Thèse présentée pour obtenir le grade de Docteur de l'Université Lumière Lyon 2

École Doctorale Informatique et Mathématiques (ED 512)
Laboratoire ERIC (EA 3083)
Discipline : Informatique

Apprentissage de Représentation dans les Réseaux de Documents : Application à la Littérature Scientifique

Robin BROCHIER

Introduction

Introduction

Contexte

Contexte de cette thèse

- Thèse démarrée en avril 2017.
- CIFRE avec DSRT (Digital Scientific Research Technology), fondée en 2014.
- Peerus : offrir aux chercheurs un outil de veille scientifique ciblé et efficace.
- Laboratoire ERIC de l'université Lyon 2.

Objectifs industriels

peerusreview

Review of *The Author-Topic Model for Authors and Documents*

Journals: Not documented

Authors: Not documented

Abstract:

We introduce the author-topic model, a generative model for documents that extends Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) to include authorship information. Each author is associated with a multinomial distribution over topics and each topic is associated with a multinomial distribution over words. A document with multiple authors is modeled as a distribution over topics that is a mixture of the distributions associated with the authors. We apply the model to a collection of 1,700 NIPS conference papers and 160,000 CiteSeer abstracts. Exact inference is intractable for these datasets and we use Gibbs sampling to estimate the topic and author distributions. We compare the performance with two other generative models for documents, which are special cases of the author-topic model: LDA (a topic model) and a simple author model in which each author is associated with a distribution over words rather than a distribution over topics. We show topics recovered by the author-topic model, and demonstrate applications to computing similarity between authors and entropy of author output.

Reviewers:

We found 9 searchers we think may review your paper.
You are required to evaluate them all.
When finished, you'll be able to recompute the result.

You are currently evaluating the result 1 / 9

Tao Dai

Affiliation: School of Software Engineering Xi'an Jiaotong University Xi'an China

Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network

on Journal of Ambient Intelligence and Humanized Computing

Abstract

Citation recommendation is the task of suggesting a list of references for an author given a manuscript. This is important for academic research for it provides an efficient and easy way to find relevant literatures. In this paper, we propose a novel probabilistic topic model to automatically recommend

...

Objectifs scientifiques

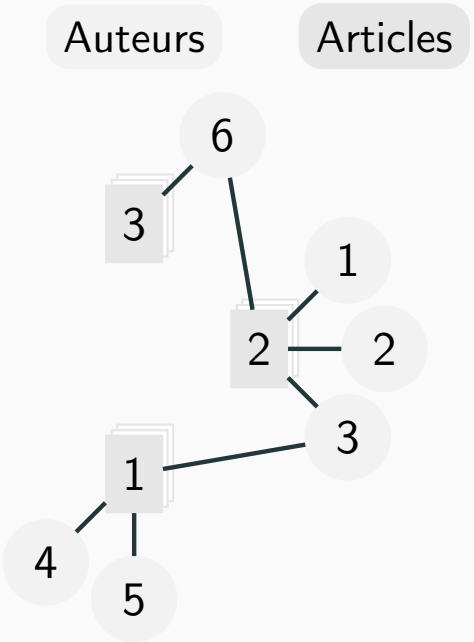
- Comment modéliser les réseaux pour la recommandation ? L'apprentissage de représentation [Bengio et al., 2013] permet-il cela ?
- Comment allier graphes et textes ? Comment ces deux domaines *a priori* distincts peuvent intéragir ?
- Peut-on appliquer l'apprentissage de représentation dans les réseaux de documents à la recherche d'expert [Balog et al., 2012] ?

Introduction

Notions générales

Les graphes

- Un graphe $G = (V, E)$ où V, l'ensemble des n_v sommets, et E, l'ensemble des n_e liens.
- A la matrice d'adjacence du graphe où $a_{ij} > 0$ ssi il existe un lien entre v_i et v_j.



$$A = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$D = \begin{bmatrix} \text{"Learning Framework..."} \\ \text{"Supervised Machine..."} \\ \text{"Scholarly Literature..."} \end{bmatrix}$$

Le traitement du langage naturel

Corpus

d_1 Vous avez des bagages Simon ?

d_2 Oh ben non. On m'a dit de venir, pas de venir avec des bagages ! Pourquoi ? Il fallait que j'en prende ?

extrait d'un dialogue du film « La Cité de la peur » (1994)

