Apprentissage de Représentation Appliqué à la Recommandation pour la Littérature Scientifique

Robin Brochier
Université de Lyon, Lyon 2, ERIC EA3083
Digital Scientific Research Technology
Directeurs de thèse: Julien Velcin et Adrien Guille

Objectifs de la thèse

Ma thèse s'effectue en collaboration CIFRE avec DSRT. Cette entreprise développe *Peerus*, une application destinée à faciliter la veille scientifique des chercheurs. L'objectif est de proposer un modèle capable de plonger dans un même espace de représentation des données aussi variées que celles de la littérature scientifique afin d'aborder différentes tâches de recommandation, dont la recherche d'experts. Pour ce faire, mes travaux se divisent en trois étapes :

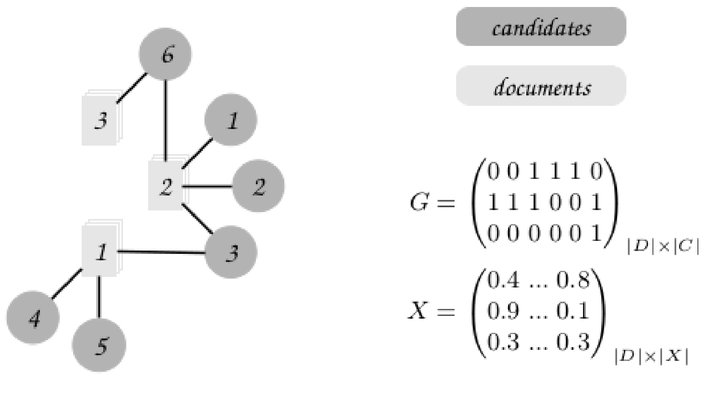
- explorer les techniques de plongement de graphe, notamment celles permettant de tirer profit de potentiels attributs liés aux nœuds
- améliorer la prise en compte d'attributs textuels en s'inspirant de récentes avancées dans le traitement automatique du langage naturel
- intégrer l'hétérogénéité dans le modèle afin de l'appliquer à diverses tâches de recommandation.

Contexte

De nombreuses applications, devenues outils du quotidien, proposent de chercher et filtrer les vastes sources de données disponibles sur le Web. En particulier, on compte une multitude de plateformes traitant de la littérature scientifique. Du simple moteur de recherche d'articles scientifiques au réseau social pour chercheurs, toutes utilisent comme données les publications quotidiennement produites à travers le monde. Pour le chercheur, faisant face à ce déluge d'information, il est devenu laborieux, voire impossible, de réaliser une veille régulière et exhaustive de ses domaines d'expertise.

Une tâche : la recherche d'experts

Le recherche d'experts consiste à soumettre une requête textuelle et à y associer un ensemble ordonné de candidats. La plupart des évaluations réalisées dans la littérature repose sur des ensembles vérités terrains de **requêtes-thématiques**, correspondant aux intitulés des domaines d'expertise. Je propose dans [1] de créer un nouvel ensemble de **requêtes-documents** pour l'évaluation d'algorithmes de recherche d'experts. Celles-ci sont échantillonnées à partir des publications des experts connus de la base de données et permettent d'évaluer non seulement la précision d'un algorithme, mais aussi sa robustesse.



Résultats sur l'évaluation de la recherche d'experts

(a) Scores avec requêtes-thématiques.

		TF-IDF	LSI
Vote	AUC	0.793±0.136	**0.857**±0.048
	P@10	0.800±0.141	0.729±0.158
	AP	0.636±0.171	**0.599**±0.123
	RR	**1.000**±0.000	**1.000**±0.000
Prop ($\eta = 0.1$)	AUC	**0.866**±0.100	0.834±0.052
	P@10	**0.843**±0.118	0.686±0.181
	AP	**0.676**±0.139	0.564±0.140
	RR	**1.000**±0.000	1.143±0.350

(b) Scores avec requêtes-documents.

		TF-IDF	LSI
Vote	AUC	**0.637**±0.129	**0.634**±0.132
	P@10	**0.417**±0.274	**0.370**±0.266
	AP	**0.338**±0.168	**0.318**±0.161
	RR	**2.752**±4.211	**3.686**±5.174
Prop ($\eta = 0.1$)	AUC	0.625±0.121	0.612±0.119
	P@10	0.381±0.283	0.339±0.278
	AP	0.319±0.163	0.298±0.158
	RR	3.557±5.171	4.558±6.141

(c) Écarts types.

