Procesamiento Eficiente de Grafos Masivos para Aplicaciones en Redes Sociales

Andrés Giordano[1], Gabriel H. Tolosa[1], Santiago Banchero[1]
Juan M. Ortiz de Zarate[2], Esteban Feuerstein[2]
{agiordano, tolosoft, sbanchero}@unlu.edu.ar; {jmoz, efeurest}@dc.uba.ar

[1]Departamento de Ciencias Básicas, Universidad Nacional de Luján
[2]Departamento de Computación, FCEyN, Universidad de Buenos Aires

Resumen

Las redes sociales digitales se han convertido sin dudas en una de las aplicaciones más populares de Internet y atraen a millones de usuarios que, de forma implícita, generan estructuras con propiedades emergentes que surgen del comportamiento global.

Existen diversos problemas interesantes a resolver como la formación de comunidades, la recomendación de enlaces y el estudio de la polarización de opiniones. En todos los casos, resulta motivador tanto el proceso de formación como el estudio de algoritmos eficientes para el procesamiento. Estos problemas se pueden abordar estudiando el grafo subyacente, el contenido de las publicaciones o combinaciones de ambos.

En este trabajo se proponen diversas líneas de investigación sobre los temas mencionados, con aplicaciones a grafos masivos y problemas reales. Se abordan tanto problemas algorítmicos en cuanto a la eficiencia como las interacciones entre usuarios y diferentes escenarios.

Palabras clave: Red social, algoritmos eficientes, comunidades, polarización, recomendación.

Contexto

Esta presentación se enmarca en los proyectos de investigación "Algoritmos Eficientes y Minería Web para Recuperación de Información a Gran Escala", "Estudio epidemiológico y seguimiento de pacientes con enfermedad celiaca (EC) asistidos por software y técnicas avanzadas de análisis de datos" del Depto. de Cs. Básicas (UNLu) y "Modelos y herramientas algorítmicas avanzadas para redes y datos masivos" del Depto. de Computación de la FCEyN (UBA).

Introducción

Las redes sociales digitales se han convertido sin dudas en una de las aplicaciones más populares de Internet y han modificado la forma en que los usuarios interactúan e intercambian información. Estas redes atraen a millones de usuarios [4, 8, 10] que, de forma implícita, generan estructuras con propiedades emergentes [1] que surgen del comportamiento global.

En general, este tipo de redes tienen a nivel estructural una topología libre de escala (*scale-free*), que se caracteriza por una distribución muy sesgada en el grado de los nodos. Además, son autosimilares, por lo que es posible estudiar porciones de la red y extraer propiedades generales. En este escenario, es posible diseñar algoritmos eficientes para compartir y distribuir la información generada. Además, es especialmente interesante si se tiene en cuenta que la red es un ambiente altamente dinámico y de gran escala.

Sin embargo, hay que considerar que estas redes son procesos humanos y no meramente tecnológicos, por lo que además de la estructura subyacente, también es de importancia en contenido generado. Su mejor comprensión posibilita aprovechar la inteligencia colectiva para mejorar servicios (por ejemplo, sistemas de recomendación) y aplicar a nuevos escenarios.

Existen diversos problemas altamente interesantes sobre una red social. Uno de estos es la formación de comunidades, es decir, grupos de usuarios que se agrupan por algún criterio. En este caso, resulta motivador no solo el proceso de formación y sus implicancias sino, además, el estudio de algoritmos eficientes para la conformación de comunidades [24].

Esta es una tarea desafiante a gran escala en aspectos que van desde el tamaño y el tipo de interacción hasta las similitudes por contenido. Si bien existen diversos métodos para analizar y modelar este tipos de redes, la necesidad de algoritmos que combinen información estructural con las propiedades de los nodos es un requisito para un amplio espectro de potenciales aplicaciones concretas. Algunos de estos problemas tienen aplicación potencial en proyectos de colaboración abierta, en la salud (grupos de personas con patologías similares) o en el caso de catástrofes [20, 2].

Otra cuestión de interés actual en redes sociales es el estudio de la polarización de opiniones. Se ha visto ampliamente en elecciones presidenciales pero existe aplicado a muchos otros casos de los cuales no se conoce aún si su comportamiento es similar y, en caso que no lo sea, qué modelos aparecen[1].

En este trabajo se proponen diversas líneas de investigación sobre los temas mencionados, con aplicaciones a grafos masivos y problemas reales. Se abordan tanto problemas algorítmicos en cuanto a la eficiencia como las interacciones entre usuarios y la formación de comunidades en diferentes escenarios.