Vocabulaire

indice	*token*
1	bagages
2	dit
3	fallait
4	pourquoi
5	prende
6	simon
7	venir

Séquences

s_1 | 1 | 6 |

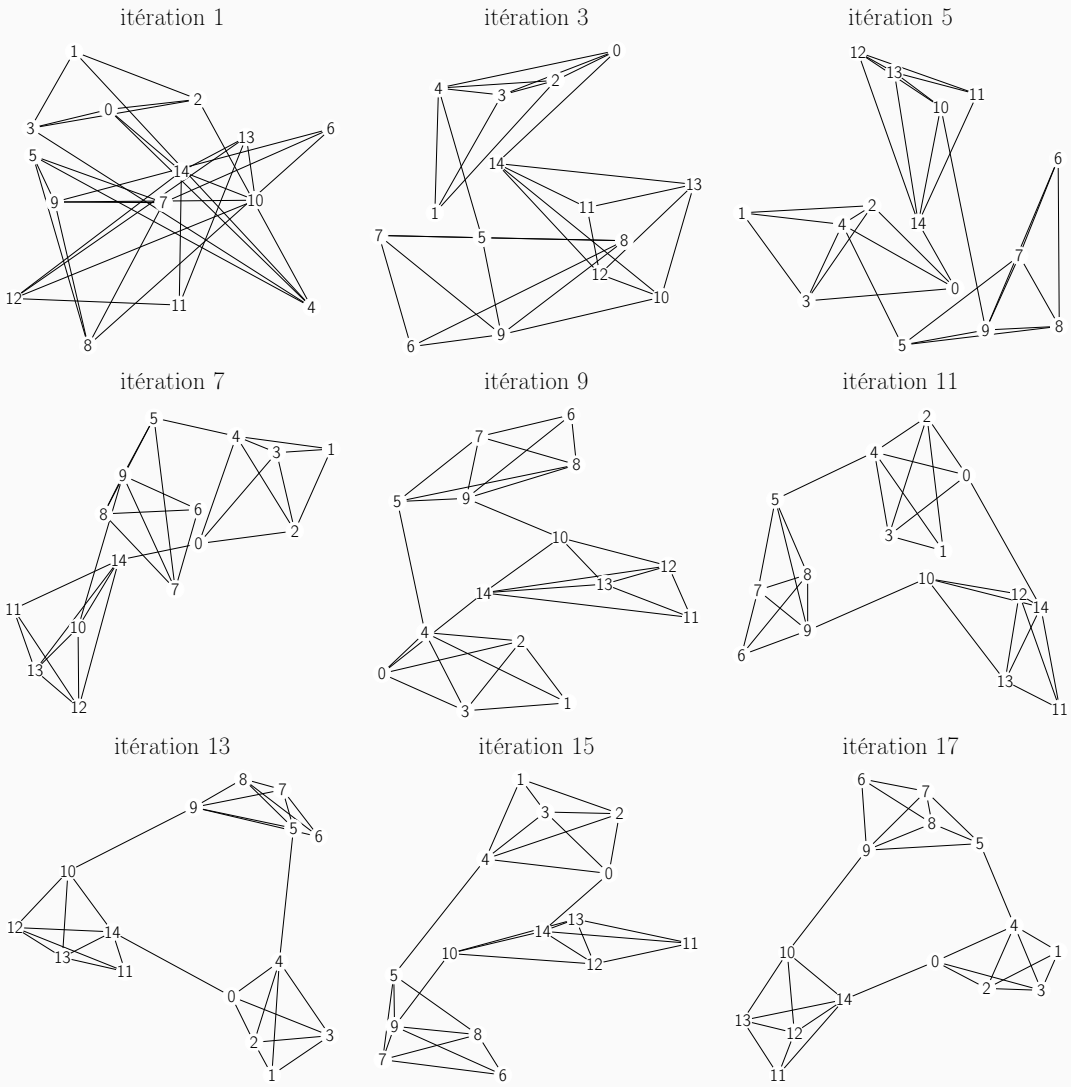
s_2 | 2 | 7 | 7 | 1 | 4 | 3 | 5 |

Matrice document-terme

	bagages	*dit*	*fallait*	*pourquoi*	*prende*	*simon*	*venir*	
	1	0	0	0	0	1	0	x_1
	1	1	1	1	1	0	2	x_2

Un calcul de similarité entre d_1 et d_2 : $\cos(\theta) = \frac{x_1 \cdot x_2}{||x_1|| ||x_2||}$

L'apprentissage de représentation



Algorithme de Fruchterman-Reingold [Fruchterman and Reingold, 1991].

Introduction

Jeux de données

Jeux de données

	# sommets	# liens	# labels	# mots	# mots par doc	densité	multi-*label*
Cora	2 211	4 771	7	4 333	67 (32)	0.20%	non
CiteSeer	3 312	4 551	6	3 505	31 (5)	0.08%	non
NYT	5 135	3 050 513	4	5 748	24 (17)	23.14%	non
Wikipedia	4 777	92 295	40	-	-	0.81%	oui
PPI	3 890	37 845	50	-	-	0.50%	oui
Gaming SE	22 872	400 664	40	15 760	53 (74)	0.15%	oui
Travel SE	15 087	465 696	60	14 539	70 (73)	0.41%	oui
Stats SE	14 834	283 885	59	15 421	103 (119)	0.26%	oui
Academia SE	20 799	622 720	55	14 034	76 (67)	0.29%	oui

Extrait d'une question posée sur `academia.stackexchange.com` **:**

Titre	How do I find the ML model referenced in a specific paper when the authors don't provide it ?
Question	I'm trying to locate the architecture of an LSTM neural network as described here, but the authors don't provide a repository for other academics. From arXiv, I see more than a dozen other researchers have referenced this paper and specifically utilized the same model architecture but it's unclear to me how they got access to the python source code. I've tried emailing some researchers but none of them have responded. Does anyone have any suggestions ?

Introduction

Plan

Plan

1. Apprentissage de Représentation des Sommets d'un Réseau
2. Apprentissage de Représentation de Documents en Réseau
3. Cas d'Application : la Recherche d'Experts
4. Conclusion et Perspectives

Apprentissage de Représentation des Sommets d'un Réseau

Apprentissage de Représentation des Sommets d'un Réseau

Introduction

Introduction

- Nombre de systèmes complexes prennent la forme de réseaux.
- L'encodage d'un graphe ne permet pas d'appliquer les algorithmes traditionnels en apprentissage automatique.

⇒ l'apprentissage de représentation :

permet d'aborder des problématiques sur les réseaux avec des méthodes d'apprentissage automatique traditionnelles éprouvées.

Objectifs

⇒ le plongement de réseau :

construire des représentations vectorielles denses, continues et de faible dimension des sommets d'un graphe.

Certains critères à respecter :

- une faible complexité algorithmique ;
- une forte capacité de parallélisation ;
- une certaine capacité de réutilisation.

Description du problème

- $G = (V, E)$ composé de n_v sommets.
- Représenter chaque sommet v_i du graphe par un vecteur u_i de dimension $\rho_v \ll n_v$, de sorte que les distances dans l'espace d'*embedding* conservent le voisinage (local et global) des sommets dans le graphe.

