Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 57-64 recibido 11-01-12 revisado 06-03-12 aceptado 18-03-12

Generación adaptativa de consultas para la recuperación temática de tweets

Adaptive query generation for topic-based tweet retrieval

Juan M. Cotelo, Fermín L. Cruz, Jose A. Troyano
Dep. de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Avda. Reina Mercedes s/n
41012 Sevilla
{jcotelo,fcruz,troyano}@us.es

Resumen: Twitter se ha convertido en un recurso con gran potencial a la hora de analizar los estados de opinión acerca de temas de actualidad. En el presente trabajo mostramos la metodología utilizada para la obtención de un corpus de mensajes de Twitter relacionados con las elecciones generales españolas del 20 de noviembre de 2011. Dado que el acceso a los mensajes en Twitter se realiza mediante consultas, hemos estudiado diversas estrategias de construcción de dichas consultas, tratando de maximizar la cobertura obtenida. Tras experimentar con diversos acercamientos, se propone un método basado en grafos que permite la captura de *tweets* relacionados con una temática determinada, adaptando dinámicamente las consultas utilizadas para incorporar automáticamente los temas relacionados que eventualmente vayan surgiendo. El recurso obtenido, de gran utilidad, entre otros, en trabajos de análisis del sentimiento, está públicamente disponible para su utilización.
Palabras clave: Recuperación de información, Twitter, análisis de grafos

Abstract: Twitter has become a resource of great potential for analyzing opinion about hot topics. In this paper we show the methodology used for obtaning a corpus of Twitter messages related to the Spanish general elections of November 20, 2011. Given that access to Twitter messages is done through querying, we have studied various strategies for building such queries, trying to maximize the coverage. After experimenting with several approaches, we propose a graph-based method that allows retrieval of tweets related to a specific topic, dynamically adapting the queries to automatically include related topics that eventually arise. The obtained resource, very useful for, among others, sentiment analysis tasks, is publicy available for use.
Keywords: Information retrieval, Twitter, graph analysis

1. Introducción

Twitter ha obtenido mucha atención en el campo investigador en los últimos años debido principalmente a dos factores:por un lado, un aumento exponencial en volumen tanto de usuarios como de mensajes; por otro lado, la disponibilidad de una API pública para acceder a los datos. Este reciente interés se ha materializado en la creación de multitud de aplicaciones web y trabajos de investigación que abordan Twitter explotando las características de red social, evolución temporal y brevedad de los mensajes que exhibe.

La captura de datos en Twitter no es trivial, permitiéndose únicamente el acceso a datos mediante consultas compuestas principalmente por términos de búsqueda. La estructura de la red es tan amplia y grande, que no es fácil determinar los términos o usuarios mas relevantes respecto a lo que se quiere capturar, siendo impracticable capturar todos los *tweets* por motivos técnicos. Por ello, la construcción de consultas sobre Twitter que permitan la captura de datos relevantes con la suficiente cobertura es un problema interesante en sí mismo.

A pesar de ello, la mayoría de los trabajos de investigación se conforman con la utili-

ISSN 1135-5948

zación de consultas simples para obtener los datos. Dependiendo del problema abordado en cada caso, esto puede ocasionar la pérdida de datos interesantes y por tanto la obtención de resultados y conclusiones parcialmente incorrectos. En trabajos recientes relacionados con Twitter ((Agarwal et al., 2011) (Davidov, Tsur, y Rappoport, 2010a), (Davidov, Tsur, y Rappoport, 2010b), (Go, Bhayani, y Huang, 2009), (Jiang et al., 2011), (Kim et al., 2009) (Pak y Paroubek, 2010), (Silva et al., 2011), (Tan et al., 2011), (Tumasjan et al., 2010), (Hong y Nadler, 2011), (Pennacchiotti y Popescu, 2011), (Congosto, Fernández, y Egido, 2011)), se utilizan una serie de listas estáticas de términos decididas manualmente; esta aproximación es demasiado simplista y no proporciona garantías de cobertura, precisión y calidad. Una excepción sucede en el trabajo de Golbeck y Hansen (2011) en la que parten del conjunto de congresistas a modo de semilla y lo expanden a través de sus seguidores para intentar determinar si existe algún enlace político entre los congresistas y los medios de comunicación usando a dichos seguidores como enlace. Sin embargo, no es un método general y mucho menos dinámico.