Líneas de I+D

En este proyecto se continúan líneas de I+D del grupo que incorporan análisis de grafos de redes sociales que permitan mejorar la calidad de algoritmos básicos y la utilidad de métricas características de estas estructuras de datos aplicadas a éstos. Se abordan problemas clásicos sobre grafos pero aplicados a gran escala. Además, se abordan problemas aplicados a situaciones concretas. En especial, las líneas de I+D principales son:

a. Estimación en Grafos Masivos

La estructura de un grafo G se define como $G = \{E, V\}$ donde V es el conjunto de nodos o vértices y E es el conjunto de aristas que los unen. La posibilidad de computar algunas métricas básicas, como por ejemplo, el cálculo de la distancia entre dos nodos arbitrarios, se ve afectada en cuanto a la eficiencia del proceso debido al tamaño de los grafos actuales. Por ejemplo, en redes sociales digitales el número de relaciones (aristas) supera ampliamente la cantidad de usuarios (nodos)[2].

El problema de la distancia entre dos nodos tiene múltiple aplicaciones prácticas, por ejemplo, para el ranking en búsquedas[3]. Se lo define como la longitud del camino mas corto entre ellos y se vuelve casi inviable si se requiere responder en pocos milisegundos. Esta métrica es usada en numerosos algoritmos que apuntan a resolver problemas como la recomendación de links [25], agrupamiento de usuarios [5], entre otros.

El cómputo exacto de la distancia es prohibitivo para aplicaciones prácticas dado el tamaño de estas estructuras y la estimación del valor es una alternativa. La reducción del error asociado a estimaciones de estos tipos de métricas a un costo computacional bajo conforma una de las líneas de investigación de este proyecto. En nuestro enfoque se utiliza un conjunto de nodos, llamados *landmarks* [17], que se toman como referencia para estimar luego la distancia entre dos nodos arbitrarios. El problema de la selección de "buenos" landmarks, es decir, aquellos que permitan minimizar el error de estimación, es una pregunta abierta ya que existen diversos criterios a aplicar que consideran grafos de diferente tamaño, densidad y dinámica.

b. Formación de Comunidades

Si bien no existe una definición exacta de *comunidad* y una técnica especifica para identificarlas, se puede decir que una comunidad esta compuesta por usuarios que comparten alguna característica en común. El interés en tener métodos que realicen esta tarea de forma efectiva surge de varias áreas como la política [14], medicina [22], etc. A través de comunidades detectadas se puede estudiar el comportamiento e interés en ciertos temas de las personas. En las redes sociales estas comunidades se pueden detectar a través de 3 técnicas básicas:

- **Análisis de la topología de la red:** este método es el mas simple y se basa solo en el grafo subyacente a la red, es decir, los usuarios y sus relaciones [18, 3, 12]. Si bien los algoritmos que aplican este enfoque son eficaces

[1] Las hinchadas de fútbol y el referéndum por la independencia de Catalunia son otros dos ejemplo que generan estructuras de opiniones polarizadas.

[2] Twitter: una imagen de esta red social contiene 81306 nodos con 1768149 aristas (`https://snap.stanford.edu/data/egonets-Twitter.html`)

[3] Un caso es la red de contactos profesionales Linkedin.

suelen agrupar usuarios que tratan de tópicos diferentes aunque densamente conectadas (carecen de alta precisión).

- **Análisis del contenido:** este enfoque explora el contenido de las publicaciones de los usuarios y no considera la información estructural de la red como lo es la densidad de las conexiones que puede existir en un conjunto de usuarios. En Twitter, por ejemplo, esto se refiere al contenido de los tweets separando texto libre de hashtags, urls y menciones [13].
- **Híbridos:** estos métodos utilizan los dos enfoque anteriores en conjunto [19, 9, 26, 23], agregan características como la similitud de contenido aplicada como peso o importancia de la relación entre un par de usuarios. Una vez generada esta estructura se aplica algún algoritmo de detección de comunidades conocido que use el peso de las aristas.

En este tema el grupo viene trabajando con un estudio puntual relacionado con el proyecto interdisciplinario mencionado en la sección Contexto, en la cual no solamente se realizan estudios médico/biológicos/químicos sino que, además, se trabaja con el tema sobre la red social Twitter. En particular, se aplican algoritmos de formación de comunidades utilizando los tres enfoques mencionados, detectando redes de usuarios con interés en la enfermedad celíaca y poder establecer autoridades en el tema y usuarios influyentes.

c. Estudio de "Jergas" y Polarización en Contenido

En esta línea de trabajo se enfoca el problema de las comunidades a partir de dos ideas diferentes a las anteriormente presentadas: uso de determinada jerga en particular y búsqueda de opiniones polarizadas (que permite identificar dos comunidades). Si bien muchos trabajos [15, 11] usan Twitter como laboratorio y a la política como contexto, no se ha enfocado el problema en la identificación de estas mediante únicamente la jerga expuesta en sus 140 (o ahora 280) caracteres. En base a la hipótesis de que como las comunidades políticas son fuertemente homofílicas y además teniendo en cuenta como esto es potenciado por las *Filter Bubbles*[16] de las redes sociales, se genera en torno a estas una jerga particular, a través de la cual podemos identificar la pertenencia del usuario a una comunidad. Es decir, en base a como "habla", se pretende predecir a que comunidad política pertenece.

