls*: Los usuarios pueden compartir recursos en sus Tweets. En esta tabla se asocian los Tweets con los recursos compartidos, para facilitar una posterior consulta.

- *tweet_tags*: Los usuarios usan una característica de Twitter como es anteponer el símbolo # antes de una palabra para identificar datos importantes, incluirse en una conversación que se origina alrededor de un concepto o participar en una tendencia. Esto se conoce como *hashtag*, palabras que anteponen a ese símbolo y son conocidas de antemano. En esta tabla se asocian todos los Tweets con los posibles *hashtags* que hay en él, de forma que se podrá consultar cual ha sido el más usado y extraer conclusiones.

Estas tablas permiten trabajar con los datos de una forma cómoda y rápida para poder extraer conclusiones sobre como se comporta Twitter ante un evento

y qué información se comparte.

3.3. Extracción de entidades

La función *parse_tweets* se activa cuando finaliza la adquisición de Tweets mediante *get_tweets*. Esta función podría funcionar de una forma paralela a la captura de datos, aunque podría producir errores de sobrecarga en el sistema, con lo que se descarta este uso. Esta función recorre toda la tabla *json_cache* extrayendo los Tweets. La estructura de Stream[11] de Tweets aporta un valor añadido entidades, ya que incluye valores tales como:

- Usuario que lo creó. Se crea un nuevo campo en la tabla *users* o se actualiza un usuario existente.
- Si es un retweet, simplemente se busca el campo que contiene el identificador del Tweet origen. Se actualiza la tabla *tweet_retweet*.
- Si existen menciones, @, se almacenan en la tabla *tweet_mention*.
- Si existen recursos compartidos se almacenan en la tabla *tweet_urls*.
- Si existen tags, #, se almacenan en la tabla *tweet_tags*.

Con todo esto, el sistema esta prácticamente preparado para recibir consultas.

3.4. Actualización usuarios y localización

Para aumentar la calidad de la herramienta se ha diseñado un sistema capaz de actualizar la información de los usuarios, ya que esta se modifica con el tiempo. Mediante el acceso a la API de Twitter *users/lookup* se descarga toda la información actualizada de los usuarios. Esta consulta, para optimizar el rendimiento, se realiza en bloques de 100 usuarios.

Un dato importante a la hora de comprender el desarrollo de un evento es situar a los usuarios en un lugar y a partir de esa información extraer conclusiones de tendencias, lugares más activos y demás información. Esta información se puede extraer de dos campos de la tabla *users* y *tweets* respectivamente:

- *users.location*: Esta información es completada por el usuario y en un amplio espectro de usuarios este campo es erróneo. Suele incluir información que no tiene que ver con la localización o incluso en muchos casos incompleta o mal estructurada.
- *tweet.geo_lat, tweet.geo_long*. Muy pocos tweets tienen activo el campo de situación lo que impide su localización.

En la Tabla 1 se pueden ver los resultados obtenidos mediante la herramienta desarrollada. Para alcanzar estas cifras, la herramienta sigue estos pasos para poder normalizar la localización de un usuario:

[11] https://dev.twitter.com/docs/platform-objects/tweets

Evento	Usuarios	Normalizados	Sin información	Inf. errónea	Tweets Geo.
#catdades	305	201 (65,9 %)	70 (22,95 %)	34 (11,15 %)	30 (2,33 %)
#OSC2012	241	167 (66,29 %)	50 (20,75 %)	24 (9,96 %)	35 (5,38 %)
#maratóTV3	18614	9432 (50,67 %)	6659 (35,77 %)	2523 (13,55 %)	361 (1,01 %)

Cuadro 1: Estadísticas localización Usuarios.

1. Si el usuario tiene un Tweet con geolocalización se manda este dato a la API de Google maps[12], la cual nos devuelve la localización, ciudad y país.

2. Para todos los usuarios sin información en el campo *users.location* la herramienta los marca como que no se han podido normalizar, por lo tanto ya no se volverán a evaluar. Si se actualiza el usuario este volverá a ser susceptible de ser localizado.