Creemos que es necesario investigar nuevos métodos de obtención de datos de Twitter que mejoren la cobertura y sean capaces de adaptarse a la naturaleza dinámica de los mensajes de Twitter.

2. Estrategias iniciales

Inicialmente, se abordó el problema de la captura de datos de Twitter mediante una serie de aproximaciones básicas como parte del trabajo inicial exploratorio. Recordemos que estas aproximaciones son, en esencia, métodos para construir consultas sobre Twitter, pues es la forma en la que éste permite el acceso a sus datos.

2.1. Etiqueta central y lista estática de términos

La primera aproximación tomada durante la captura de *tweets* fue simple y directa. De manera intuitiva, se modela la existencia de un término "central" que representa la temática en cuestión a ser explorada, siendo este término muy frecuente y utilizado por la comunidad para dicha temática. En Twitter es común la utilización de etiquetas o *hashtags*, términos ligeramente desambiguados que hacen referencia a algún tema determinado. Estos términos comienzan por el carácter `'#'`.

La temática objetivo utilizada en este trabajo son las elecciones generales españolas del 20 de noviembre de 2011. Para ello se decidió escoger como término central la etiqueta `#20N`, siendo esta la etiqueta recurrente para hablar de las elecciones en Twitter.

Esta técnica *naïve* nos sirve como primera aproximación al problema de explorar el espacio de datos y como punto central en dicho espacio. Sin embargo, a pesar de su prolífico uso en Twitter, la cobertura es insuficiente. Si nos quedamos con solo los datos obtenidos de esa consulta, podemos perder información altamente relacionada con las elecciones que no está especificada de manera directa en los *tweets* de esta etiqueta. Por ejemplo, es posible que perdamos mensajes importantes de los candidatos o de los partidos políticos.

Una opción para solventar el problema de la cobertura sería la generación de una lista de términos adicionales relacionados con la temática en lugar de un solo término central, pero la cobertura dependería exclusivamente del criterio del experto que genera la lista. Además no se contemplarían posibles términos que harían referencia a temas relacionados importantes que pueden aparecer a lo largo del tiempo. En la siguiente sección planteamos un método que pretende solucionar este problema.

2.2. Listas de términos generadas dinámicamente

La generación de una lista estática no responde bien ante la naturaleza temporal de Twitter. El uso de los diferentes términos y etiquetas varía significativamente a lo largo del tiempo, reduciendo la capacidad de cobertura de una lista estática a lo largo de una ventana de tiempo relativamente grande. Sin embargo, lo mas grave es la incapacidad de adaptarse para capturar eventos impredecibles que afectan a la temática explorada en cuestión.

Para ilustrar esta situación y tomando como ejemplo la temática de las elecciones generales, supongamos que el gobierno actual decide instaurar una ley polémica o reacciona ante algún evento de forma igualmente polémica, lo que puede terminar afectando significativamente a la opinión publica sobre las futuras elecciones. Es muy probable que

surgiera una nueva etiqueta durante un tiempo que representara ese evento, por lo que los *tweets* usarán esa etiqueta para referirse a ese evento.

Estos eventos de alta relevancia pero relativa corta vida son muy importantes en Twitter y pueden ser de distinta naturaleza como debates, desastres naturales, ataques terroristas, crisis económica o acciones políticas.

Una estrategia para adaptarse al carácter temporal de Twitter es la de generar las listas de términos dinámicamente, analizando los resultados tal como se van capturando y pudiendo determinar qué términos son los más relevantes en cada momento. Para el problema que se describe en este artículo, la técnica de *bootstrapping* nos permite ir actualizando el conjunto de términos de búsqueda de forma regular e iterativamente sobre sí mismo. Cada cierto periodo estipulado, se realiza un cómputo del nuevo conjunto de términos de búsqueda a partir de los datos capturados anteriormente durante una ventana de tiempo determinada. A la técnica descrita le falta un componente fundamental: el algoritmo o método que se aplica a cada iteración para determinar qué términos son más relevantes en cada momento.

