UdeAnalytics: Encontrando comunidades cerradas en Twitter

Herrera C.[1], Marulanda J.P.[2], Roquemen V.[3], Silva D.[4]
[1−3] *Grupo de Óptica y Fotónica,* [2] *Grupo de Cosmología y Gravitación*
[4] *Grupo de Instrumentación y Microelectrónica,* [4] *Grupo de Estado Sólido*
Instituto de Física, Universidad de Antioquia, A.A. 1226, Medellín, Colombia.

11/12/2019

Resumen

En el presente proyecto se realizó un estudio con el cual se pretendió encontrar comunidades cerradas en Twitter. Se usaron como herramientas el API de Twitter, Tweepy e Infomap. Para el análisis se definió una métrica que contuviera la interacción entre los usuarios, basada en el número de *retweets*, respuestas, seguidores y seguidos. Por medio de la métrica se encontró no sólo una comunidad alrededor de un centroide, que para este estudio fue Daniel Quintero, candidato a la Alcaldía de Medellín para el periodo 2020-2023, sino también subcomunidades alrededor de usuarios importantes de la comunidad principal. Al revisar los usuarios que fueron los centroides de su comunidad cerrada, se confirmaron los hallazgos al ser estos los miembros más activos e importantes de la comunidad.

Palabras Clave: comunidades cerradas, Twitter, análisis de grafos, métrica, Infomap.

---

1. Introducción

Las redes sociales han proporcionado desde su aparición un nuevo medio donde personas de todo el mundo pueden interactuar entre sí. En plataformas como Twitter, donde la interacción principal entre los usuarios consiste en el intercambio de opiniones, resulta interesante estudiar cómo esto puede afectar la dinámica entre ellos dentro de la red. Bajo esta premisa, el estudio de la dinámica de posibles sub-comunidades dentro de una comunidad principal puede dar una idea de, por ejemplo, el panorama político en una región al tomar en cuenta si esas interacciones son positivas o negativas.

Una comunidad cerrada puede encontrarse maximizando la modularidad de la misma. La *ecuación de mapa* proporciona un método efectivo para dicho fin. Esta se basa en la dualidad que existe entre comprimir la información de un conjunto de datos y encontrar estructuras en ellos, lo cual direcciona el problema al uso de la teoría de la información [1]. Concretamente, partiendo de trayectorias aleatorias guiadas por los pesos de las conexiones en una red, se quiere asignar a los nodos de dicha red un código que describa las trayectorias, lo cual debe aprovechar las regularidades presentes en la red [2].

Para determinar entonces una modularidad apropiada en la red, se quiere que esta esté agrupada de tal forma que el código necesario para describir las trayectorias men-

cionadas anteriormente sea óptimo. En este código, que es una secuencia de números binarios, se encuentra recopilada la información por donde pasa la trayectoria [2]. La *ecuación de mapa* permite hacer justamente esto, sin tener que construir el código de forma explícita. El método para encontrar la modularidad óptima de la red consiste entonces en formar distintas particiones en la red, evaluando la *ecuación de mapa* con el objetivo de minimizarla. Así, los módulos resultantes de esta optimización se toman como las comunidades cerradas dentro de la red.

La *ecuación de mapa* requiere de una métrica apropiada para describir las relaciones de cercanía de los datos a ser tratados. Aquí se propone una variación a la métrica utilizada por Shaham en [3], donde damos prioridad a los seguidores y seguidos de un usuario. Además, aunque en una muestra de naturaleza política hay gran contenido de *tweets* con connotación negativa, siguen siendo interacciones que contribuyen a las posibles comunidades cerradas y no deben excluirse.

En este proyecto se pretende encontrar una comunidad cerrada alrededor de alguno de los tres candidatos más prometedores a la Alcaldía de Medellín para el período 2020-2023, estos son: Daniel Quintero, Luis Alfredo Ramos y Santiago Gómez.

2. Metodología

Los datos se obtienen usando la API de Twitter por medio de la librería Tweepy de Python [4].