Apprentissage de Représentation des Sommets d'un Réseau

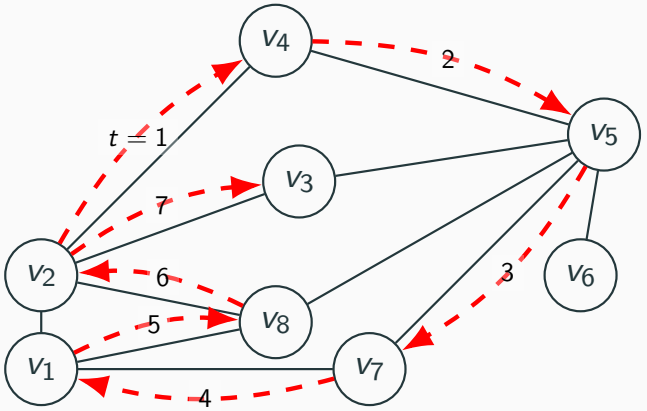
État de l'art

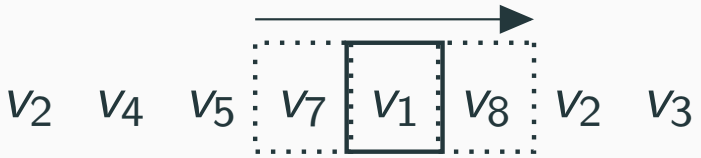
Parcours de graphe

- A : matrice d'adjacence de $G = (V, E)$ (binaire pour l'exemple).
- A^2 : nombres exactes de voisins en commun pour chaque paire de sommets (proximité d'ordre 2).
- A^ℓ : similarité selon les chemins de longueur ℓ.
- Les marches aléatoires utilisées dans de nombreux domaines [Page et al., 1999, Andersen et al., 2006, Fouss et al., 2007].

DeepWalk [Perozzi et al., 2014]

1)





$$C = \begin{array}{cc} \text{cible} & \text{contexte} \\ \left[\begin{array}{cc} v_2 & v_4 \\ v_4 & v_2 \\ \dots & \dots \\ v_2 & v_3 \\ v_3 & v_2 \end{array}\right. & \end{array}$$

2) *Skip-Gram with Negative Sampling (SGNS)*
[Mikolov et al., 2013a, Mikolov et al., 2013b]

$$\sum_{(v_i,v_j)\in C} \Big(\log \sigma(h_j \cdot u_i) + \sum_{k=1}^{n_k} \mathbb{E}_{v_k \sim q(v_k)} \big[\log \sigma(-h_k \cdot u_i) \big] \Big).$$

Apprentissage de Représentation des Sommets d'un Réseau

Global Vectors for Node Representation

Idées derrière GVNR [Brochier et al., 2019a]

- S'inspirer de GloVe [Pennington et al., 2014] plutôt que de SGNS pour le plongement de réseau.
- Factorisation d'une matrice S des comptages des co-occurrences des sommets.
- Seuillage qui diminue la complexité en temps de l'algorithme.
- Prise en compte de valeurs nulles améliorant les représentations.

Matrice factorisée

μ marches de longueurs ℓ sont opérées et une fenêtre de taille τ incrémente les co-occurrences de sommets dans la matrice $S \in \mathbb{R}^{n_v \times n_v}$ en utilisant une pondération $\lambda = \frac{1}{q}$. Les valeurs de S inférieures à s_{min} sont ensuite mises à zéros.

$$S = \begin{pmatrix} 0 & 4.2 & 8.5 & 0.5 & 0 & 0 \\ 4.2 & 0 & 0 & 3.1 & 0.25 & 0.5 \\ 8.5 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 3.1 & 0 & 0 & 0 & 6 \\ 0 & 0.25 & 0 & 0 & 0 & 1 \\ 0 & 0.5 & 0 & 6 & 1 & 0 \end{pmatrix} \xrightarrow{\text{seuillage } s_{\text{min}}} \begin{pmatrix} 0 & 4.2 & 8.5 & 0 & 0 & 0 \\ 4.2 & 0 & 0 & 3.1 & 0 & 0 \\ 8.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.1 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 6 & 1 & 0 \end{pmatrix}$$

Factorisation

La factorisation s'écrit :

$$\underset{U,H,b^u,b^h}{\operatorname{argmin}} \sum_{i=1}^{n_v}\sum_{j=1}^{n_v} \underbrace{\delta(s_{ij})}_{\text{échantillonnage}} \Big(\overbrace{u_i \cdot h_j + b_i^u + b_j^h}^{\text{représentations et leur biais}} - \underbrace{\log(1+s_{ij})}_{\text{matrice factorisée}} \Big)^2, \quad (1)$$

avec

$$\delta(s_{ij}) = \begin{cases} 1 & \text{si } s_{ij} > 0, \\ m_i & \text{sinon, où } m_i \sim \text{Bernoulli}(\alpha_i). \end{cases} \quad (2)$$

En notant $p_i = \frac{|s_{i.}|_{\neq 0}}{n_v}$ la densité de la $i^{\text{ème}}$ ligne de S, α_i est calculé selon la cote :

$$\alpha_i = \begin{cases} \mathbf{n_k} \times \frac{p_i}{1-p_i} & \text{si } p_i \leq (n_k+1)^{-1}, \\ 1 & \text{sinon.} \end{cases} \quad (3)$$