2.2.1. Selección usando heurísticas basadas en la frecuencia

Para poder seleccionar los términos más relevantes en cada momento, una de las ideas mas simples que surgen es la de utilizar heurísticas basadas en frecuencias, partiendo de la etiqueta central como semilla para la recogida de *tweets* y durante una ventana de tiempo razonable.

Se realizó un análisis que consistió en la elaboración de listas de términos ordenadas por relevancia, calculada mediante distintas heurísticas basadas en frecuencia: *log-verosimilitud*, la *información mutua puntual* y el *test exacto de Fisher*. Se realizó una limpieza de los datos de entrada para eliminar elementos no deseados como *palabras huecas*. El texto fue procesado para dejar sólo nombres, adjetivos y verbos, además de los constructos sintácticos específicos de Twitter tales como las menciones(`@`) y las etiquetas (`#`).

Se obtuvieron listas de unigramas, bigramas y trigramas (con una ventana de tamaño 4 para bigramas y trigramas) ordenadas según las heurísticas anteriores, comprobándose los 100 primeros elementos de cada lista. La lista de unigramas fue bastante satisfactoria, encontrando en ella muchos elementos relacionados con las elecciones que, a priori, no se hubieran obtenido a través de una elaboración manual. Aunque de la lista de bigramas se obtuvieron asociaciones entre términos muy interesantes y significativas con las elecciones como pueden ser $\langle gobierno, crisis \rangle$, $\langle PP, recortes \rangle$ o $\langle cambio, gobierno \rangle$, la lista se encontraba llena de bigramas que contenían algún término de la lista de unigramas o que eran de escasa utilidad debido al ruido. La lista de trigramas no aportó información realmente interesante.

Simplemente tomando los términos más relevantes de lista de unigramas (en este caso los 100 primeros) como una consulta combinada disyuntiva en Twitter, podemos capturar *tweets* relacionados con el tema representado por `#20N` (elecciones generales) con una gran cobertura y de forma automática.

2.2.2. Selección usando medidas de relevancia en grafos

El método anterior tiene ciertas desventajas, siendo la mas importante el no aprovechar la estructura de grafo que exhibe Twitter, tratando a las etiquetas y menciones como simples términos. Es más, un análisis detallado de las listas reveló que la mayoría de los términos relevantes son etiquetas o menciones a usuarios importantes, por lo qué un análisis mediante algoritmos de grafos puede dar mejores resultados. Por ello, surge la necesidad de utilizar otras estrategias que analicen la estructura de grafos de Twitter, tomando como semilla o punto de referencia la etiqueta central.

Dado que Twitter es en realidad un grafo inmenso con varios tipos de relaciones, usaremos *PageRank* como herramienta de análisis de enlaces para procesar dichas relaciones. *PageRank* nos permite obtener un *ranking* de nodos de un grafo en función de su relevancia dentro del mismo.

El grafo sobre el que se computa *PageRank* se construye a partir de los datos capturados de la ventana de tiempo especificada, generando nodos para las etiquetas y los usuarios (`#` y `@`). La aristas del grafo entre nodos de usuarios y etiquetas representan lo siguiente:

- $Usuario \longrightarrow Usuario$: menciones
- $Usuario \longrightarrow Etiqueta$: uso de etiqueta
- $Etiqueta \longleftrightarrow Etiqueta$: coocurrencia

El grafo resultante tiene aristas tanto dirigidas como no dirigidas y el peso de cada una de esas aristas viene determinado por el volumen de *tweets* de la situación representada. Con esta ponderación que hace uso de las relaciones entre usuarios y etiquetas, se aplica *PageRank* sobre el grafo ponderado y se obtiene el ranking de nodos más relevantes en la red, siendo cada nodo un término que puede ser tanto etiqueta como usuario.

