Recomendación de Usuarios en Twitter basada en Topología de la Red

Universidad Nacional del Centro de la Provincia de Buenos Aires

Propuesta de Tesis de Grado

Integrante: Roo, Joan Sol

Director: Dr. Marcelo Armentano

Fecha: 06/05/2013.

1. Introducción

Twitter es tanto una red social en línea como así también un servicio de *microblogging* que permite a sus usuarios enviar y leer publicaciones de texto de hasta 140 caracteres, conocidos como "tweets".

Creada en 2006, la red creció considerablemente en estos 8 años, alcanzando más de 500 millones de usuarios, de los cuales 200 millones[1] publican activamente un promedio de400 millones[2] de tweets diarios; también cabe destacar la cantidad de aplicaciones que utilizan la API de Twitter, que ha crecido de 150.000[3] a 1.5 millones en el mismo período de tiempo[4].

Los *tweets* pueden tener cualquier contenido, sin embargo existen usuarios particulares que publican sobre un tema específico. Nos referiremos a estos usuarios que tratan temas específicos como fuentes de información, o simplemente, fuentes.

Un usuario de Twitter puede "seguir" a otro usuario, lo que implica que desea recibir sus nuevas publicaciones, una forma de suscripción. Las relaciones seguidor-seguido pueden ser simétricas, lo que puede considerarse una relación de "amistad". Sin embargo estas relaciones son poco comunes en Twitter [1][2][3]. En contraste, muchos usuarios que utilizan Twitter (alrededor del 67.6%) son buscadores de información sobre sus temas de interés, siguiendo las fuentes de información que hablan de éstos.

Dada la gran cantidad de usuarios que publican información constantemente desde todas partes del mundo, es prácticamente imposible que un usuario pueda conocer a todas aquellas personas que publican información de su interés. Es por esto que los usuarios de Twitter se beneficiarían con un sistema de recomendación de dichas fuentes de información.

Los sistemas recomendadores de información brindan sugerencias personalizadas acerca de lo que el usuario puede encontrar interesante. Este tipo de sistemas es aplicable a las redes sociales como Twitter, por ejemplo, para recomendar a un usuario otros usuarios que publiquen información de su interés. Esto puede ser de gran utilidad para nuevos usuarios de la red, que no saben a qué usuarios comenzar a seguir. También es útil para recomendar nuevos usuarios a usuarios antiguos.

A fin de recomendar, es necesario construir un perfil con los intereses del usuario, para lo que existen dos enfoques: explícito (indicado manualmente por el usuario) e implícito (inferido a partir de su interacción con el sistema); el proyecto utilizará este último.

Recomendar usuarios de interés en Twitter no es una tarea trivial, dadas las características heterogéneas de los intereses que muestran los interesados, algunos de los cuales puede no encontrarse reflejados aun. A esto se suma la selección de candidatos que, dadas las dimensiones de la red y la escasa información que puede extraerse de entradas de 140 caracteres, vuelve inviable la evaluación exhaustiva.

---

[1] http://mashable.com/2012/12/18/twitter-200-million-active-users/
[2] http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/
[3] http://blog.twitter.com/2011/08/your-world-more-connected.html
[4] http://thenextweb.com/socialmedia/2012/03/13/twitters-ecosystem-passes-1-5-million-apps/

En 2010 Twitter incorporó un recomendador basado en la popularidad global y relaciones entre usuarios (recomendando seguidos de los actualmente seguidos)[5]; este algoritmo depende fuertemente de las relaciones de amistad y, tal como hemos expuesto previamente, estas relaciones son poco frecuentes en Twitter. En este trabajo, motivados por lo expuesto anteriormente, se propone un recomendador de fuentes de información, basado en las relaciones asimétricas entre los usuarios, priorizando sus características topológicas: su presencia en la sub-red local del usuario, así como su peso en la red total).

El objetivo es desarrollar un recomendador que permita recomendar usuarios cuyo contenido sea de interés, sin que esté necesariamente relacionado a los intereses actualmente reflejados por el usuario.

2. Trabajos Relacionados

Existen múltiples estudios que fundan las bases para este trabajo: Hong et al.[4] estudian las temáticas y categorías de temáticas dentro de Twitter, mientras que Java et al.[1] estudian las comunidades implícitas que existen en Twitter, que tratan temáticas relacionadas; dentro de estas comunidades los usuarios pueden categorizarse en: fuentes, buscadores de información, o bien amigos.