Complexité

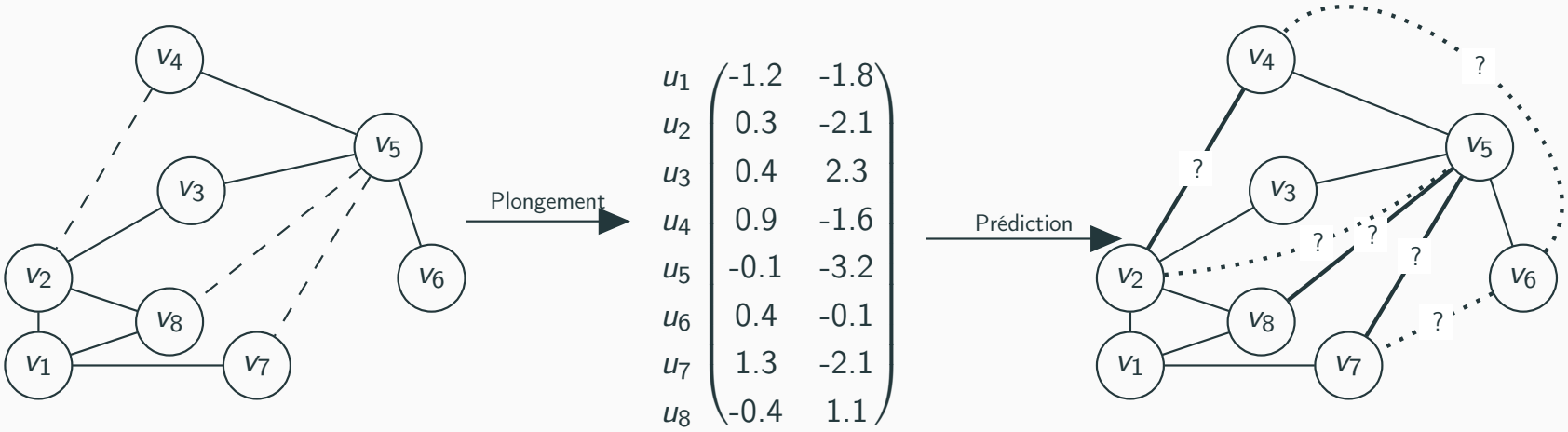
Comparaison des nombres de mises à jour des paramètres dans SGNS, GloVe et GVNR selon différents jeux de données. GVNR utilise $n_k = 1$ échantillon négatif et un seuillage s_{min}=1.

Jeu de données	n_v	n_e	SGNS	GloVe	GVNR
Cora	2 211	9 542	67 214 400	858 315	**793 326**
CiteSeer	3 312	9 102	100 684 800	**431 660**	493 024
NYT	5 135	6 101 026	156 104 000	**6 106 114**	12 205 092
Wikipedia	4 777	184 590	145 220 800	10 158 398	**5 741 006**
PPI	3 890	75 690	118 256 000	8 055 838	**7 230 492**
Gaming SE	22 872	801 328	695 308 800	12 259 288	**7 754 520**
Travel SE	15 087	931 392	458 644 800	9 921 904	**5 752 762**
Stats SE	14 834	567 770	450 953 600	11 436 290	**7 376 892**
Academia SE	20 799	1 245 440	632 289 600	25 544 057	**13 129 272**

Apprentissage de Représentation des Sommets d'un Réseau

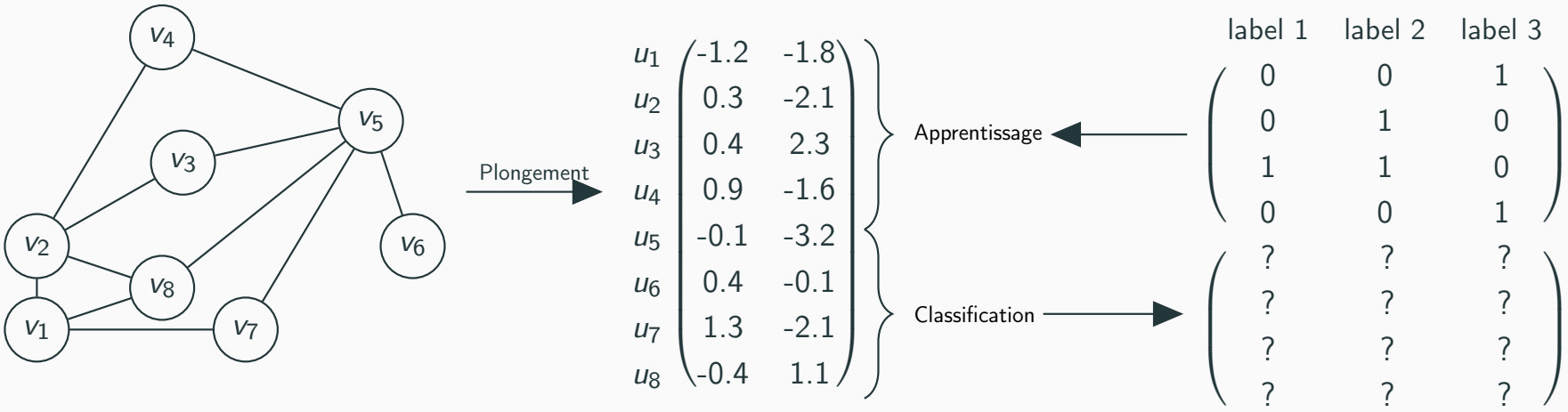
Évaluation

Prédiction de liens



Protocole d'évaluation utilisé pour évaluer GVNR sur la prédiction de liens.

Classification de sommets



Protocole d'évaluation utilisé pour évaluer GVNR sur la classification des sommets.

Résultats expérimentaux

Cora	Classification										Prédiction
	F1					AUC					AUC
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	50%
Aléatoire	14.44	13.62	13.85	14.26	13.18	52.61	51.71	51.24	51.60	51.59	51.24
NetMF	68.22	74.97	77.60	79.80	81.05	89.38	92.81	93.77	94.70	95.69	66.53
DeepWalk	74.09	76.98	78.53	78.36	79.82	94.21	95.20	95.70	95.80	96.13	70.50
GloVe	65.01	65.82	68.95	68.68	69.64	90.42	90.78	92.37	92.52	93.34	71.97
GVNR ($s_{\text{min}} = 0$)	74.04	77.12	78.64	78.60	79.86	94.07	95.22	95.88	96.04	96.39	**78.10**
GVNR ($s_{\text{min}} = 1$)	**76.01**	**78.50**	**80.13**	**80.54**	**81.45**	**95.12**	**95.99**	**96.41**	**96.55**	**96.90**	76.26