La ventaja que tiene el análisis de grafos frente el análisis estadístico es que al analizar las relaciones entre los nodos podemos descubrir elementos de alta relevancia pero de bajo volumen de ocurrencias que un análisis basado en frecuencia no captaría. Este método detecta elementos como `#anguitaporcórdoba` que tienen poder mediático pero no se incluirían si tenemos en consideración solamente el volumen; se incluyen porque al aplicar *PageRank*, estos coocurren o son mencionados por elementos que tienen una relevancia directa mayor. En pocas palabras, se favorecen los elementos que son referenciados por otros elementos ya relevantes.

Al aplicar el método a intervalos de tiempo, utilizando en cada iteración el conjunto completo de tweets obtenidos en el paso anterior, se observa un comportamiento no deseado: el conjunto de términos diverge demasiado, separándose de la semilla inicial cada vez más.

Para ilustrar este efecto, se muestran los 10 primeros términos de las primeras cuatro iteraciones del método, ejecutadas cada 60 minutos y usando como semilla `#20N`:

1. #20n #15m @metroscopia @marianorajoy #15o #globalchange #occupyspain @llamamatrimonio @occupyspain @conrubalcaba
2. #fb @upyd @conrubalcaba #20n @anapastor_tve #15m #t #15o #pp @marianorajoy
3. #fb #nowplaying #thaiflood #sumatealrosa #nf #20n @conrubalcaba @1dupdates #t #15o
4. #nowplaying #fb #sumatealrosa #nf #thaiflood #in #facebook @gllamazares @no_al_cancer_ @marianorajoy

Se observa que cada iteración se desvía aún mas de la temática a explorar, pues elementos como `#nowplaying` o `#thaiflood` no están relacionados con lo que queremos capturar. Esto se debe a diversas causas:

- En las listas de términos aparecen usuarios que además de hablar de la temática central (elecciones/política), pueden comentar sobre otras temáticas.
- Algunos temas de los que hablan los usuarios explorados son temas muy populares en ese momento: `#thaiflood`, `#cancerdemama` o `#spotify`. Dichos temas acumulan muchas menciones de manera temporal, entrando en la lista de términos mas relevantes mediante puro volumen.
- Una vez que un término no relacionado con la temática central gana una posición en la lista de los mas relevantes, se capturan los *tweets* relacionados con ese término provocando una "inundación" de información no relacionada con la política.

Para solucionar este problema, se propone una pequeña variante del método en la cual se evita que la lista diverja con respecto a la semilla central. En lugar de utilizar todos los *tweets* recuperados en el último paso para la construcción del grafo, se utilizan sólo los *tweets* que contienen al término semilla (en nuestro caso, `#20N`). Para una mejor comprensión, la figura 1 muestra un diagrama intuitivo del método finalmente propuesto.

Con este método se consigue, con un alto grado de satisfacción, solventar la problemática existente sobre la generación de consultas para la captura de datos en Twitter, todo ello de forma iterativa, explotando la topología de grafo que proporciona Twitter y controlando la inducción y cobertura de los términos.

3. Recurso generado

Realizando una captura en tiempo real de los *tweets* mediante el API de streaming que Twitter ofrece y usando el método de generación de consultas expuesto en la sección 2.2.2 se generó un recurso que contiene información relevante a las actualizaciones de estado o *status updates* de ese periodo. Un *status update* es una actualización del estado del usuario ya sea mediante *tweets*, *retweets* o respuestas (*replies*).