En primer lugar se recogen *tweets* que contengan ciertos usuarios clave, durante cierto periodo de tiempo. Con los datos obtenidos se determina un centroide apropiado para el estudio, teniendo como criterio el usuario con más menciones. A partir de este, se recolectan los usuarios más activos en la comunidad de acuerdo a las menciones del centroide en sus *tweets*. Una vez obtenida la comunidad de usuarios, y excluyendo al centroide, se recolecta información de estos, como sus listas de seguidos y seguidores y los *tweets* que se hicieron durante el tiempo elegido.

Con la anterior información se construye una matriz de cercanía, la cual cuantifica la interacción de cada usuario con cualquier otro de la comunidad por medio de una métrica dada por:

$$EdgeW = W_1 \cdot RT + W_2 \cdot REP + W_3 \cdot FF \qquad (1)$$

donde W_1, W_2 y W_3 son coeficientes de peso que se definieron arbitrariamente, tal que $W_3 > W_2 > W_1$, RT es el promedio de *retweets* entre dos usuarios, REP es el promedio de respuestas entre los mismos, y FF se define como 0 si los usuarios no se siguen, 1 si sólo uno de ellos sigue al otro, o 2 si se siguen recíprocamente.

Además de la métrica, se define la importancia que tiene cada usuario como:

$$NodeW = \frac{NTweetsCentroide}{NTweets} \qquad (2)$$

donde $NTweetsCentroide$ es el número de *tweets* dirigidos al centroide sobre el cual se está estudiando la comunidad y $NTweets$ es el número total de *tweets* del usuario en cierto período de tiempo.

Con esta información se construye un grafo de la red donde cada usuario tiene, en cada una de sus conexiones, un peso determinado por la métrica.

Una vez obtenido el grafo, se usa el algoritmo de *Infomap*, de forma tal que se obtienen aglomeraciones de usuarios dentro de la red que, como se definió anteriormente, maximizan la modularidad de ésta y se consideran como comunidades cerradas.

3. Resultados y discusión

Se recolectó información entre el 18 y el 19 de octubre de 2019, usando como usuarios clave los candidatos a la alcaldía de Medellín ya mencionados. Por los resultados ilustrados en la figura 1 se concluyó que Daniel Quintero (@QuinteroCalle) era el candidato más apropiado para ser el centroide a estudiar. El grafo generado con Infomap y los resultados estadísticos obtenidos para los usuarios estudiados pueden verse en las figuras 2 y 3 respectivamente.

Se puede ver que el parámetro más representativo es el relacionado con que un usuario siga a otro, lo que justifica la jerarquía escogida para el tamaño de los pesos en la ecuación 1.

Se encontraron 23 aglomeraciones, denominadas subcomunidades, que contenían al menos dos usuarios y que

bajo las consideraciones del proyecto se toman como comunidades cerradas. Los 132 nodos restantes corresponden a usuarios con poca o nula interacción con la comunidad y se interpretaron como personas aisladas de ella.

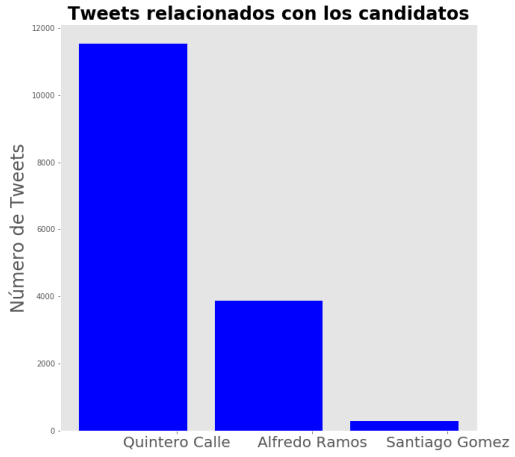

Figura 1: Cantidad de *tweets* donde se menciona a cada candidato.

