Métriques pour l'analyse de grands graphes bipartis

Raphaël Fournier-S'niehotta, Cédric Du Mouza, Tiphaine Viard

Contexte

Les systèmes de recommandation sont apparus il y a vingt-cinq ans, afin d'aider les utilisateurs à filtrer les masses d'information disponible et de leur permettre de trouver facilement des contenus qui les intéressent. Aujourd'hui, ces recommandations se trouvent au cœur de nombreux systèmes : lecture d'actualités, achats en ligne, recherche d'information. Les algorithmes permettant de mettre ainsi en relation des *utilisateurs* avec des *articles* ont significativement évolué. L'idée du filtrage initial ("les utilisateurs qui ont interagi avec cet élément ont aussi interagi avec ces éléments") est restée, mais les systèmes de recommandation reposent désormais sur des techniques plus évoluées : factorisation matricielle [2], filtrage collaboratif [3] ou apprentissage automatique [4].

Ces méthodes sont également très utilisées dans le cadre de l'analyse de traces réseau. Ces traces prennent typiquement la forme d'une séquence de paquets (t, u, v), signifiant que la machine u a envoyé un paquet à la machine v au temps t. L'analyse de telles traces permet de répondre à des questions majeures en termes de sécurité, et permet de mieux comprendre l'utilisation qui est faite du réseau. Traditionnellement, ces traces sont analysées par le biais de statistiques plus ou moins agrégées, qui sont ensuite modélisées comme des séries temporelles, ou intégrées à des algorithmes d'apprentissage (*machine learning*).

Dans ces deux cas, bien que très naturelle, la modélisation de ces systèmes comme des graphes reste assez peu répandue, d'une part car ce sont des objets de grande dimension difficiles à manipuler, et d'autre part car les méthodes de l'état-de-l'art donnent des résultats satisfaisants. Dans cette thèse, nous proposons d'étudier la conception de métriques adaptées à l'analyse de tels graphes. Explorer avec discernement des bibliographies vastes, au croisement de la théorie des graphes, de la recommandation et de l'analyse de trafic IP sera donc un aspect crucial du travail.

La thèse se centrera donc sur l'étude de graphes bipartis dans le contexte de la recommandation. Afin d'assurer la solidité et la généricité des métriques définies sur ces graphes, l'analyse de trafic IP, dont la mesure est traditionnellement bipartie, sera un cas d'application secondaire.

Les données disponibles pour la thèse proviennent de jeux de données déjà collectés. Pour la partie recommandation, des données provenant de MovieLens, Lastfm ou Amazon sont régulièrement utilisées. Pour la partie trafic IP, la base de trafic MAWI fait référence dans le domaine de l'analyse de traces réseau. Dans ces ensembles, de nombreuses données sont manquantes ou erronées, et il faut prendre ces erreurs en compte.

Enjeux

Quelques méthodes de l'état-de-l'art étudient un système de recommandation comme un graphe. Cependant, peu de propriétés pertinentes existent pour étudier ces graphes intrinsèquement bipartis, et les approches de l'état-de-l'art étudient majoritairement les propriétés de ses projetés (voir Figure 1), dans lequel des utilisateurs sont reliés s'ils ont au moins 1,2,...,k nœuds en commun, afin de revenir à un graphe uniparti. Cependant, cette projection induit une importante perte d'information, et il n'existe pas de consensus sur la manière la plus adéquate de projeter.

Par ailleurs, là où obtenir une similarité de vecteurs est relativement simple, la question de similarités de graphes est un problème de recherche à part entière. Il existe quelques indicateurs qui peuvent être testés pour obtenir des similarités entre utilisateurs (ou articles) à partir du graphe (ou du projeté), par exemple la distance d'édition [5], qui compte le nombre de liens qu'il faut ajouter ou retirer pour passer d'un graphe G_1 à un graphe G_2. Une fois la similarité choisie et calculée, on obtient des groupes d'utilisateurs proches de l'utilisateur courant, dans les profils desquels il est pertinent de rechercher des produits à recommander.

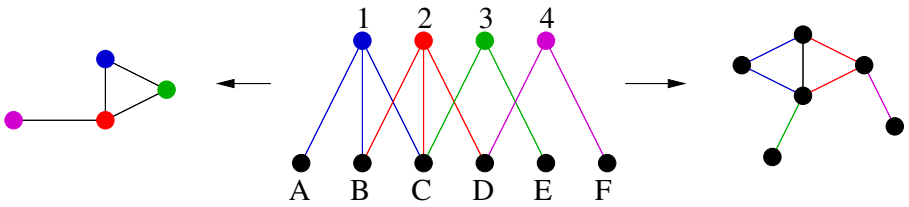

Figure 1: **Centre :** Un graphe biparti ($\top = \{1, 2, 3, 4\}, \bot = \{A, B, C, D, E, F\}, E$) (au centre) et ses deux projections. **Gauche :** Projection sur l'ensemble de nœuds $\top = \{1, 2, 3, 4\}$, c'est-à-dire qu'il y a un lien entre deux nœuds u et v de $\top$ si et seulement si u et v ont (au moins) un voisin en commun dans $\bot$. Ainsi, il y a un nœud entre 1 et 3 dans le $\top$-projeté car 1 et 3 ont C en commun dans $\bot$. **Droite :** Projection sur l'ensemble de nœuds $\bot$.

Un premier enjeu de cette thèse consiste donc à explorer la diversité d'indicateurs existants et à examiner leurs limites dans le contexte spécifique de la recommandation.

Un deuxième enjeu sera de poursuivre cette étude en proposant de nouveaux indicateurs dédiés à ce contexte bien précis. En particulier, ces métriques devront être adaptées aux données incomplètes et aux biais que cela induit.

Enfin, un dernier enjeu consistera à assurer l'aspect algorithmique du passage à l'échelle : comment calculer et mettre à jour ces indicateurs dans les grands graphes dynamiques efficacement.

Pré-requis et déroulement de la thèse

Des connaissances précises du domaine de la recommandation, de l'analyse réseau ou des graphes sont les bienvenues mais ne sont pas requises pour démarrer cette thèse. En revanche, des capacités d'abstraction, de formalisation et d'adaptation sont souhaitées.

Ouverture à l'international

Le projet s'insère dans un contexte de collaborations internationales, notamment avec le National Institute of Informatics (NII), à Tokyo, qui est à l'initiative de la base de trafic *MAWI*.

References

[1] Tiphaine Viard, Matthieu Latapy, Clémence Magnien. Computing maximal cliques in link streams. Theoretical Computer Science, 609, 245-252, 2016.

[2] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8).

[3] Schafer, J. H. J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. The adaptive web, 291-324.

[4] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook (pp. 1-35). Springer US.

[5] Gao, X., Xiao, B., Tao, D., & Li, X. (2010). A survey of graph edit distance. Pattern Analysis and applications, 13(1), 113-129.