Travel SE	Classification										Prédiction
	F1					AUC					AUC
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	50%
Aléatoire	7.25	6.89	6.67	6.91	6.61	58.47	54.14	53.66	53.20	52.47	52.03
NetMF	10.12	11.57	13.02	13.88	13.88	58.16	63.12	64.00	66.23	66.68	97.99
DeepWalk	**12.39**	**13.28**	13.42	**14.90**	**15.16**	59.54	**63.45**	**66.48**	**67.86**	**68.22**	97.44
GloVe	8.85	9.85	10.44	10.32	10.61	53.62	55.60	59.64	62.41	63.70	98.82
GVNR ($s_{\text{min}} = 0$)	9.96	11.50	12.13	13.17	13.54	58.31	61.62	63.95	65.13	65.65	**98.94**
GVNR ($s_{\text{min}} = 1$)	9.77	12.56	**13.48**	14.57	14.45	**63.64**	62.99	63.55	65.85	65.81	98.82

Résumé des résultats

- GVNR réalise généralement de meilleurs scores que GloVe, sauf sur PPI et Wikipedia.
- Le seuillage de la matrice S apporte constamment une nette amélioration des représentations apprises.
- En prédiction de liens, GVNR produit de meilleurs scores sur l'ensemble des jeux de données.

Apprentissage de Représentation des Sommets d'un Réseau

Conclusions et perspectives

Conclusion et perspectives

- GVNR constitue une bonne alternative aux algorithmes de l'état de l'art, souvent basés sur SGNS.
- Réduction de la complexité en temps de GloVe avec gain de performances s'appuyant sur :
  - un seuillage de la matrice S ;
  - un échantillonnage différent des valeurs de S.
- Peut-on étendre facilement ce modèle pour les réseaux de documents ?
- Un avantage est la construction explicite de S. Peut-on la construire différemment ? Pour capturer les rôles structurels des sommets ?

Apprentissage de Représentation de Documents en Réseau

Apprentissage de Représentation de Documents en Réseau

Introduction

Introduction

- Le plongement de réseau permet d'aborder efficacement des tâches telles que la classification des sommets et la prédiction de liens.
- Les articles scientifiques sont structurés.

⇒ Plongement de documents en réseau :

comment combiner ces deux sources d'information (graphe et texte) ?

Objectifs

Plusieurs objectifs possibles :

- l'enrichissement des représentations (GVNR-t) ;
- l'apprentissage d'un modèle du langage (IDNE - induction) ;
- l'interprétation (GVNR-t et IDNE) ;
- l'explicabilité (MATAN [Brochier et al., 2019b]).

Description du problème

- $G = (V, E, X)$ composé de n_v sommets/documents, où $X \in \mathbb{R}^{n_d \times n_\omega}$ matrice des termes de $n_d = n_v$ documents construite à partir d'un vocabulaire de n_ω mots.
- Représenter chaque document d_i du graphe par un vecteur u_i de dimension ρ_d tel que $\rho_d \ll n_d$ et $\rho_d \ll n_\omega$.

Apprentissage de Représentation de Documents en Réseau

État de l'art

Text-Associated DeepWalk

- TADW (*Text-Associated DeepWalk*) [Yang et al., 2015] constitue le modèle de référence en plongement de documents en réseau.
- Le modèle étend DeepWalk, qui réalise implicitement la factorisation d'une matrice [Levy and Goldberg, 2014] $M \simeq \frac{A+A^2}{2}$, en une trifactorisation intégrant les attributs X des sommets.
- $\underset{U,H}{\operatorname{argmin}} \left\| M - U^\top HX \right\|_F^2 + \frac{\lambda}{2} \left(\|U\|_F^2 + \|H\|_F^2 \right)$.

Apprentissage de Représentation de Documents en Réseau

Contributions

GVNR-t [Brochier et al., 2019a]

- GVNR construit des représentations U et H de dimension $n_v \times \rho$ en factorisant S construite par marches aléatoires.
- Hypothèse : le sens d'un document d_j peut être capturé par la moyenne de ses vecteurs mots $x_j W$ où $W \in \mathbb{R}^{n_\omega \times \rho}$.

$$\underset{U,W,b^u,b^h}{\operatorname{argmin}} \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \delta(s_{ij}) \big(u_i \cdot \underbrace{\mathbf{x_j W}}_{\text{contexte, modélisé par } h_j \text{ dans GVNR}} + b_i^u + b_j^h - \log(1 + s_{ij}) \big)^2.$$

L'attention en NLP (Transformer [Vaswani et al., 2017])

Phrase	self	attention	mechanism
Plongements	$x_1 \in \mathbb{R}^{\rho_\omega}$	$x_2 \in \mathbb{R}^{\rho_\omega}$	$x_3 \in \mathbb{R}^{\rho_\omega}$
Requêtes	$q_1 = x_1 W^q$	$q_2 = x_2 W^q$	$q_3 = x_3 W^q$
Clefs	$k_1 = x_1 W^k$	$k_2 = x_2 W^k$	$k_3 = x_3 W^k$
Valeurs	$v_1 = x_1 W^v$	$v_2 = x_2 W^v$	$v_3 = x_3 W^v$
Scores	$q_1 \cdot k_1 = 29.61$	$q_1 \cdot k_2 = 31.02$	$q_1 \cdot k_3 = 26.79$
Normalisation ($\sqrt{\rho_w}$)	20.94	21.93	18.94
Poids (softmax)	$\alpha_{12} = 0.26$	$\alpha_{12} = 0.70$	$\alpha_{13} = 0.04$
Pondérations	$0.26 \times v_1$	$0.70 \times v_2$	$0.04 \times v_3$
Sommes	$y_1 = \sum_{i=1}^{3} \alpha_{1i} v_i$	$y_2 = \sum_{i=1}^{3} \alpha_{2i} v_i$	$y_3 = \sum_{i=1}^{3} \alpha_{3i} v_i$