3. Para el resto de usuarios, con valor de *users.location* distinto de vacío y no normalizado se limpia la localización de símbolos que puedan interpretarse erróneamente y se dejan sólo caracteres alfanuméricos. Esta información, ya limpia, se manda a la API *geonames*[13], la cual nos devuelve ciudad y país normalizados o ,en caso de no encontrarlo, error. Se actualizan los usuarios según el caso.

4. Existe un caso especial de valor en *users.location* en el que los datos son los valores de longitud y latitud separadas por una coma. Se extrae esta información y se actual igual que con los Tweets.

Como se verá en la extracción de información esta información nos permite situar perfectamente a un usuario en el mapa. La Tabla 1 nos muestra muy claramente que la tasa de normalización se ve muy penalizada por usuarios sin información de localización y en menor medida con información errónea. La localización de los Tweets como se puede ver es muy baja prácticamente inexistente, llegando en el caso de #maratóTV3 a ser de un despreciable 1 %.

3.5. Experimentos realizados

Se han desarrollado diferentes consultas para extraer la información más relevante de los datos capturados. Es interesante observar que los datos permiten, al estar ya estructurados, mediante sencillas consultas a la base de datos extraer información muy valiosa que nos permite comprender el desarrollo de un evento. A continuación se muestran unos ejemplos indicando cómo se ha implementado. Hay que indicar que se hace uno de PHP y consultas Mysql sobre un entorno HTML y se hace uso de diferentes API. Como ejemplo se toman los datos extraídos de *I Jornada de Catalunya Dades* con el *hashtag* #catdades.

[12] http://maps.googleapis.com
[13] http://api.geonames.org

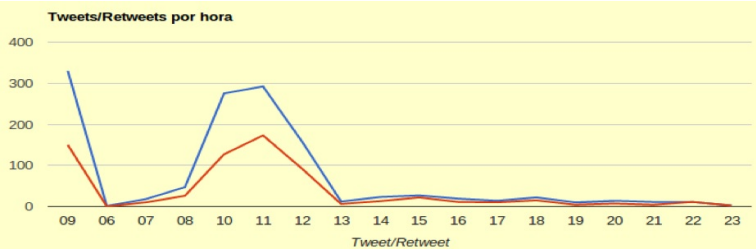


Figura 3: Evolución Tweets y Retweets.

Para comprender la evolución de los Tweets es necesario mostrar una evolución en el tiempo de los mismos. En la Figura 3 se puede comprobar como la evolución en el tiempo tanto de los Tweets como de los Retweets es pareja, es decir cuanto más Tweets más Retweets se producen. Además, en esta gráfica muestra que los puntos en los que hay más Tweets son los momentos en los qué la conferencia estaba en pleno apogeo. Se puede deducir sólo viendo la gráfica en que momentos se estaba realizando la conferencia.

Mediante una sencilla extracción de datos del servidor BBDD y la utilización de la API Google Chart [14] (en concreto *LineChart*) se nos permite mostrar estos datos de una forma rápida.

La herramienta desarrollada es capaz de detectar las palabras que más se han repetido y seleccionar las palabras que no son tan habituales y, por lo tanto, que pueden aportar más información. Se muestra un ejemplo en la Figura 4.

Estas palabras nos dan un pequeño resumen de los conceptos más repetidos en los eventos. La herramienta recoge todas la palabras en un array y almacena la cantidad de repeticiones. Para facilitar el trabajo se eliminan acentos y símbolos externos a alfanuméricos. De este array se eliminan las palabras que aparezcan en los corpus. Para el desarrollo de este proyecto se han creado tres corpus de palabras más repetidas del catalán, castellano e inglés, que corresponden a las 1000 palabras más usuales de cada uno de los idiomas. Mediante este sistema se eliminan las palabras más usuales y se buscan las palabras con una repetición alta pero que no sean de uso habitual. Sólo viendo la imagen se puede deducir entorno a que giraba la Jornada: “Jornada”, “Dades”, “Obertes”, “Reutilitzacio”, ”Catalunya”, ”2013”...

