Résumé projet magistère - Théorie des réseaux

Valentine Carteron
Alexis Mareau
Lien GitHub : *https://github.com/valentinecarteronGitHub/Projet-Magist-re_Graphes*

Introduction

Dans le cadre du projet de magistère sur la théorie des réseaux, nous avons décidé d'étudier la base de données de Netflix France. Netflix est le premier service de streaming, et étant utilisateurs de ce service, nous trouvions intéressant de travailler sur cette base de données filmiques. Nous nous sommes demandés comment recommander, à l'aide de graphes, du contenu à un utilisateur à partir de son historique ?

Notre projet se base sur 3 parties principales :

- Les données (l'extraction et le nettoyage) ;
- Les graphes (la création, la visualisation, la description et les centralités) ;
- Les systèmes de recommandation.

I - Les données

Nous avons à notre possession deux jeux de données. Le premier est une base de données de Netflix France qui a été conçu par nos soins en *scrapant* trois sites web. Pour le second, il s'agit de l'historique de visionnage d'un utilisateur.

La base de données Netflix est composée de 3996 individus et de 6 variables. Celles-ci sont :

- `Titre`, `Identifiant`, `Casting`, `Type du contenu` (film ou série) : obtenues par web scraping sur *fr.flixable.com*
- `Genres` : obtenue par web scraping sur *https://www.netflix.com/fr*
- `Note IMDb` : obtenue par web scraping sur *https://www.imdb.com/*

Tout utilisateur de Netflix possède un historique de visionnage. Ce dernier est récupérable en suivant ce lien : *https://www.netflix.com/viewingactivity*. Une fois le fichier csv téléchargé, il faut le nettoyer à l'aide d'expressions régulières de façon à garder une liste unique de films et séries (car il apparaît dans l'historique chaque épisode vu d'une même série).

II - Les graphes

Concernant les graphes, leurs nœuds correspondent aux identifiants des films/séries (`Identifiant`), avec pour poids la valeur de leur `Note IMDb`. Une arête existe entre deux noeuds s'ils ont au moins un genre ou un acteur en commun. Le poids des arêtes correspond à la somme des genres et du casting en commun entre deux films.

Nous avons effectué six études différentes sur nos données (deux seront proposées lors de notre présentation) :

- Sur notre base de données complète (films et séries) ;
- Sur seulement la base de données des films ;
- Sur seulement la base de données des séries ;
- Sur notre base de données (films et séries) avec suppression des trois genres les plus fréquents ;
- Sur seulement la base de données des films avec suppression des trois genres les plus fréquents ;
- Sur seulement la base de données des séries avec suppression des trois genres les plus fréquents.

Le choix de supprimer les trois genres les plus fréquents (Comédie, Drames et Films documentaires) s'appuie sur le fait que le graphe de base (films et séries) est hyper-connecté. Après

mise en place de cette décision, nous avons remarqué que cela ne changeait pas grand chose, on garde un graphe hyper-connecté, du fait du gros volume de données que nous étudions.

Pour chacune de ces études, nous avons créé le graphe, regardé sa description (le nombre de nœuds, la densité, le degré minimum, maximum et moyen) et calculé quatre centralités différentes.

La suppression des 3 genres les plus fréquents a pour conséquence une diminution du degré maximal (de 1714 à 1100), du degré moyen (de 692 à 414) et de la densité du graphe (de 0,17 à 0,1). Le fait que la densité soit proche de 0 signifie que nos nœuds sont assez isolés les uns aux autres. Au contraire, une valeur de 1 signifie que chaque nœud est relié aux autres nœuds.

Les quatres centralités calculées sont :

- La centralité de degré qui mesure la popularité ;
- La centralité de proximité qui mesure la proximité ;
- La centralité d'intermédiarité qui mesure la facilité de connexion ;
- La centralité de vecteur propre qui mesure l'influence du nœud dans le réseau.

III - Les systèmes de recommandation

Nous avons créés quatre systèmes de recommandations de type *content-based*. C'est à dire que la recommandation est faite selon le contenu visionné par un utilisateur et non d'après une ressemblance de comportement d'utilisateurs. Parmi ces systèmes, il y en a deux qui utilisent le coefficient de Jaccard comme mesure de similarité et deux qui utilisent l'indice Adamic/Adar. Pour chacune de ces mesures, il y a un système qui prend en compte le poids des arêtes et l'autre non.

Le coefficient de Jaccard entre un noeud *u* et un noeud *v* se calcule comme le rapport entre l'ensemble des voisins communs à *u* et *v* et l'ensemble des voisins de *u* et *v*. Cette indice est une mesure classique, souvent utilisé et facile à comprendre, c'est pour cela que nous l'avons utilisé. L'indice d'Adamic/Adar entre un noeud *u* et un noeud *v* est définit comme la somme de la centralité de degré logarithmique inverse des voisins partagés par *u* et *v*. Une telle définition est basée sur le concept que les éléments communs à *u* et *v* avec de très grands voisinages sont moins significatifs par rapport aux éléments communs à *u* et *v* avec de très petits voisinages.

Les contenus recommandés par ces systèmes étant propres à chaque utilisateur, on ne peut pas conclure de manière générale sur la pertinence des résultats obtenus sans l'avis de cet utilisateur.

Conclusion

Les systèmes de recommandation semblent correctes à priori. Cependant, un des inconvénient est que nous n'avons aucun moyen de vérifier leur précision. Ainsi, il aurait été bien d'avoir un retour sur la recommandation proposé à l'utilisateur.

Parmi les améliorations futures, il serait intéressant de construire notre système de recommandation en prenant en compte le poids des noeuds, c'est-à-dire les notes IMDb des films. Sauf que pour certains films nous ne disposions pas de notes. Une valeur manquante concernant une note pour un film aurait pu être remplacé par la moyenne des notes de ses voisins, ou par la note correspondant au film dont il est le plus similaire selon l'indice de Jaccard. Nous aurions pu également changer le poids des liaisons, par exemple mettre une importance plus forte pour le genre que pour le casting ou bien séparer les films et les séries pour la recommandation.