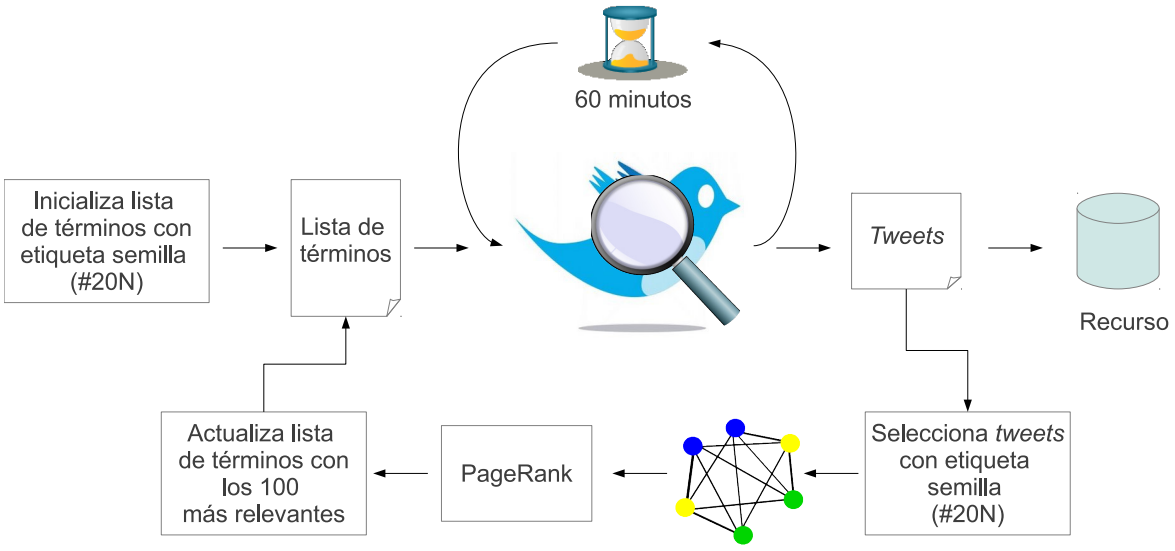

Figura 1: Diagrama del método finalmente propuesto

Campo	Descripción
truncated	Indica si el texto fue truncado por exceso de caracteres
retweetCount	Indica el nº de *retweets* respecto al original
id	Identificador numérico del tweet
createdAt	Fecha y hora de creación

Tabla 1: Atributos del elemento `<status>`

3.1. Descripción del recurso

El recurso es una colección de archivos XML , cada uno de los cuales contiene los *tweets* capturados durante una hora, a lo largo de todo el periodo de las elecciones españolas de 2011 (precampaña, campaña, día de las elecciones y el día después de las elecciones), comprendiendo del 21 de octubre al 21 de noviembre.

Cada fichero posee la siguiente estructura general dividida en secciones:

- `<filterQuery>`: Consulta de filtrado usada (lista de términos).
- `<statuses>`: Colección de elementos `<status>` capturados a lo largo del periodo.

El elemento `<status>` representa una actualización de estado. La lista de atributos del elemento `<status>` se muestra en la tabla 1 y la lista de subelementos se muestra en la tabla 2.

Sub-elemento	Descripción
inReplyToStatusId	Identificador del *status* al que este *status* hace una respuesta
inReplyToUserId	Identificador del usuario al que este *status* hace una respuesta
hashtags	Hashtags referenciados
source	Medio de origen
text	Texto procesado final
urlEntities	URL que aparecen
user	Usuario que realiza la actualización de estado
userMentions	Menciones a otros usuarios que aparecen

Tabla 2: Descripción de los subelementos del elemento `<status>`

En esencia, el recurso generado conforma un corpus de Twitter que engloba el periodo de las elecciones. En la tabla 3 se muestra una serie de métricas de interés sobre las características más generales del recurso, siendo éstas principalmente volumétricas. En la gráfica 2 se muestra de forma más intuitiva la distribución global de los *updates* según su naturaleza básica: *retweets*, *replies* (sean especificas a un *update* concreto o no) o "*tweets* sencillos".