Resultados preliminares muestran que las estructuras de los grafos generados por los distintos conjuntos de tweets presentan patrones en común que abren las puertas a nuevas preguntas que pueden ser abordadas de manera interdisciplinaria entre la computación y las ciencias sociales.

d. Clustering

Las técnicas de clustering permiten agrupar items con características similares. Existen diversas técnicas que se aplican a casos puntuales con diverso grado de eficacia. Esta línea complementa uno de los puntos anteriores ya que se puede aplicar para la formación de comunidades donde las observaciones u objetos a agrupar son los usuarios y las dimensiones de los datos están dadas por las características de los mismos extraídas de la topología y/o el contenido publicado. Un cluster detectado será considerado comunidad pero no al revés, ya que un conjunto de usuarios que conformen una comunidad no necesariamente serán agrupados por estas técnicas. En concreto se intenta establecer las características y los algoritmos a utilizar para que la detección sea tanto eficaz como eficiente por tratarse de grafos masivos.

e. Recomendacion de Enlaces

La recomendación de enlaces propone buscar y sugerir a ciertos usuarios aquellos links no establecidos que tienen alguna probabilidad de ocurrir en un futuro (por algún criterio). Aplicaciones ampliamente conocidas, como Facebook y Twitter, aplican estas técnicas de forma sistemática. Algunos trabajos proponen primero la detección de comunidades [21, 6] para luego hacer recomendación de links. Esta linea de investigación propone mejorar la calidad y eficiencia de estas recomendaciones a través del análisis de las redes aplicando algoritmos de detección de comunidades, análisis de sentimiento, métricas sobre estructuras derivadas del contenido o híbridos de estas técnicas, así como también analizar el cambio estructural y de flujo de contenidos que genera la aceptación de sugerencias en estas redes.

f. Caracterización de Objetos

Las redes sociales no solo permiten la publicación de texto sino también de imágenes y videos. Además, a estas publicaciones se le agregan metadatos como la geo-localización, fecha de creación y datos del usuario como su lugar de residencia, fecha de nacimiento, etc. Algunos trabajos proponen usar el contenido, los metadatos y la estructura de la red social para detectar objetos o eventos y caracterizarlos en base al contenido compartido por los usuarios. Hotchman et. al [7] logra encontrar ciertos patrones en las métricas de las imágenes (matiz, brillo promedio, etc) tomadas en diferentes ciudades. Este tema resulta de gran interés para el grupo y forma parte de nuestras lineas de investigación. Por ejemplo, se puede extraer información sobre ciudades y sus habitantes (aspectos culturales y sociales) a partir de analizar publicaciones de fotografías.

Resultados y Objetivos

El objetivo principal de la propuesta es estudiar, desarrollar, aplicar, validar y transferir modelos, algoritmos y técnicas que operen eficaz y eficientemente sobre grafos masivos (redes sociales digitales). Se propone profundizar sobre el estado del arte y definir, analizar y evaluar nuevos enfoques incorporando las técnicas de minería de datos. Específicamente,

- Definir nuevas estrategias de selección de nodos landmarks para la estimación de distancia entre nodos cuyo cómputo sea rápido y requiera poco espacio de información extra. Una cuestión a tener en cuenta es poder relacionar diferentes métricas con propiedades estructurales de diferentes grafos y utilizar las apropiadas en cada caso.
- Diseñar técnicas de formación de comunidades que apliquen eficaz y eficientemente a problemas concretos (por ejemplo, la comunidad de celíacos, como se mencionó), a través de enfoques híbridos principalmente.
- Definir modelos de detección de comunidades con opiniones polarizadas a partir del estudio de las jergas que se establecen alrededor de un tema o de un conjunto particular de usuarios.
- Estudiar y adaptar los algoritmos de clustering al problema anterior, con énfasis en la performance para escalar a grafos masivos.
- Estudiar y adaptar técnicas de recomendación de enlaces y analizar el comportamiento estructural de una red bajo ciertas configuraciones de aceptación de las recomendaciones.
- Comenzar a explorar el problema de caracterización de ciudades/ciudadanos con técnicas basadas en redes sociales de imágenes. El énfasis inicial es realizar una prueba de concepto con ciudades de Argentina.