Otros autores han encontrado que hay información útil en la topología de la red por sí misma: Chen et al. [5] compararon algoritmos basados en relaciones y en contenido para recomendaciones de usuarios, encontrando que el primero es mejor detectando contactos conocidos mientras que el segundo en encontrar nuevas relaciones.

Por otro lado existen trabajos que abordan el problema de recomendación de usuarios dentro de Twitter: Sun et al. [6] utilizaron un algoritmo basado en difusión para obtener un grupo de usuarios que toman el rol de reporteros en emergencias; más relacionado a nuestro trabajo, están los algoritmos de recomendación de usuarios en Twitter a partir de un subconjunto de usuarios presentado por Hannon et. al [7]. Estos autores consideraron múltiples estrategias de generación de perfiles de acuerdo a cómo los usuarios son representados en un enfoque basado en contenidos (sus tweets y los de los usuarios con los que se relacionan), un enfoque basado el filtrado colaborativo (basados en los IDs de sus seguidores y/o de sus seguidos), así como enfoques híbridos.

Tal como hemos mencionado, Twitter ha implementado en 2010 un algoritmo basado en relaciones entre usuarios, pero, dado el recorrido utilizado, depende en gran medida de las relaciones de amistad, y existe evidencia que estas son poco frecuentes.

El trabajo propuesto es una alternativa al algoritmo propuesto por Twitter, un caso particular de filtrado colaborativo, que explota las características de las comunidades, en particular, temáticas e intereses en común; mediante el estudio de las relaciones actuales, se da lugar a un perfil de usuario[8], a mayor cantidad de fuentes siga el usuario, mejores serán los resultados obtenidos.

[5] https://blog.twitter.com/2010/discovering-who-follow

3. Trabajo Propuesto

Luego de realizar un relevamiento bibliográfico, y aproximaciones iniciales a los distintos aspectos del problema, se propone desarrollar una aplicación que evalúe una sub-red local al usuario interesado mediante sucesivos filtrados, seleccionando a cada paso los usuarios más relacionados, y clasificándolos según sus estadísticas de presencia en la red. A continuación en la figura 1 puede observarse las etapas de filtrado del algoritmo.

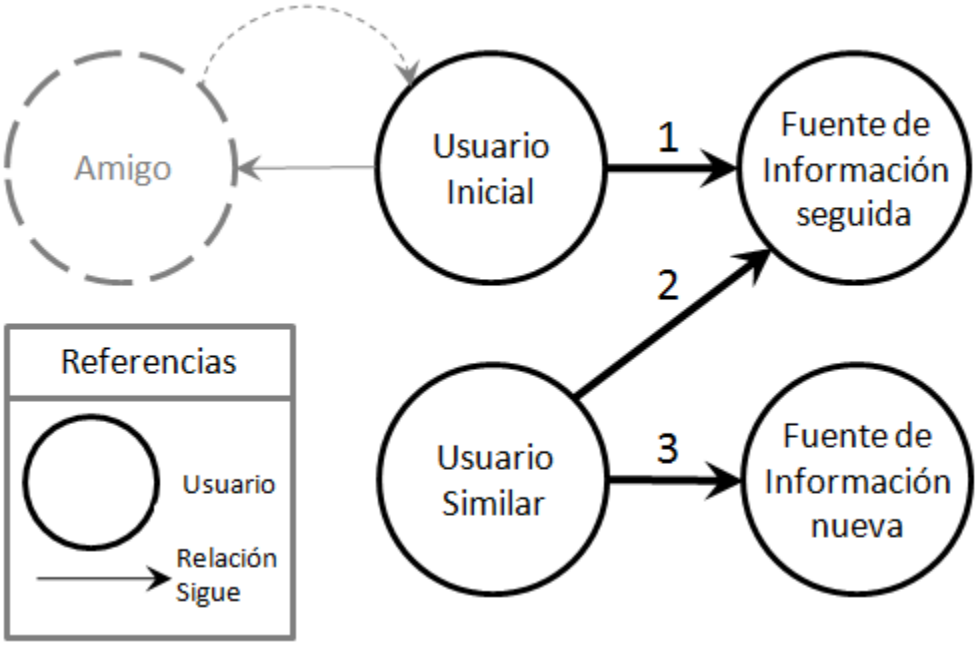

Figura 1: algoritmo de recorrido de la red para recomendar nuevas fuentes

En la última etapa del algoritmo se obtiene un conjunto de usuarios candidatos, pesados según su presencia en la subred recorrida. Sobre este conjunto se aplicarán distintas estrategias de ranking a fin de obtener los usuarios a recomendar.