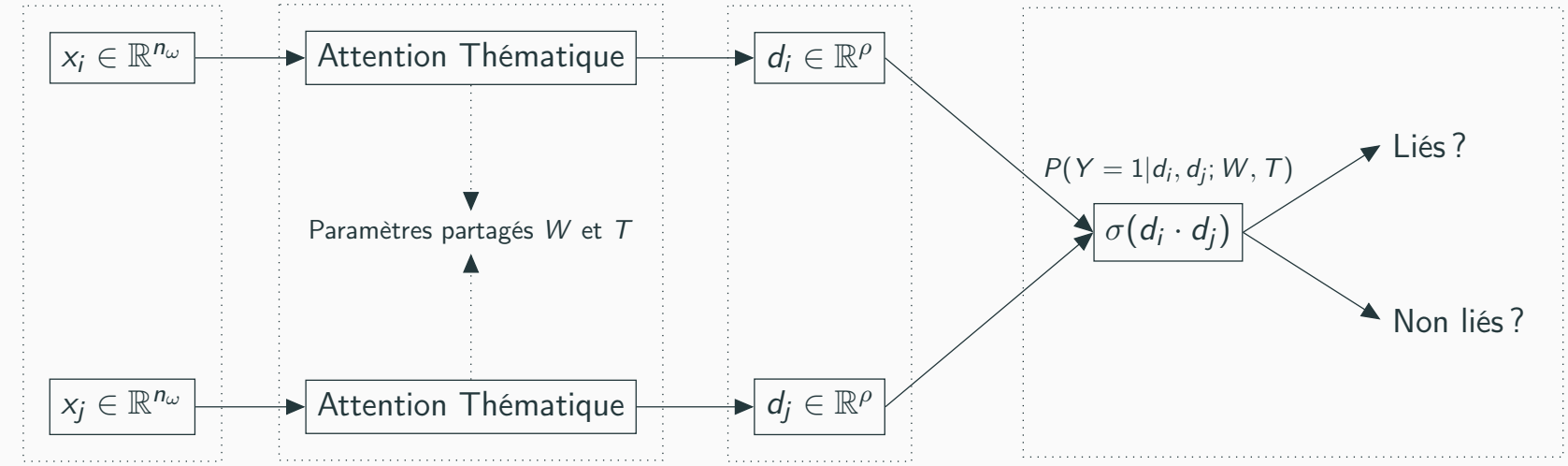
IDNE [Brochier et al., 2020b]

- Un problème avec GVNR-t est son caractère transductif.
- L'induction est une caractéristique appréciée pour la résolution du problème de « démarrage à froid » [Schein et al., 2002].
- L'interprétabilité [Goebel et al., 2018] devient un enjeu important pour de nombreux systèmes de recherche d'information.

⇒ IDNE (*Inductive Document Network Embedding*) :

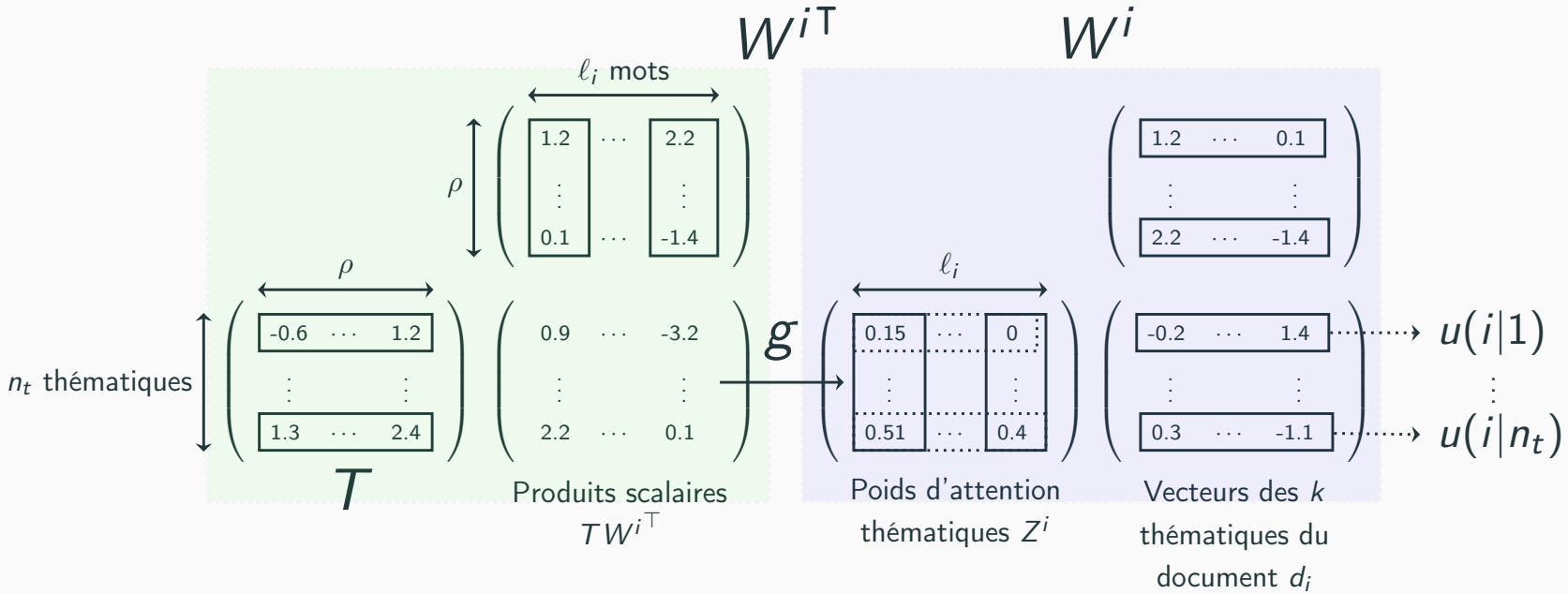
- inductif ;
- apprentissage à partir du réseau ;
- thématiques latentes interpretables.

Vue d'ensemble



L'attention thématique

Inspirée par le *Set Transformer* [Lee et al., 2019].



$$u_i = \sum_k^{n_t} \frac{u(i|k)}{|x_i|_1}.$$

Construction de $u_i = f(T, W^i)$ par le mécanisme d'attention thématique.