En [4],[5] se describen las bases de como identificar las entidades y en [9] se describe como detectar palabras con una cadencia de aparición no habitual. Estos conocimientos se han aplicado en este proyecto.

Como se ha comentado a lo largo de ester trabajo otro punto importante a la hora de extraer información es mostrar los *hashtag* que nos permitan encontrar nuevas tendencias o incluso nuevos campos de investigación gracias a las micro-comunidades que se generan alrededor de un *hashtag*. Se muestra un ejemplo en la Figura 5.

Gracias a las facilidades que aporta una base de datos es sencillo extraer esta

[14] https://developers.google.com/chart/

Figura 4: Nube de frecuencias de aparición de palabras.

Figura 5: Nube de frecuencias de aparición de hashtags.

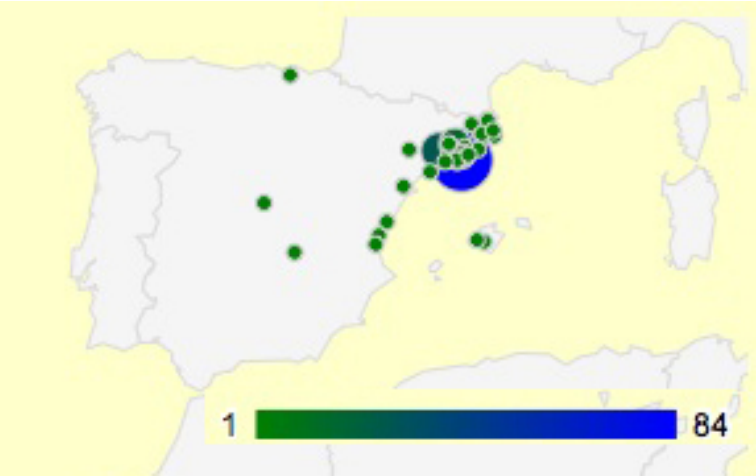


Figura 6: Mapa de situación geográfica de los usuarios.

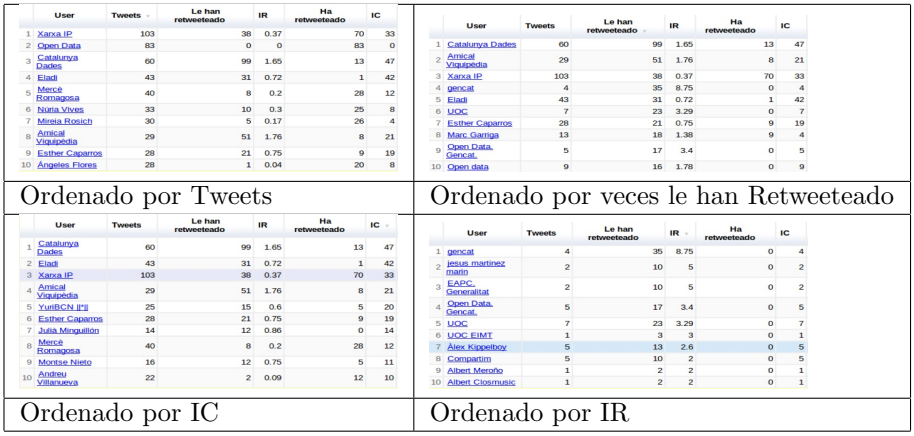
información desde la tabla *tweets_tags*. Alrededor del *hashtag* principal aparecen otros como "Opendata", "Openacces", "Drupal", "Xarxa", "TV3"... que facilitan la búsqueda de información que se genera alrededor de los eventos.

En el apartado 3.4 se comentó la necesidad de buscar a partir de los datos que se disponen la información de la localización de los usuarios para poderlos situar en el mapa. Esto nos permite mostrar de donde son los usuarios así como poder ver con facilidad núcleos centrales de envío de Tweets sobre un evento. En la Figura 6 se puede ver un ejemplo de la situación de los usuarios.