Dadas características limitantes del API provista por Twitter, se contemplarán distintas estrategias de caché y persistencia de las consultas realizadas, buscando la mejor alternativa para el modelo de datos utilizado. Dadas limitaciones impuestas por el Api de Twitter, se implementarán estrategias para cacheo de consultas, a fin de disminuir el impacto en los tiempos de los algoritmos.

En lo que respecta a los detalles de implementación, se generará una interfaz web a través de la cual el usuario podrá solicitar recomendaciones, permitiéndole configurar los algoritmos utilizados en el proceso, así como observar estadísticas tanto de su perfil, como del recorrido del algoritmo y de los resultados obtenidos.

Una vez generada la aplicación web, se procederá a evaluar el rendimiento de los algoritmos en comparación con otros recomendadores, tanto basados en contenido como el algoritmo propuesto por Twitter, utilizando la técnica de hold-out: se oculta un porcentaje de las fuentes originalmente seguidas por el usuario al cual se desea recomendar y se evalúa que porcentaje de éstas son recuperadas por el algoritmo.

4. Plan de Trabajo

A continuación se detallan las tareas a realizar, con el tiempo estimado de su desarrollo:

1. Relevamiento bibliográfico en el área de Minería de Texto (text mining), Predicción de conexiones en redes sociales (link prediction) y Sistemas de Recomendación (recommender systems). (Tiempo estimado: 1 mes)
2. Análisis de algoritmos existentes en cada una de las áreas del punto anterior orientados a los objetivos particulares de este trabajo. (Tiempo estimado: 1 mes)
3. Diseño e implementación de un algoritmo que seleccione un subconjunto pequeño de usuarios potencialmente interesantes en base a la exploración de la topología de la red a partir de un usuario objetivo. (Tiempo Estimado: 1 mes)
4. Evaluación de distintas alternativas de persistencia de datos para reducir el número de consultas al API de Twitter. (Tiempo estimado: 1 mes)
5. Implementación de una aplicación web que permita testear los algoritmos diseñados. (Tiempo estimado: 2 meses)
6. Realización de pruebas y recolección de resultados para posterior análisis. (Tiempo estimado: 1 ½ mes)
7. Análisis de los resultados obtenidos y comparación con otros sistemas de recomendación que funcionan sobre Twitter. (Tiempo estimado: 1 ½ mes)
8. Redacción del informe final. Tiempo estimado: en forma paralela a las actividades anteriores.

Tiempo total: 9 meses

Bibliografía

1. Java, X. Song, T. Finin and B. Tseng. Why we twitter: understanding microbloging usage and communities. En Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pages 56-65. 2007.
2. B.Krishnamurthy, P.Gill, and M. Arlitt. A few chirps about Twitter. En Proceedings of the 1st WorkShop on Online Social Networks (WOSP '08), pages 19-24, Seattle, WA, USA, 2008.
3. H.Kwak, C. Lee, H. Park and S. Moon. What is Twitter, a social network or a news media?. En Proceedings of the 19th International Conference on World Wide Web (WWW'10), pages 591-600, Raleigh, North Carolina, USA, 2010.
4. L. Hong, B.D. Davison. Empirical study of topic modeling in Twitter. En Proceedings of the SIGKDD Workshop on SMA. (2010)
5. 5. J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. En Proceedings of the 27th International Conference on Human Factors in Computing Systems, pages 201-210, Boston, MA, USA, 2009.
6. A. R. Sun, J. Cheng, and D.D. Zeng. A novel recommendation framework for micro-blogging based on information diffusion. En proceedings of the 19th Workshop on Information Technologies and Systems, 2009.
7. J. Hannon, M Bennett, and B. Smyth. Recommending Twitter users to follow using content and collaborative filtering approaches. En Proceedings of the 4th. ACM Conference on Recommender Systems(Rec Sys '10). Pages 199-206, 2010.
8. S. Schiaffino, A. Amandi. Intelligent User Profiling. En M. Bramer (Ed.): Artificial Intelligence, LNAI 5640, pp. 193 – 216, 2009.