Apprentissage guidé par le réseau

- Estimation de W et T en exploitant les liens entre les documents.
- Matrice binaire S :

$$s_{ij} = \begin{cases} 1 & \text{si } (A + A^2)_{ij} > 0, \\ 0 & \text{sinon.} \end{cases}$$

- La probabilité de deux documents, selon le modèle, d'être liés :

$$P(Y = 1 | u_i, u_j; W, T) = \sigma(u_i \cdot u_j).$$

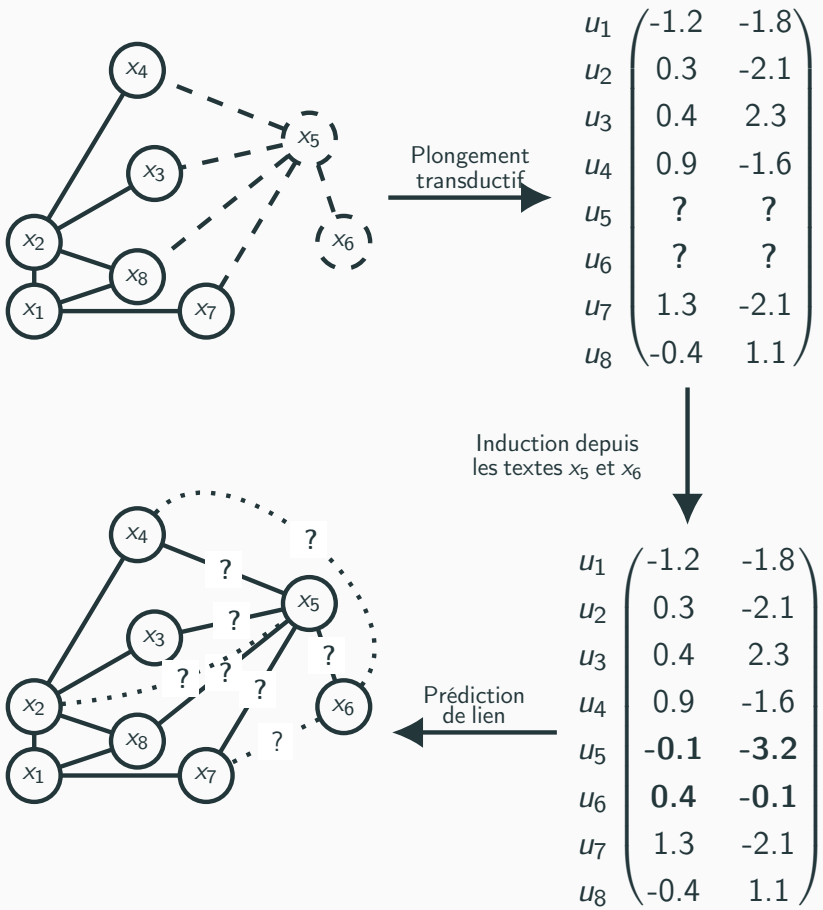
- Minimisation de la vraisemblance de S étant donnés W, T :

$$L(W, T) = \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} s_{ij} \log \sigma(u_i \cdot u_j) + (1 - s_{ij}) \log \sigma(-u_i \cdot u_j).$$

Apprentissage de Représentation de Documents en Réseau

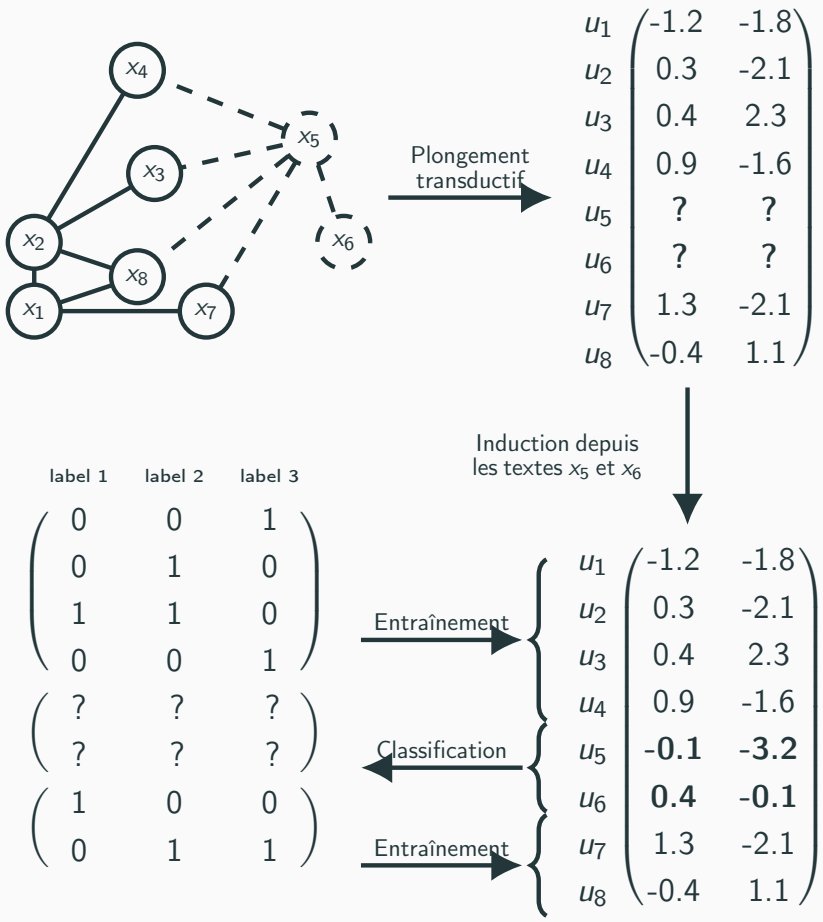
Évaluation

Prédiction inductive de liens



Évaluation des plongements de réseau de documents pour la prédiction inductive de liens.