Se observa que el principal foco de usuarios se sitúa donde se celebra el evento. A partir de este punto central se muestran diferentes focos donde aparecen usuarios. Se dispone de la información sobre la localización ya normalizada, lo que nos permite realizar una extracción de la información de forma sencilla. Para mostrar esta información sólo hay que seguir las pautas que indica la API Google Chart para mostrar la información en el mapa. Los datos que requiere esta API es una tabla con la localización y cantidad de usuario para esa localización.

Ya con toda esta información nos queda profundizar en los usuarios, la parte más importante de Twitter. Para este menester se ha creado una tabla dinámica que permite ordenar los usuarios mediante diferentes criterios.

- Tweets generados en el evento.
- Cuantas veces le han retweetado.
- IR: El índice de resonancia del usuario indica la relevancia o impacto en la conferencia de un usuario. Un Tweet del usuario cuantos ReTweets provoca.
- IC: El índice de creatividad identifica a los usuarios más creativos com el número de Tweets originales del usuario.



	User	Tweets	Le han retweeteado	IR	Ha retweeteado	IC
1	Xarxa IP	103	38	0.37	70	33
2	Open Data	83	0	0	83	0
3	Catalunya Dades	60	99	1.65	13	47
4	Eladi	43	31	0.72	1	42
5	Mercè Romagosa	40	8	0.2	28	12
6	Núria Vives	33	10	0.3	25	8
7	Mireia Rosich	30	5	0.17	26	4
8	Amical Viquipèdia	29	51	1.76	8	21
9	Esther Caparros	28	21	0.75	9	19
10	Angeles Flores	28	1	0.04	20	8

Ordenado por Tweets

	User	Tweets	Le han retweeteado	IR	Ha retweeteado	IC
1	Catalunya Dades	60	99	1.65	13	47
2	Amical Viquipèdia	29	51	1.76	8	21
3	Xarxa IP	103	38	0.37	70	33
4	gencat	4	35	8.75	0	4
5	Eladi	43	31	0.72	1	42
6	UOC	7	23	3.29	0	7
7	Esther Caparros	28	21	0.75	9	19
8	Marc Garriga	13	18	1.38	9	4
9	Open Data. Gencat.	5	17	3.4	0	5
10	Open data	9	16	1.78	0	9

Ordenado por veces le han Retweeteado

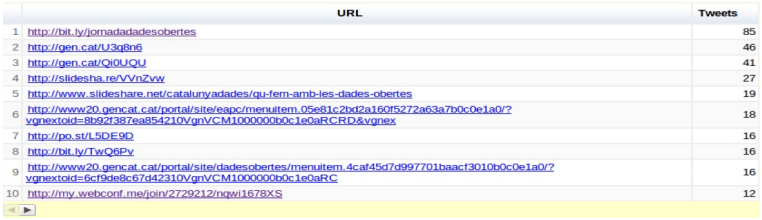
	User	Tweets	Le han retweeteado	IR	Ha retweeteado	IC
1	Catalunya Dades	60	99	1.65	13	47
2	Eladi	43	31	0.72	1	42
3	Xarxa IP	103	38	0.37	70	33
4	Amical Viquipèdia	29	51	1.76	8	21
5	YuriBCN II*II	25	15	0.6	5	20
6	Esther Caparros	28	21	0.75	9	19
7	Julià Minguillon	14	12	0.86	0	14
8	Mercè Romagosa	40	8	0.2	28	12
9	Montse Nieto	16	12	0.75	5	11
10	Andreu Villanueva	22	2	0.09	12	10

Ordenado por IC

	User	Tweets	Le han retweeteado	IR	Ha retweeteado	IC
1	gencat	4	35	8.75	0	4
2	jesus martinez marin	2	10	5	0	2
3	EAPC. Generalitat	2	10	5	0	2
4	Open Data. Gencat.	5	17	3.4	0	5
5	UOC	7	23	3.29	0	7
6	UOC EIMT	1	3	3	0	1
7	Àlex Kippelboy	5	13	2.6	0	5
8	Compartim	5	10	2	0	5
9	Albert Meroño	1	2	2	0	1
10	Albert Closmusic	1	2	2	0	1

Ordenado por IR

Cuadro 2: Ordenación de usuarios según índice.



