Técnicas Basadas en Grafos Aplicadas al Procesamiento del Lenguaje Natural

Gabriel Vázquez Torres

14 de Enero de 2018

1. Introducción

En esta práctica se van a comparar y describir cuatro grafos con el objetivo de conocer algunas de sus características. Para ello utilizaremos la herramienta **Gephi**.

2. Tabla de comparación entre grafos

A continuación se detallarán los datos pedidos en la práctica mediante la tabla 1.

Grafo/Propiedades	EuroSiS Generale	RetGraph	Les Miserables	Boundary Countries
Nodos	1285	1300	77	146
Aristas	7524	6708	254	1290
¿Es dirigido?	Dirigido	Mixto	No dirigido	Dirigido
¿Tiene pesos?	Sí	Sí	Sí	No
Longitud media de los caminos	4,943	3,337	2,641	1
Coeficiente de Clustering Medio	0,213	0,009	0,736	0
Diámetro de la Red	14	5	5	1
Grado medio	11,711	10,32	6,597	17,671
Grado medio con pesos	5,904	0,52	21,299	8,836
Componentes fuertemente conexas	511	-	-	146
Componentes debilmente conexas	6	1	1	1

Cuadro 1: Tabla comparativa entre grafos.

3. Descripción de diferentes grafos

A continuación se hará un breve análisis comparativo entre los distintos grafos.

3.1. Red de información: EuroSiSGenerale

EuroSiSGenerale se trata de un ejemplo de grafo tipo **red de información**. El grafo representa las Relaciones que tienen lugar entre las páginas web de diferentes fuentes de divulgación científica (SiS:Science in Society) de 12 países europeos. En Cada nodo se representa una institución o actor, mientras que las aristas representan relaciones entre ellos. Los nodos contienen diferentes tipos de información como son el nombre del actor, el país, el tipo de actor (Universidad, ONG, empresa, medio de comunicación, etc.) y las disciplinas o campos en los que actúa (por medio de etiquetas binarias). Este conjunto de datos permiten diferentes posibilidades de análisis y visualización según el tipo de información que se desee investigar o comunicar. En el caso estudiado se ha llevado a cabo una organización por países, pero sería igualmente posible estudiar las relaciones entre los diversos tipos de actores. Las relaciones entre instituciones son complejas, siendo caro y difícil tipificarlas (¿relación de cooperación, de competencia o ambas?) o medir su intensidad. Debido a esta problemática se ha llevado a cabo el estudio de una sola variable simplista y fácilmente observable: la presencia de hipervículos en la página web de cada actor hacia otros .Las aristas del grafo son dirigidas y representan hipervínculos, en los que el actor (la página) que contiene el

enlace es el nodo fuente y el actor hacia el que dirige el enlace es el nodo objetivo. Cada arista puede tener asociado un valor discreto entre 1-2. En el ejemplo los nodos se han coloreado por países. Las aristas adquieren el mismo color que el nodo fuente. De esta manera los distintos países aparecen como clústers o subcomponentes con el mismo color, subrayando el hecho de que son más frecuentes los enlaces entre actores de un mismo país. Es decir, si tomamos solo los nodos de un país, el coeficiente medio de agrupamiento casi siempre es más alto que el del grafo completo. Por ejemplo, cuando este coeficiente es más bajo, como en los subcomponentes de Armenia y Montenegro,se está indicando una integración pobre entre las distintas instituciones del país. Un coeficiente de agrupamiento más alto para un subcomponente (e.g. Bélgica o Finlandia) indica que los miembros de este subcomponente están muy integrados entre sí. A partir del diámetro de la red para estos subcomponentes nos da una idea muy similiar: para Finlandia es de 4 mientras que para Polonia es de 13, casi el de la red completa, 14. El grafo tiene 6 componentes, pero solo 13 de los 1285 nodos no pertenecen al componente "gigante" o, lo que lo mismo, los grupos de actores no integrados con el resto son muy pequeños, de 1–3 actores. El número de componentes fuertemente conexos es de 511, mucho mayor. Esto se debe a la naturaleza de los enlaces representados, que normalmente no son recíprocos. Los actores más pequeños referencian a los más importantes, no viceversa. Los nodos aparecen representados por círculos cuyo radio va a depender del grado del nodo, de forma que los actores principales aparecen destacados. En este caso Los nodos de mayor grado corresponden a instituciones de Bélgica y Finlandia. De estos, destacan por tener un grado de entrada (más referenciados por otros) la Universidad Libre de Bruselas y la European Science Foundation, mientras que un grado de salida alto es indicativo de un hub (referencia a muchos otros actores) como Research.be. En el caso de Bélgica, la presencia de nodos con un grado de salida o de entrada mucho mayor que el resto indicaría una estructura interna más jerarquizada, la cual tiene sus funciones más definidas (una Universidad principal, una Sociedad Nacional de Ciencia en Sociedad). Por otro lado la red de Finlandia, cuyo coeficiente de agrupamiento es incluso más alto, posee una estructura más horizontal, con menos diferencias entre actores principales y secundarios. Los actores internacionales se representan en el centro del grafo, al estar conectados a nodos de varios países. Entre estos enlaces podemos intuir la presencia de muchos puentes locales. Si llevamos a cabo un estudio de la centralidad de alguno subcomponentes como Francia o Italia podríamos concluir que se trata de redes más integradas a nivel internacional que nacional. Es importante notar que el grafo solo contiene instituciones de 12 países, lo cual es un sesgo: muchos nodos pueden aparecer como débilmente conectados si la mayoría de sus relaciones se dirigen a países que no aparecen en el grafo, lo cual puede no ser representativo de la realidad.