3.2. Utilidad del recurso

A continuación se muestran a modo de ejemplo una serie de resultados de análisis para dar ejemplo de la utilidad y la capacidad de

Métrica	Valor
# de *users*	1.587.930
# de *hashtags*	195.985
# de *updates*	5.724.612
# de *retweets*	1.362.785
# de *replies*	202.778
# de *status replies*	683.305

Tabla 3: Métricas generales del recurso

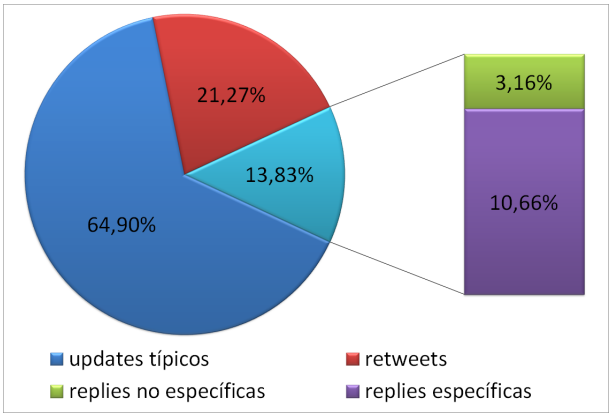

Figura 2: Distribución del volumen total de *updates* según su naturaleza.

explotación del mismo.

Una de las formas más simples y directas de analizar el recurso es mediante la comparación de las diferentes listas de términos utilizadas a lo largo del periodo. De esta forma, de un simple vistazo, podemos observar cuáles son los términos mas relevantes para la comunidad twittera, correspondientes a las elecciones.

Hemos escogido tres momentos interesantes inicio de la precampaña, el debate de los candidatos del PSOE y el PP a presidente y el día de las elecciones. La tabla 4 muestra los 10 términos mas relevantes de estos tres puntos temporales.

La visualización expuesta en la figura 3 representa el grafo generado durante todo el debate de los candidatos a presidente tal y como se especifica en la sección 2.2.2. Se han etiquetado los nodos más relevantes, cada nodo está coloreado y escalado en función de su puntuación de *PageRank* y las aristas están coloreadas según el promedio de los nodos que une.

Twitter posee una naturaleza claramente temporal que puede ser explotada a través del recurso. Para ilustrar un caso de análisis temporal, tomaremos la gráfica de la figura 4 correspondiente al debate de los candidatos a Presidente del Gobierno por parte de los partidos PSOE y PP. Esta gráfica muestra la evolución temporal de la relevancia de los usuarios `@conrubalcaba` y `@marianorajoy` y las etiquetas `#rubalcaba`, `#debate` y `#rajoy`, donde las líneas finas muestran el valor real mientras que las líneas gruesas muestran la tendencia utilizando una media móvil de 6 valores.

Una de las primeras cosas que podemos observar es la correcta detección de la etiqueta `#debate` y su relevancia debido a la aplicación del método expuesto basado en *PageRank*, siendo ésta relevante solo en ese periodo y llegando a superar al resto de términos durante el propio debate. Se detecta el auge de etiquetas como `#rajoy` o `#rubalcaba` que experimentan un claro aumento durante el debate y exhiben una interesante correlación con respecto a `#debate`, mientras que los términos `@conrubalcaba` y `@marianorajoy` no sufren alteración alguna y mantienen su relevancia relativamente estable a lo largo del periodo.

Como último apunte, se observa que al final del periodo observado los términos referentes al candidato Mariano Rajoy (`@marianorajoy` y `#rajoy`) quedan ligeramente por encima de los términos referentes al candidato Rubalcaba (`@conrubalcaba` y `#rubalcaba`), tomando como referencia el valor absoluto (línea fina) sobre la tendencia, coincidiendo con los resultados de las encuestas sobre el ganador del debate.

Hay que tener en cuenta que ambas gráficas mostradas están en escala logarítmica para una mejor observación y presentación.

4. *Conclusiones*

El método propuesto permite recuperar los *tweets* relacionados con un tema determinado, incluyendo automáticamente en las consultas utilizadas las nuevas etiquetas y usuarios relevantes que vayan apareciendo en el periodo en el que se realice la captura. Se trata de una mejor solución a la empleada en otros trabajos recientes relacionados con Twitter, especialmente en términos de cobertura y adaptación a posibles eventos novedosos relacionados.