	URL	Tweets
1	http://bit.ly/jornadadadesobertes	85
2	http://gen.cat/U3g8n6	46
3	http://gen.cat/Qi0UQU	41
4	http://slidesha.re/VVnZvw	27
5	http://www.slideshare.net/catalunyadades/qu-fem-amb-les-dades-obertes	19
6	http://www20.gencat.cat/portal/site/eapc/menuitem.05e81c2bd2a160f5272a63a7b0c0e1a0/?vgnextoid=8b92f387ea854210VgnVCM1000000b0c1e0aRCRD&vgnex	18
7	http://po.st/L5DE9D	16
8	http://bit.ly/TwQ6Pv	16
9	http://www20.gencat.cat/portal/site/dadesobertes/menuitem.4caf45d7d997701baacf3010b0c0e1a0/?vgnextoid=6cf9de8c67d42310VgnVCM1000000b0c1e0aRC	16
10	http://my.webconf.me/join/2729212/nqwi1678XS	12

Figura 7: Listado de recursos compartidos.

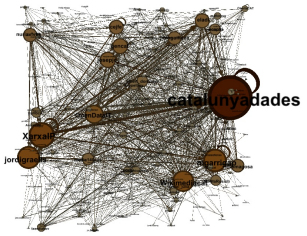
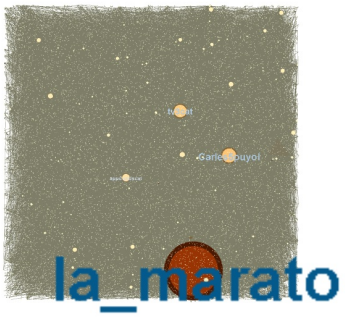
- Influencia: La influencia que tiene el usuario sobre Twitter. Se calcula como followers X tweets + 2 X menciones + 2 X Retweets.

- Menciones, followers, Friends y status. Otros campos que nos permiten ordenar los resultados según criterios.

Dependiendo de lo que el usuario de la herramienta este buscando podrá seleccionar el tipo de usuario que le interesa. Ya sea los que han tenido más impacto en el evento o en influencia, con más menciones o incluso con más Tweets referentes al evento. Como se puede ver en la Tabla 2 es sencillo su uso y facilita poder seguir a los usuarios simplemente pinchando en el nombre del mismo. En la Tabla 2 se pueden ver cuatro ejemplos de ordenación por diferentes índices.

Otro punto importante es mostrar que recursos han sido compartidos por los usuarios. Se ha desarrollo una tabla con la ayuda de la API Google Chart para listar de una forma ordenada los recursos según el número de veces que se ha incluido en los Tweets capturados. Se considera que que un recurso es más importante a mayor cantidad de referencias desde los Tweets capturados. En la Figura 7 se muestra una captura de los diez primeros recursos ordenados, lo cual facilita su consulta.

Como ha quedado demostrado con estos ejemplos las posibilidades que apa-

	
#catdades	#maratóTV3

Cuadro 3: Menciones entre usuarios extraidas de Gephi.

recen cuando se estructura una información que fluye por Tweeter sobre un evento son inmensas, además esta información se puede extraer y representarla en otro formato para que pueda ser interpretada por otras herramientas externas. Existen muchos ejemplos como Graphviz[15], Gephi[16], NodeXL[17] o las Fusión Tables de Google[18] cada uno con sus características propias que permiten mostrar las relaciones que se generan alrededor de un evento. En este proyecto hemos probado los cuatro, pero hemos seleccionado el proyecto de Software Libre Gephi. En el siguiente dirección [19] se pueden ver las posibilidades que aporta esta herramienta. Los datos que se muestran en el vídeo se extraen de la tabla tweet_mentions, las menciones entre usuarios, que incluye a información de usuarios origen y destino en formato CSV.

