Travaux Pratique

Graph Mining

Advanced Data Mining Report

Master Degree in Big Data Management and Analytics

Prepared by: *Supervised by:*

Ferreira, Jessé *RAMEL, Jean-Yves*

Academic Year : 2019-2020

UFR des sciences et techniques - Site de Blois

Contents

List of Figures

1 TP Questions

1. Question: Si on considère un graphe G pour lequel :
   1. les nœuds sont les personnages et les comics
   2. les arêtes sont les liens (personnage, comic)

   Ce graphique ressemble à un graphique mixte, puisqu'un personnage peut apparaître dans plusieurs comics et plusieurs comics peuvent montrer (graph oriente) et aussi faire référence à plusieurs personnage, par contre ce personnage peut n'apparaît pas dans les comics qu'il etait cité (graph non oriente).

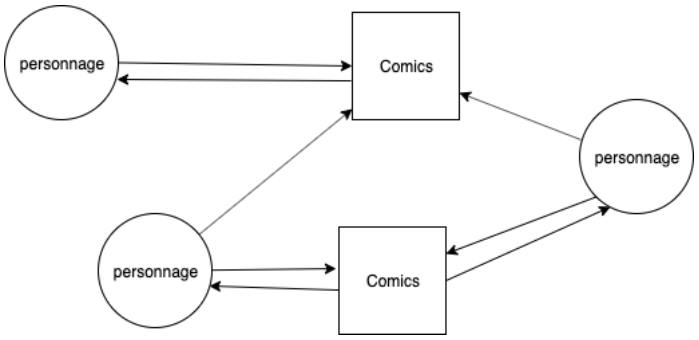

Figure 1: Graph de type Mixte (Oriente et non oriente)

   Pour transformer G* en un graphe de type 'réseau social' à partir du jeu de donnée 'Marvel Universe', on doit change les arêtes (personnage, comic) par (personnage, personnage) et attribué le comic à les atttribu des personnage. Ainsi, le graph sera un graph orienté.

   Donc, le poids de G* pourrait etre associé à le nombre de fois qu'un personnage apparaît dans un comics.

2. Question: La réponse est sur le code.
3. Question: La réponse est sur le code.
4. Question: La réponse est sur le code.
5. Question: La réponse est sur le code.

6. Question: Reponse: Cela permet de meilleur voir les pattern et la distribution des noeuds dans le canvas en ameilerant la visualisation.

7. Question: la réponse est sur le code.

8. Question: la réponse est sur le code.

9. Question:

```
pop_heros = []

foreach pop_hero in edges:
    if pop_hero.weight >= 50:
        popularity = sum of unique weights
        of pop_hero edges
        pop_heros.push({ pop_hero, popularity })

return sortedDesc(pop_heros)
```

10. Question: la réponse est sur le code.

11. Question: Il a y encore des noudes que empêche d'avoir une visualisation propre de l'autres noudes seletioné, dont il fault supprimer les noudes sans edges (target et source).

12. Question: Cette fonction récupérera les autorités en vérifiant les statistiques de hits dans chaque nœud. L'objectif est de rechercher les nœuds les plus pertinents. Ensuite, la liste des autorithies est triée pour donner la couleur aux top noeuds sélectionnés.

    Les 5 personnages les plus importants, selon le critère HITS sont:
    ‘‘‘
    HITS 0 : CAPTAIN AMERICA
    HITS 1 : HAWK
    HITS 2 : ANT-MAN/DR. HENRY J.
    HITS 3 : THING/BENJAMIN J. GR
    HITS 4 : HULK/DR. ROBERT BRUC
    ‘‘‘

13. Question: les 10 personnages les plus importants, selon le critère des degrés élevés sont:

    ‘‘‘
    HITS 1 : CAPTAIN AMERICA
    HITS 2 : HAWK

HITS 3 : ANT-MAN/DR. HENRY J.
HITS 4 : THING/BENJAMIN J. GR
HITS 5 : HULK/DR. ROBERT BRUC
HITS 6 : QUICKSILVER/PIETRO M
HITS 7 : HUMAN TORCH/JOHNNY S
HITS 8 : WONDER MAN/SIMON WIL
HITS 9 : MR. FANTASTIC/REED R
HITS 10 : IRON MAN/TONY STARK
‘‘

