Propuesta para la descarga masiva de seguidores

Eduardo Villalpando Mello

La API tiene tres endpoints para obtener información acerca de relaciones entre usuarios:

GET followers/ids Regresa 5,000 id's por minuto de los seguidores de una cuenta.	GET friends/ids Regresa 5,000 id's por minuto de los usuarios que siga una cuenta.	GET friendship/show Permite hacer 12 consultas por minuto para determinar la relación entre dos cuentas.

Considerando que los usuarios influyentes (aquellos con más de 10,000 seguidores) acumulan un total de 4 mil millones de seguidores, sería necesario hacer 800 mil consultas al endpoint GET followers/ids, por lo que no representa una opción viable. Estos usuarios también acumulan 70 millones de cuentas seguidas, que pueden recuperarse en menos de 48 horas mediante el endpoint GET friends/ids. Sin embargo, esta información no muestra la influencia de los usuarios influyentes sobre el resto de la red. Por último, el endpoint GET friendships/show da total control sobre las comparaciones, pero es imposible comparar todos los usuarios entre sí.

Ante esto, la metodología propuesta es descargar todas las cuentas seguidas por los usuarios influyentes, y comparar a los usuarios influyentes con el resto para que aquellos con la mayor probabilidad de relación sean consultados.

La literatura sugiere diversas metodologías (AlMahmoud & AlKhalifa, 2018) (Zhang, Wu, & Yang, 2012) para comparar la conexión o similitud que existe entre dos usuarios. Este tipo de algoritmos son utilizados (aunque tomando en cuenta muchas más variables) por Twitter (Goel, Sharma, Wang, Yin, & Twitter) para sugerir cuentas qué seguir, mostrar anuncios relevantes a usuarios, ordenar la lista de tweets del feed, entre otros.

Se toman en cuenta señales como: (i) usuarios seguidos y seguidores, (ii) menciones, (iii) retweets, (iv) hashtags y (v) metadatos de perfil para cuantificar la relación que existe entre dos usuarios. Se propone desarrollar este análisis de modo que solo los usuarios con un alto grado de conexión sean consultados con la API, garantizando que se identifiquen las conexiones más fuertes en la base de datos.

El algoritmo toma las señales anteriormente mencionadas y mediante una función sigmoide (para ajustarse mejor a variables discretas como valores verdadero/falso) devuelve la probabilidad de que dos usuarios se sigan.

$$\text{sig}(z) = \frac{1}{1 + e^{-z}}$$

Donde

$$\begin{aligned} z(i,j) = &\gamma_c \text{sim}_{\text{comunidad}}(i,j) + \\ &\gamma_m \text{sim}_{\text{menciones}}(i,j) + \\ &\gamma_r \text{sim}_{\text{retweets}}(i,j) + \\ &\gamma_h \text{sim}_{\text{hashtags}}(i,j) + \\ &\gamma_p \text{sim}_{\text{perfil}}(i,j) \end{aligned}$$

Primero se realizan las comparaciones para obtener los valores de cada una de las variables ($\text{sim}_{\text{comunidad}}, \text{sim}_{\text{menciones}}, \text{sim}_{\text{retweets}}, \text{sim}_{\text{hashtags}}, \text{sim}_{\text{perfil}}$) y se almacenan. Usando la metodología actual, este proceso puede demorar una semana aproximadamente.

Posteriormente, se selecciona un cierto número de filas (sin considerar los resultados arrojados, pero procurando que haya diversidad en los mismos) para realizar la comprobación con la API y almacenar también el resultado (verdadero/falso)

Con esa información es posible usar un modelo de regresión logarítmica para estimar los valores de los pesos. A partir de ello es posible obtener la probabilidad de que dos usuarios se sigan. Solamente si este valor es mayor a 0.5, se realiza la verificación con la API.

Igualmente, tras cierto número de verificaciones es posible volver a entrenar el modelo de regresión logarítmica para que los pesos garanticen siempre el mínimo margen de error.

Usando esta metodología es posible identificar las principales conexiones entre usuarios en menos de dos semanas.

Actualmente se han realizado 8.1 millones de comparaciones, de las cuales solo 3 mil arrojaron datos con valores mayores a 0 (esto debido a que la base de datos no cuenta con todo el universo de información en Twitter). Tras entrenar el modelo de regresión logarítmica con 2 mil resultados, este fue capaz de predecir, con un nivel de confianza del 80%, si dos cuentas se seguían. Se estima que, con mayor cantidad de datos, este valor pueda incrementar considerablemente.

Esta información resulta útil no solo para reducir la cantidad de llamadas necesarias a la API, sino también para identificar las cuentas con mayor grado de interacción entre sí.

Referencias

AlMahmoud, H., & AlKhalifa, S. (4 de Octubre de 2018). TSim: a system for discovering similar users on Twitter. *Journal of Big Data*. doi:https://rdcu.be/cxVUg

Goel, A., Sharma, A., Wang, D., Yin, Z., & Twitter, I. (s.f.). Discovering Similar Users on Twitter. Obtenido de http://snap.stanford.edu/mlg2013/submissions/mlg2013_submission_20.pdf

Iqbal, M. (5 de Julio de 2021). *Twitter Revenue and Usage Statistics (2021)*. Obtenido de Business of Apps: https://www.businessofapps.com/data/twitter-statistics/

Wojcik, S., & Huges, A. (24 de Abril de 2019). *Sizing Up Twitter Users*. Obtenido de Pew Research Center: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Zhang, Y., Wu, Y., & Yang, Q. (2012). Community Discovery in Twitter Based on User Interests. *8*(13), 991-1000. Obtenido de https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.9055&rep=rep1&type=pdf