3.2. Red social libre de escala: Les Miserables

En Este tipo de grafo aparece una representación de un texto literario, la obra Les Miserables de Victor Hugo. Los nodos representan personajes de la novela (solo están etiquetados con el nombre del personaje), mientras que las aristas representan la co-ocurrencia de dos personajes en el mismo capítulo y su peso corresponde al número de capítulos en que aparecen ambos personajes. Podemos hablar en este caso de un grafo de co-ocurrencia, no dirigido. Se podría decir que el grafo representa una red social ficticia, pero hay que tener en cuenta que una representación estática de una red social en la mayoría de casos sólo representa un instante dado, mientras que en este grafo las relaciones ocurren a lo largo de una línea temporal (ficticia). Esta basado en una compilación de datos por parte de Donald Knuth. El tamaño de los nodos depende nuevamente de su grado. El nodo de mayor grado (36) aparece en el centro del grafo, y corresponde al personaje protagonista, Valjean. Este personaje aparece a lo largo de toda la obra y por tanto ocurre junto a casi la mitad del total de personajes. Alrededor del nodo central aparecen nodos cuyo grado supera la media (6,6), como Fantine (15), Javert (17), Thenardier (16), Gavroche (19), Cosette (11), Marius (19) o Myriel (10). Cada uno de estos personajes es el centro de su propio subcomponente en el grafo, lo cual revela que estos personajes son personajes principales en alguna parte de la trama ya que la obra tiene varios volúmenes. En la representación también aparecen un grupo de personajes secundarios situado alrededor de los nodos principales . en ocasiones Estos grupos suelen estar poco o nada relacionados entre sí, ni tampoco con el personaje principal. Esto hace referencia a que cada uno de los protagonistas cobra importancia en una parte de la historia, dando acceso a un grupo de personajes secundarios que solo se relacionan con él. un ejemplo de esta estructura es Fantine: en la vista del grafo, los nodos de color azul-cyan representan al entorno de Fantine, están todos relacionados entre sí y por tanto su coeficiente de agrupamiento es 1. Solo Tholomyes está

relacionado con personajes de fuera de este entorno, y por tanto su coeficiente de agrupamiento es menor. Podemos decir pues que cuanto mayor es el coeficiente mayor es la relación. Este modelo de red concuerda con las características de una red social libre de escala.

3.3. Red aleatoria: RetGraph

En este tipo de grafos se muestra la co-ocurrencia entre palabras en una misma frase sobre un conjunto de documentos . en este caso a diferencia de la regla general las aristas son dirigidas mostrando así una relación semántica no recíproca entre los nodos por ejemplo una relación de hiperonimia. Los nodos sólo aparecen etiquetados su número identificador, por lo que no conocemos las palabras que representan Podríamos inferir que los 7 nodos con grado de salida son posiblemente hiperónimos, y los 6 con grado de entrada son posiblemente hipónimos. El coeficiente de agrupamiento es muy bajo, siendo para dos tercios del total de nodos (que no tienen vecinos conectados entre sí). Sin embargo, el grafo solo tiene 1 componente. La distribución de grado que aparece es una distribución normal. Todas estas características son propias de una red aleatoria.

3.4. Red de afiliación: BoundaryCountries

Este tipo de grafo recibe el nombre de bipartido ya que los nodos pertenecen a dos categorías separadas, representando países o distritos. Las diferentes aristas son dirigidas, siendo el nodo fuente siempre la categoría "País", y el nodo objetivo de la categoría "Distrito Cada nodo País enlaza a su vez con 10 nodos Distrito, de modo que su grado de salida es 10 y su grado de entrada 0. Por su parte, los nodos Distrito tienen un grado de salida 0, y su grado de entrada, es decir, el número de países que contienen, oscila entre 72 y 81. Esto hace que el grafo sea casi bi-regular. La propia distribución del grado permite identificar visualmente las categorías iniciales del conjunto de datos. Otra característica del grafo es que Al ser un grafo digirido, la longitud de los caminos es siempre 1, es decir, no permite enlazar un país con otro debido a la dirección de las aristas. Al no existir aristas entre nodos de la misma categoría, el coeficiente de agrupamiento es 0. En este grafo se muestra una estructura típica de red de afiliación ya que al Seleccionar los distintos distritos podemos observar los países que engloban. Es decir, se representa una relación entre países de co-pertenencia a uno o más distritos, mientras entre dos o varios distritos podrían computarse relaciones de solapamiento.