El recurso obtenido está disponible para su utilización pública[1]. Actualmente, esta-

[1] `http://www.lsi.us.es/~fermin/index.php/Datasets`

Principio	Debate	Elecciones
#20n	#20n	#20n
@crparlamentaria	#debate	#elecciones20n
#15m	#eldebate	#elecciones
#nolesvotes	#caraacara	@antoniofraguas
@psoe	@conrubalcaba	@kurioso
@elconfidencial	#reiniciaeldebate	#votar
@marta_llorens	@marianorajoy	@el_pais
@gad3_com	@ramonlobo	#mesas20n
@ppopular	#seacaboelcirco	@marianorajoy
#pp	@otrodiademierda	@la_ser

Tabla 4: Términos relevantes en tres instantes temporales

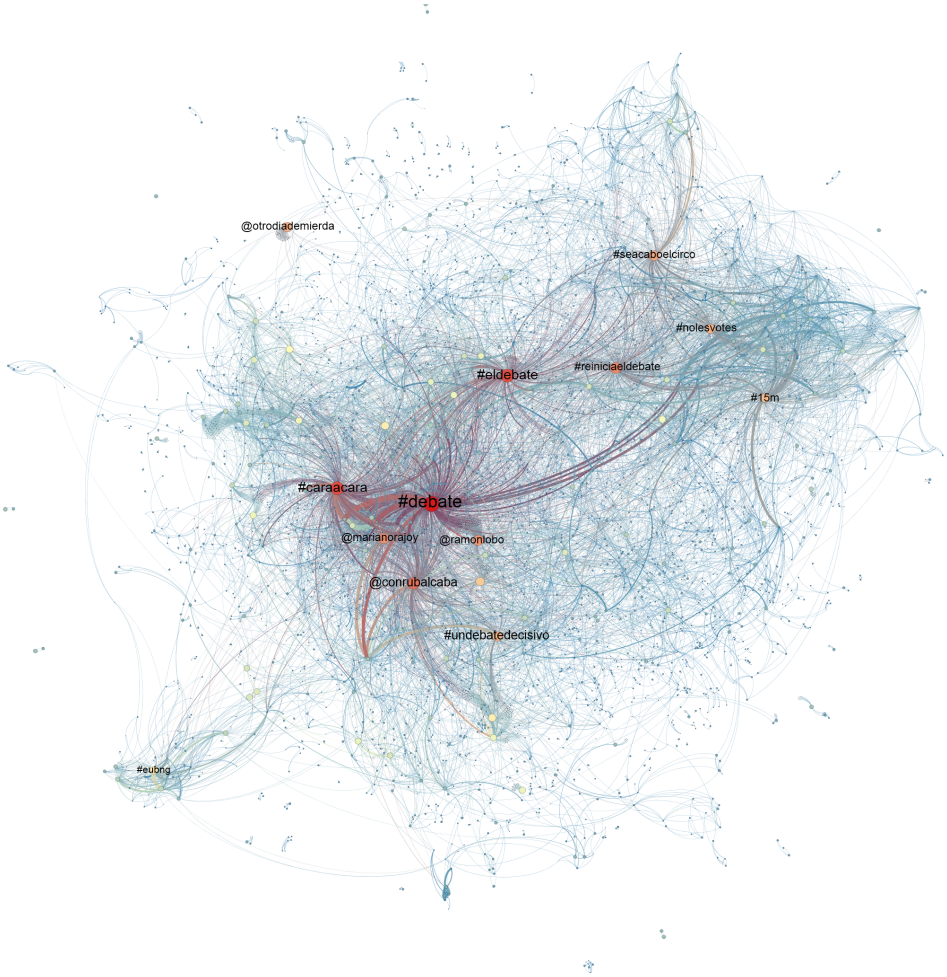

Figura 3: Visualización del grafo correspondiente al debate

mos empleando el recurso para llevar a cabos experimentos de análisis de opinión y detección de comunidades.

Bibliografía

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, y Rebecca Passonneau. 2011. Sentiment analysis of twitter data. En *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, páginas 30–38, Portland, Oregon, Junio. Association for Computational Linguistics.

Congosto, M. Luz, Montse Fernández, y Esteban Moro Egido. 2011. Twitter y política: Información, opinión y ¿predicción? *Cuadernos de Comunicación Evoca*, 4.