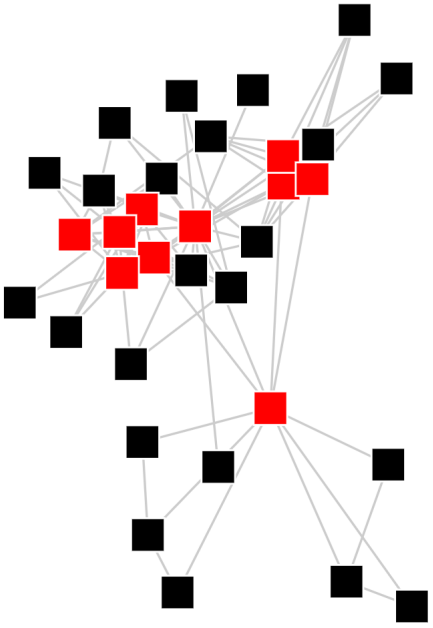

Figure 2: The top 10 highest degree nodes.

14. Question:
    Part 1:
    Les colour des noudes sont attribué en basant sur les classes des communautés de chaque noudes qui est generé par l’index calculé sur le methode statistc du

louvain.

Part 2:
les repartions sont ordenées et mapped dans l'order croissante donc l'algorithme de Louvain reorganise les graphs en basant sur le clustering hierarchique, en regroupant les noudes voisins en communautées.

15. Question:

À un moment donné peut etre exister deux graphes $G_1(N_a, E_a)$ et $G_2(N_b, E_b)$ avec éventuellement un nombre différent de nœuds et d'arêtes, et il peuvent avoir des correspondance entre graph, c'est-à-dire, avoir la même quantité de noudes voisins, dont les voisins aussi peuvent avoir les mêmes nombre des proichain voisins et etc. Donc, pour mesure la similarité on peut comparer les ensemble de voisin de chaque G (G_1 et G_2), par contre, ça peut deviens coute.

L'intéret en cet type de analyse est qu'on peut trouve de heros que peuvent avoir de mêmes amis et aussi enimis, et donc ils peuvent etre enssemble pour sauve la vie humane contre Thanos, comme dans le film "Avengers: Endgame".

```
voisans1 = getNeighbours(sig_GA, nodeId_GA)
voisans2 = getNeighbours(sig_GB, nodeId_GB)

for i in voisans1.nodes:
    for j in voisans2.nodes:

        if i != j
            dissimilarity_flag = 0
        else
            similarity_flag = 1
```

16. Question:
Là l'ideé est un peu pareil à cela sur la dernier question, mais là on cherche les communauté que se rasenssembles à la fois de concentré de l'analyse sur chaque individuos, maintenant on cherchee le similarité de sub-graph. On peut faire cela aprés le partition hierarchique, et donc, parcourir les enssemble de sub-graph en demandant et comparant les labels.

L'intéret en cet type de analysis est de trouver de mêmes caracteristques entre

des communautes.

```
for i in graphA.nodes:
    for j in graphB.nodes:

        if node[i]._louvain != node[j]._louvain
            dissimilarity_flag = 0
        else
            similarity_flag = 1
```

17. Question 17 : Aprés la repartions de communautées, l'algorithme va mesures les poids entre noudes de une communautées et entre communautées. Oú le densité intra-cluster c'est le nombre de arcs interne à une communautées par quantite de possibles noudes, dont de global internal_density sera la moyenne entre les repartions. Et inter-cluster c'est la maximum distance entre les communautées.

    $intracluster = (Sumdes|E|/|V| * |V| - 1)/$ nb repartions total
    $intercluster = |E| - |ExternalE|/(|V| * |V| - 1) -$ quantite total de possibles noudes

    Nb internal edges = 80
    Internal density = 0.25774198568316214
    External density = 0.0016790597265531303

18. Question: La réponse est sur le code. Attention, il y a bug dans cette function. Pour l'utilise, il faut d'abord click dans le respective button et donc choisir deux noudes à voir la mieux distance. Puis, il faut click à neuveux pour afficher le resulté sur console et sur le tab.