En la Tabla 3 se pueden ver las relaciones que se han establecido entre los usuarios mediante las menciones. Una corresponde al vídeo anterior donde claramente se pueden ver los actores principales del evento y las relaciones que se establecen y la otra Figura simplemente muestra que hay un par de nodos centrales y uno muy importante y el resto son actores que giran alrededor de este central. En la segunda Figura se pude ver la cantidad de relaciones que han aparecido en este evento proporcional a la cantidad de Tweets del mismo.

4. Comparación de los experimentos

Para evaluar el comportamiento de la herramienta desarrollada se han realizado tres experimentos con connotaciones diferentes e idiomas diferentes para poder comprobar que la extracción de información se comporta correctamente. Se han seleccionado dos conferencias y un evento con una carga de Tweets mu-

[15] http://www.graphviz.org/
[16] https://gephi.org/
[17] http://nodexl.codeplex.com/
[18] https://www.google.com/fusiontables
[19] https://www.dropbox.com/s/089odkycnf81m48/catdades.avi

hashtag	Tweets	Retweets	Tweet_url	Tweet_tags	Tweet_mentions
#catdades	1286	683 (53,11 %)	566	1727	1804
#OSC2012	650	298 (45,85 %)	160	975	487
#maratóTV3	35841	22242 (62,06 %)	7497	47292	31662

Cuadro 4: Tabla de comparación entre eventos.

cho más alta, lo que nos permite comprobar el funcionamiento del sistema ante mucha información.

1. #catdades I Jornada de Catalunya Dades[20], 9 de noviembre del 2012.

2. #OSC2012 Open Source conference 2012[21], 14 de diciembre del 2012

3. #maratóTV3 Marató TV3 2012[22], 16 de diciembre del 2012

En la tabla 4 se puede ver los datos extraídos de cada una de los eventos capturados. De una forma general hay que indicar que hay un peso muy importante en el flujo de datos de Twitter de los retweets, prácticamente la mitad de datos que fluye son retweets lo que implica que original sólo hay un 50 % de datos. Es interesante observar que más o menos de cada dos Tweets uno de ellos tiene más de un hashtag. Los datos demuestran que habitualmente se incluye uno o dos hashtags en cada Tweet.

En la captura de Tweets de #catdades se extrae que la mitad de los tweets que circularon compartían recursos, , a diferencia del resto de eventos, quizás gracias al carácter de datos abiertos. También, esta conferencia se diferencia del resto ya que hay muchas más menciones lo que indica mucha más relación entre los usuarios participantes del hashtag. Además, la media de Tweets por usuario 4,22 con mucha diferencia superior a las otras dos 2,70 y 1,93 respectivamente, lo que indica una mayor participación por usuario.

5. Conclusiones y trabajo futuro

A lo largo de este artículo se ha demostrado que una buena estructura de datos permite una extracción sencilla de datos permitiendo su interpretación o transformación. En este proyecto se ha comprobado que partiendo de una estructura claramente heterogénea de información y sin un nexo de unión aparente, se ha conseguido estructurar un sistema capaz de facilitar la extracción de datos.

A partir de tres experimentos se han dado las bases para comprender como se desarrolla un evento desde el prisma de Twitter: usuarios, Tweets, recursos... Twitter es una fuente inagotable de información susceptible de ser interpretada

[20] http://catalunyadades.wordpress.com/
[21] http://www.opensourceconference.nl/
[22] http://www.tv3.cat/marato/

bajo diferentes enfoques. Se ha visto que existen patrones que se pueden asociar a todos los eventos y otros por el contrario son específicos de cada evento.

Se plantean diferentes retos como trabajo futuro, principalmente realizar más experimentos en diferentes eventos para comprobar que patrones o estadísticas son más repetidas, por ejemplo porcentaje de retweets. Otro camino de investigación que se contempla es el desarrollo de un sistema de reputación que capacite a los usuarios en base a las aportaciones que realicen en Twitter bajo diferentes parámetros y faciliten a los usuarios interesados en seleccionar a los usuarios más experimentados y con mejor reputación. Los parámetros a incluir son número de apariciones en los diferentes eventos, número de Tweets parciales y totales en los eventos, número de menciones, número de retweets que les han hecho.