Formación de Recursos Humanos

Este proyecto brinda un marco para que algunos docentes auxiliares y estudiantes lleven a cabo tareas de investigación y se desarrollen en el ámbito académico. Actualmente, se están dirigiendo cuatro trabajos finales correspondientes a la Lic. en Sistemas de Información (UNLu) y un estudiante de doctorado del Departamento de Computación, FCEyN (UBA). Además, hay dos pasantes alumnos y un becario CIN (Beca de Estímulo a las Vocaciones Científicas). Se espera dirigir al menos dos estudiantes más por año y presentar dos candidatos a becas de investigación.

Referencias

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74, 2002.

[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, A. Gionis, and S. Leonardi. Online team formation in social networks. In *Proc. of the 21st International Conference on World Wide Web*, WWW '12, New York, NY, USA, 2012. ACM.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[4] D. M. Boyd and N. B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 2007.

[5] J. Edachery, A. Sen, and F. J. Brandenburg. Graph clustering using distance-k cliques. In J. Kratochvíyl, editor, *Graph Drawing*, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[6] D. Feltoni Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti. Enhancing social recommendation

with sentiment communities. In J. Wang, W. Cellary, D. Wang, H. Wang, S.-C. Chen, T. Li, and Y. Zhang, editors, *Web Information Systems Engineering – WISE 2015*, pages 308–315, Cham, 2015. Springer International Publishing.

[7] N. Hochman and L. Manovich. Zooming into an instagram city: Reading the local through social media. *First Monday*, 18(7), 2013.

[8] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, New York, NY, USA, 2007. ACM.

[9] M. N. Kewalramani. *COMMUNITY DETECTION IN TWITTER.pdf*. PhD thesis, University of Maryland Baltimore County, 2011.

[10] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of the 19th International Conference on World Wide Web*, WWW '10, 2010.

[11] K. H. Lim and A. Datta. Following the follower: Detecting communities with common interests on twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 317–318, New York, NY, USA, 2012. ACM.

[12] K. H. Lim and A. Datta. A topological approach for detecting twitter communities with common interests. In M. Atzmueller, A. Chin, D. Helic, and A. Hotho, editors, *Ubiquitous Social Media Analysis*, pages 23–43, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[13] K. H. Lim and A. Datta. An interaction-based approach to detecting highly interactive twitter communities using tweeting links. *Web Intelligence*, 14(1):1–15, 2016.

[14] M. Ozer, N. Kim, and H. Davulcu. Community detection in political twitter networks using nonnegative matrix factorization methods. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 81–88, Aug 2016.

[15] M. Ozer, N. Kim, and H. Davulcu. Community detection in political twitter networks using nonnegative matrix factorization methods. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 81–88, Aug 2016.

[16] E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011.

[17] M. Potamias, F. Bonchi, C. Castillo, and A. Gionis. Fast shortest path distance estimation in large networks. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 867–876, New York, NY, USA, 2009. ACM.

[18] M. Rosvall, D. Axelsson, and C. T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, Nov 2009.

[19] Y. Ruan, D. Fuhry, and S. Parthasarathy. Efficient community detection in large networks using content and links. *CoRR*, abs/1212.0146, 2012.

[20] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proc. of the 19th International Conference on World Wide Web*, WWW '10, New York, NY, USA, 2010. ACM.

[21] S. Soundarajan and J. Hopcroft. Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 607–608, New York, NY, USA, 2012. ACM.

[22] D. Surian, Q. D. Nguyen, G. Kennedy, M. Johnson, E. Coiera, and G. A. Dunn. Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. *J Med Internet Res*, 18(8):e232, Aug 2016.

[23] E. Vathi, G. Siolas, and A. Stafylopatis. Mining and categorizing interesting topics in twitter communities. *Journal of Intelligent and Fuzzy Systems*, 32(2):1265–1275, 2017.

[24] M. Wang, C. Wang, J. X. Yu, and J. Zhang. Community detection in social networks: An in-depth benchmarking study with a procedure-oriented framework. *Proc. VLDB Endow.*, 8(10), 2015.

[25] Y. Zhang and J. Pang. Distance and friendship: A distance-based model for link prediction in social networks. In R. Cheng, B. Cui, Z. Zhang, R. Cai, and J. Xu, editors, *Web Technologies and Applications*, Cham, 2015. Springer International Publishing.

[26] Y. Zhang, Y. Wu, and Q. Yang. Community discovery in twitter based on user interests. *Journal of Computational Information Systems*, page 2012.