Davidov, Dmitry, Oren Tsur, y Ari Rappoport. 2010a. Enhanced sentiment learning using twitter hashtags and smileys. En *Proceedings of the 23rd*

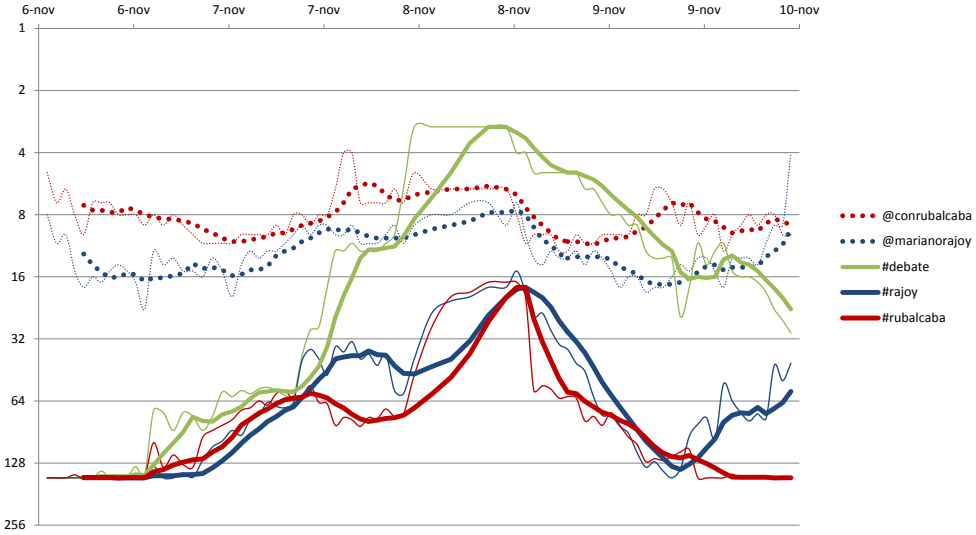

Figura 4: Evolución temporal de los términos `@conrubalcaba`, `@marianorajoy`, `#debate`, `#rajoy` y `#rubalcaba` antes y después del debate de los candidatos. Escala logarítmica en base 2 ($y' = log_2(y)$)

International Conference on Computational Linguistics: Posters, COLING '10, páginas 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Davidov, Dmitry, Oren Tsur, y Ari Rappoport. 2010b. Semi-supervised recognition of sarcastic sentences in twitter and amazon. En *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, páginas 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Go, Alec, Richa Bhayani, y Lei Huang. 2009. Twitter sentiment classification using distant supervision. En *Processing*.

Golbeck, Jennifer y Derek Hansen. 2011. Computing political preference among twitter followers. En *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, páginas 1105–1108, New York, NY, USA. ACM.

Hong, Sounman y Daniel Nadler. 2011. Does the early bird move the polls?: the use of the social media tool 'twitter' by u.s. politicians and its impact on public opinion. En *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, dg.o '11, páginas 182–186, New York, NY, USA. ACM.

Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, y Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, páginas 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kim, E, S Gilbert, M J Edwards, y E Graeff. 2009. Detecting sadness in 140 characters: Sentiment analysis of mourning michael jackson on twitter.

Pak, Alexander y Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Pennacchiotti, Marco y Ana M. Popescu. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. En *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, páginas 430–438, New York, NY, USA. ACM.

Silva, Ismael Santana, Janaína Gomide, Adriano Veloso, Wagner Meira, Jr., y Renato Ferreira. 2011. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. En *Proceedings of the 34th international ACM SIsGIR conference on Research and development in Information*, SIGIR '11, páginas 475–484, New York, NY, USA. ACM.

Tan, Chenhao, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, y Ping Li. 2011. User-level sentiment analysis incorporating social networks. En *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, páginas 1397–1405, New York, NY, USA. ACM.

Tumasjan, Andranik, Timm O Sprenger, Philipp G Sandner, y Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Word Journal Of The Intenational Linguistic Association*, páginas 178–185.