Otro campo de investigación es la detección de sentimientos alrededor de un concepto dado, en especial, sobre un tema de un evento. En [10] se muestra que a partir de una muestra significativa de conceptos positivos y negativos el sistema es capaz de clasificar un mensaje según el sentimiento que tiene sobre el tema, es decir, positivo, negativo o neutro. Sería un gran campo de investigación para mejorar que camino llevan las opiniones sobre un punto de un evento.

Un trabajo futuro interesante es acoplar las estructura de datos generadas a las especificaciones de SIOC-project[23] que define un esquema de datos que facilita el acceso, comunicación e intercambio de información entre diferentes sistemas. Cumpliendo con estas especificaciones se podrá unir esta nueva web semántica a una red ontológica para facilitar su acceso. Lo que nos permitirá mantener la comunicación con diferentes sistemas en línea permitiendo una búsqueda estructurada dentro de las comunidades online, en especial Twitter. Estas estructuras se hacen necesarias ya que las comunidades online, ya que son poco estructuradas y dificultan su acceso y/o consulta.

Por último es muy interesante aplicar una nueva arquitectura a la herramienta de captura de datos. Mediante pequeños clientes alojados en diferentes IP que capturen de una forma continua los mismos datos. En [1] se desarrolla un sistema de multiples clientes y un servidor central. que es susceptible de ser aplicado en este proyecto ya que en esta investigación se han detectado perdidas de Tweets. Por ejemplo en #maratóTV3 se estimó una pérdida aproximada de 0,05 % de Tweets que se transmitieron ese día, aproximadamente unos 182 sobre 35841. Aunque no es una cifra muy alta, si se instalaran un par de nodos cliente y se juntaran los datos capturados, el sistema sería mucho más robusto y más seguro ante la perdida de conexión o caída del sistema. Es interesante observar que sobre este tema en [6] se indica que con una captura lo suficientemente grande es posible reproducir una información fiable aún a pesar de la perdida de mensajes que se producen por la sobrecarga de mensajes que fluyen en Twitter.

[23] http://sioc-project.org/

Agradecimientos

Este artículo ha estado parcialmente financiado por el proyecto MAVSEL (ref. TIN2010-21715-C02-02).

Referencias

[1] Matko Bošnjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. Twitterecho - a distributed focused crawler to support open research with twitter data. *Proc. of the Intl. Workshop on Social MediaApplications in News and Entertainment (SMANE2012), at the ACM 2012 International World Wide WebConference, WWW 2012*, 2012.

[2] Axel Bruns. How long is a tweet? mapping dynamic conversation networks on twitter using gawk and gephi. *Information, Communication & Society*, 15(9):1323–1351, 2012.

[3] Julie Letierce, Alexandre Passant, Stefan Decker, and John G. Breslin. Understanding how twitter is used to spread scientific messages. *Web Science Conf.*, 2010.

[4] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: Segment-based event detection from tweets. *In Proc. of ACM Conference on Information andKnowledge Management (CIKM'12)*, 2012.

[5] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. Twiner: Named entity recognition in targeted twitterstream. *Proceedings of the 35th international ACM SIGIRconference on Research and development in information retrieval*, pages 721–730, 2012.

[6] Luis F. Lopes, Joao M. Zamite, Bruno C. Tavares, Francisco M. Couto, Fabricio Silv, and Mario J. Silva. Automated social network epidemic data collector. *INForum informatics symposium.*, 2009.

[7] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Processing and visualizing the data in tweets. *ACM SIGMOD Record*, 40(4):21–27, 2011.

[8] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. *Proceedings of the 2011 annual conference on Humanfactors in computing systems*, pages 721–730, 2011.

[9] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. *Proceedings of the 2010 international conference onManagement of data*, pages 1155–1158, 2010.

[10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinionmining. *Proceedings of the Seventh International Conference onLanguage Resources and Evaluation (LREC'10)*, 2010.