Se nota también que el usuario con el mayor valor promedio de la métrica fue @J_Upegui, a pesar de no ser el centroide de la sub-comunidad más numerosa. Este usuario fue particularmente activo en las fechas de recolección de datos, pues todos sus *tweets* en ese rango de tiempo hicieron referencia a @QuinteroCalle.

4. Conclusiones

Del análisis de comunidades cerradas, pudieron encontrarse, en efecto, sub-comunidades de usuarios que, si bien su actividad estaba direccionada al centroide escogido, su interacción se veía contenida en un grupo bien definido. Además, se encontró que dentro de estas sub-comunidades había también un centroide, por medio del cual se daban la mayoría de interacciones entre los usuarios pertenecientes al grupo.

El resultado que arroja el algoritmo de modularización de la red depende en gran medida de la métrica que se defina para modelar la interacción entre los usuarios. La métrica definida para este problema, dada en la ecuación 1, resultó ser idónea para identificar esas comunidades ya que, al estudiar en detalle las sub-comunidades en-

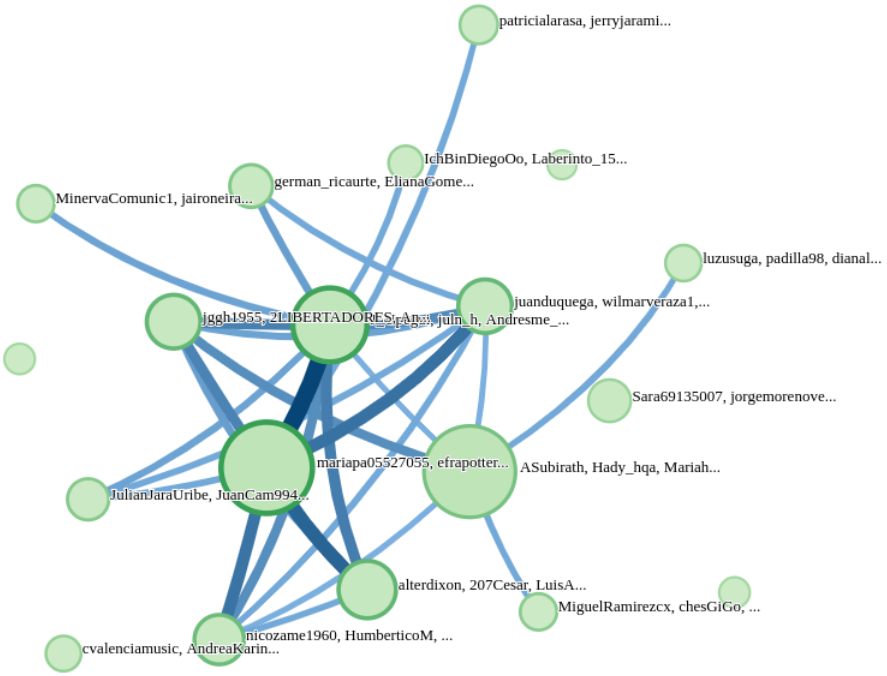

Figura 2: Grafo generado por el algoritmo de *Infomap*. Los nodos que se ven son comunidades cerradas que identificó el algoritmo y se componen de más nodos agrupados.

Figura 3: Datos estadísticos de la comunidad.

contradas, se confirmó que los miembros dentro de esta eran activos y, más importante, se evidenció la importancia del centroide secundario dentro de estos grupos.

Además, la métrica permitió identificar a aquellos usuarios que, si bien dentro de las fechas en que se recolectaron los datos mencionaron a @QuinteroCalle en sus *tweets*, no eran parte de la comunidad.

Referencias

[1] L. Bohlin, D. Edler, A. Lancichinetti, and M. Rosvall, "Community detection and visualization of networks with the map equation framework," in *Measuring Scholarly Impact*, pp. 3–34, Springer, 2014.

[2] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, 2009.

[3] Shaham, "Generating a twitter ego-network detecting communities.."

[4] A. Moujahid, "An introduction to text mining using twitter streaming api and python."