TF-IDF	LSI
0.056	*0.058*
0.112	*0.115*
0.073	*0.071*
0.908	*1.307*
0.074	0.079
0.181	0.195
0.104	0.108
2.206	3.320

Apprentissage de représentation dans les réseaux de documents

La structure des liens au sein d'un réseau renferme d'importantes informations sur ses nœuds. Une approche pour faciliter le traitement de ces informations consiste à apprendre des représentations de ces nœuds en utilisant des techniques originellement appliquées au plongement de mots dans un espace de faible dimension. Dans [2], je propose un algorithme de plongement de graphe, *GVNR* (Global Vectors for Node Representation), qui factorise la matrice des comptes de co-occurrences des nœuds lors de marches aléatoires. En formulant le problème comme une régression sur les valeurs non nulles de la matrice et sur des valeurs nulles aléatoirement échantillonnées puis en intégrant des représentations des mots associés aux nœuds dans la factorisation, ce modèle plonge dans un même espace mots, nœuds et contenus textuels.

$$\text{GVNR} : \underset{U,V,b^U,b^V}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{n} s(x_{ij})\left(u_i \cdot v_j + b_i^U + b_j^V - \log(c + x_{ij})\right)^2$$

$$\text{GVNR-t} : \underset{U,W,b^U,b^V}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{n} s(x_{ij})\left(u_i \cdot \frac{\delta_j\, W}{|\delta_j|_1} + b_i^U + b_j^V - \log(c + x_{ij})\right)^2$$

$$s(x_{ij}) = \begin{cases} 1 & \text{si } x_{ij} > 0, \\ m_i & \text{sinon, avec } m_i \sim \text{Bernoulli}(\alpha_i). \end{cases}$$

$$\text{où } \alpha_i = \begin{cases} k \times \frac{p_i}{1-p_i} & \text{si } p_i \leq (k+1)^{-1}, \\ 1 & \text{sinon} \end{cases}$$

Résultats sur l'apprentissage de représentations

Mesures F_1 sur un réseaux de citations d'articles scientifiques (graphe seulement) associés à des conférences.

	% of training data				
	10%	20%	30%	40%	50%
GloVe	57.7	62.4	69.5	72.8	73.8
GVNR ($x_{\min} = 0$)	58.5	62.5	70.7	73.4	75.0
NetMF	65.7	**72.9**	**76.4**	**78.6**	79.4
DeepWalk	67.8	71.6	74.5	75.8	79.2
GVNR ($x_{\min} = 1$)	**69.5**	72.6	75.9	78.1	**80.2**

Mesures F_1 sur un réseaux de collocations de mots associés à des étiquetages morpho-syntaxiques.

	% of training data				
	10%	20%	30%	40%	50%
GloVe	**34.0**	**44.1**	**46.7**	**47.7**	**48.6**
($x_{\min} = 0$)	31.7	40.7	43.2	44.7	45.1
NetMF	27.5	33.5	36.2	37.7	38.7
DeepWalk	33.6	43.6	46.2	47.6	48.2
GVNR ($x_{\min} = 1$)	32.2	41.7	44.0	45.2	46.1

Mesures F_1 sur un réseaux de citations d'articles scientifiques (graphe et texte) associés à des conférences.

	% of training data				
	10%	20%	30%	40%	50%
LSA	54.7	61.0	62.4	63.0	62.8
DeepWalk+LSA	73.8	77.9	78.4	78.1	78.1
TADW	77.1	78.8	78.2	78.8	78.6
GVNR-t	**79.3**	**80.7**	**80.8**	**81.4**	**81.1**

Perspectives

- Avant l'été 2019 : explorer et améliorer le traitement du texte dans les réseaux de documents, initié avec les mécanismes d'attention dans [3].
- Avant fin 2019 : intégrer l'hétérogénéité dans le modèle pour aborder diverses tâches de recommandation.
- Construire un jeu de données complet de type *requêtes-documents* pour la recherche d'experts.
- Évaluer des modèles d'apprentissage de représentation sur des tâches de recommandation.

Références

[1] Robin Brochier, Adrien Guille, Benjamin Rothan, and Julien Velcin.
Impact of the query set on the evaluation of expert finding systems.
In *BIRNDL 2018 (SIGIR 2018)*, 2018.

[2] Robin Brochier, Adrien Guille, and Julien Velcin.
Global vectors for node representations.
In *Proceedings of the 2019 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2019.

[3] Robin Brochier, Adrien Guille, and Julien Velcin.
Link prediction with mutual attention for text-attributed networks.
In *Companion of the The Web Conference 2019 on The Web Conference 2019*. International World Wide Web Conferences Steering Committee, 2019.

Contact

- Web : http://eric.ish-lyon.cnrs.fr/11-FR-membre-Robin.BROCHIER
- Email : robin.brochier@univ-lyon2.fr

peerus
INSPIRING RESEARCH