Creación de playlists a partir de un grafo

Mario Becerra Tania Mendoza Miguel Vilchis

Octubre de 2016

1. Introducción

El estudio de redes ha sido de gran interés en los últimos años, debido a que representan un conjunto de actores y las relaciones entre ellos de una manera intuitiva y se puede tener una representación visual. Muchos sistemas usados diariamente pueden ser modelados por medio de redes, como la relación entre las páginas de internet, redes de transporte y servicios diversos, redes de parentescos y *Facebook*. Las redes comúnmente son modeladas por medio de grafos, ya sean direccionales y no direccionales.

El análisis de los grafos generados a partir de distintas redes, pueden ser utilizados para encontrar características particulares en las redes. Por ejemplo, encontrar las páginas web más importantes sobre algún tema en internet, definir comunidades en un grafo social como *Facebook* o *Twitter*, o incluso en una red telefónica; o calcular el camino más corto de un punto a otro en una red de transporte.

Este trabajo se centra en las redes sociales, las cuales tienen ciertas características que deben cumplir para ser catalogadas como tal [**?**]. Las principales son:

- Existe una colección de entidades que participan en la red, que comúnmente son personas.
- Existe por lo menos una relación entre las entidades de la red.
- Se asume que la localidad de los nodos no es aleatoria. Se debe entender por localidad la posición que ocupan los nodos con relación a los demás, comúnmente se tienden a juntar más con nodos que comparten características similares.

Un claro ejemplo de red social es *Facebook* que se puede representar como un grafo donde los usuarios son los nodos y sus conexiones son las relaciones de amistad entre ellos. De una red con éstas características se puede obtener información de las comunidades que lo integran, que nodos son más relevantes y la similitud entre ellos.

Se analizó una red social de *Twitter*, creada a partir de *retweets* y *replies*. El objetivo es encontrar comunidades a partir de las conexiones que existen entre los usuarios.

La información generada del análisis de redes de este tipo ayuda a entender el comportamiento de las entidades que lo conforman y en el caso de medios como la mercadotecnia y el comercio, pueden ser utilizadas para obtener alguna ventaja.

1.1. Comunidades en las redes sociales

Podemos definir comunidad como un conjunto de entidades que tienen alguna atributo en común y por eso tienen a juntarse en el mismo grupo. Aplicar métodos como *k-means* no capta la esencia de este tipo de redes, ya que cada entidad se asigna sólo a una comunidad y nada más. Las entidades que forman parte de redes sociales comúnmente forman parte de más de una comunidad. El análisis de una red social difiere de las demás redes en que las entidades pueden formar parte de diferentes comunidades, por ejemplo una persona puede ser parte de una institución, un grupo de amigos, pero también pertenece a una familia.

2. Metodología

El principal objetivo es identificar comunidades en el grafo. Existen varios tipos de algoritmos de detección de comunidades, algunos son divisivos, en el sentido que detectan ligas inter-comunidad y después los quitan de la red; otros son aglomerativos, que van juntando nodos recursivamente; y otros están basados en la maximización de una función objetivo.

2.1. Modularidad

Una medida muy usada para hacer conglomerados en redes sociales es la modularidad, la cual es un escalar entre -1 y 1 que mide la fuerza de la división de una red en conglomerados. Una red con modularidad alta tiene conexiones fuertes dentro de las comunidades, pero débiles entre las comunidades. Muchos de los algoritmos basados en optimización como los mencionados anteriormente buscan encontrar particiones que maximicen la modularidad.

En particular, la modularidad Q está definida como la fracción de arcos que están dentro de cada comunidad menos el número esperado de arcos en cada comunidad de un grafo aleatorio con la misma distribución de grados de entrada y salida que el grafo que se estudia. Matemáticamente, esto se ve como[**?**]

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j), \tag{1}$$

donde A_{ij} representa el peso del arco entre los nodos i y j, $k_i = \sum_j A_{ij}$ es la suma de los pesos de los arcos que salen del nodo i, c_i es la comunidad que se le asigna al vértice i, $\delta(u, v)$ es 1 si $u = v$ y 0 en otro caso, y $m = \frac{1}{2} \sum_{ij} A_{ij}$.

Una desventaja de utilizar la modularidad como función objetivo, es que puede fallar en encontrar comunidades pequeñas en una red muy grande, esto debido a la naturaleza de la función de modularidad, que resta el número esperado de vértices en una red aleatoria, el cual va disminuyendo mientras la red va creciendo; por lo que esto puede ser menor que uno, entonces la modularidad puede interpretar esto como signo de correlación fuerte entre dos comunidades, por lo que las juntaría en una sola comunidad.

2.2. *Betweenness*

Betweenness o intermediación es una medida del número de veces que un nodo actúa como puente en el camino más corto entre dos nodos. Es una forma de cuantificar el control que tiene un nodo en la comunicación existente entre otros. La idea básica detrás de esta medida es que los nodos con mayor *betweenness* son los que aparecen con mayor probabilidad en los caminos más cortos, de esta forma, es una medida de centralidad en una red. Formalmente se puede definir como [**?**],

$$C_{BET}(i) = \sum_{j,k} \frac{b_{jik}}{b_{jk}} \tag{2}$$

donde b_{jk} es el número de caminos más cortos desde el nodo j hasta el nodo k, y b_{jik} el número de caminos más cortos desde j hasta k que pasan a través del nodo i.

Los nodos con un alto valor de intermediación son muy importantes en la estructura de una red, ya que comunican comunidades con otras. Comúnmente los valores más altos de *betweenness* son obtenidos por los nodos que están en los bordes de las comunidades. Si uno nodo con alta intermediación desaparece, las comunidades podrían quedar incomunicadas. Calcular el *betweenness* resulta ser una tarea complicada y tardada, pues se recorre toda la red nodo por nodo.

Esta noción de centralidad en los nodos se puede extender a las aristas, y de esta forma el *betweenness* de una arista es el número de caminos más cortos entre el par de nodos que corren a través de esta arista. Así, las aristas que conectan comunidades tendrán mayor nivel de *betweenness*, pues al quitar estas aristas, las comunidades quedarían separadas una de la otra. Esta noción de *betweenness* de aristas se puede explotar para encontrar comunidades en la red. El algoritmo Girvan-Newman hace esto siguiendo los siguientes pasos:

1. Se calcula el *betweenness* de cada arco
2. Se quita el arco con mayor *betweenness*
3. Se recalcula el *betweenness* de los arcos afectados por la acción de haber quitado el arco
4. Se repiten los pasos 2 y 3 hasta que no queden más arcos

Este algoritmo devuelve un dendrograma en el cual las hojas son los nodos, con esto se pueden asignar comunidades a los nodos.