Classification inductive de sommets



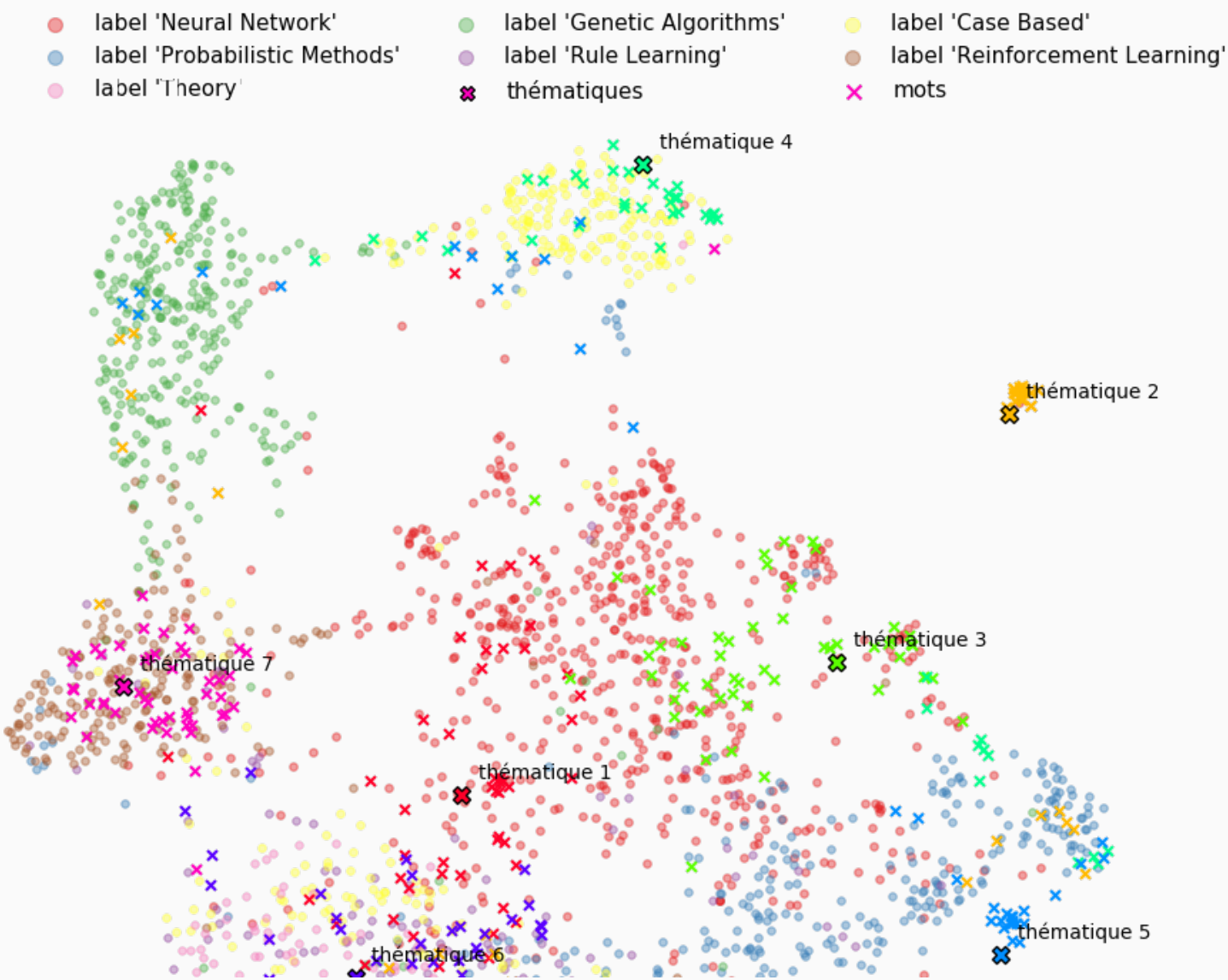
Évaluation des plongements de réseau de documents pour la classification inductive des sommets.

Résultats expérimentaux

Cora	TC										IC		TP	IP
	F1					AUC					F1	AUC	AUC	AUC
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	90%	90%	50%	90%
Aléatoire	13.85	13.70	12.66	12.83	12.53	51.51	51.96	51.60	51.22	51.22	11.80	48.84	49.87	50.27
TF	68.84	74.69	77.20	78.90	80.05	92.13	94.63	95.64	96.06	96.55	81.98	97.16	83.44	84.88
TF-IDF	72.56	76.99	80.13	80.66	81.84	93.11	95.24	96.38	96.74	97.20	**83.24**	**97.59**	85.17	85.02
LSA	72.46	77.22	79.92	80.56	81.16	93.65	95.47	96.63	96.95	97.17	81.26	97.35	87.17	88.63
DW+LSA	77.52	80.70	83.40	83.12	85.21	95.31	96.47	97.24	97.23	97.79	-	-	82.67	-
TADW	61.77	67.02	70.84	71.62	73.13	89.01	91.25	93.22	93.37	94.20	79.55	96.35	81.59	84.71
G2G	79.38	82.19	83.57	84.08	85.03	96.14	97.39	97.65	97.89	98.13	71.98	94.57	86.36	74.72
MATAN	75.48	76.65	77.58	79.20	78.30	95.19	95.79	96.23	96.41	96.69	76.13	94.91	82.72	71.47
GVNR-t	**82.20**	83.49	**85.26**	**85.82**	**86.67**	97.10	97.53	98.00	98.08	**98.48**	79.91	97.21	**94.31**	**92.44**
IDNE	80.41	**83.83**	84.79	84.67	86.06	**97.27**	**97.79**	**98.28**	**98.23**	98.45	82.25	97.57	92.90	88.56

Stats SE	TC										IC		TP	IP
	F1					AUC					F1	AUC	AUC	AUC
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	90%	90%	50%	90%
Aléatoire	7.15	6.57	6.34	6.33	6.27	60.08	55.65	53.05	52.58	53.09	6.11	52.05	50.28	49.85
TF	44.73	51.09	53.49	54.63	56.33	79.06	83.43	87.04	87.16	87.74	58.04	88.81	67.13	67.31
TF-IDF	38.29	50.49	53.70	54.90	**57.74**	72.11	74.10	79.20	80.82	83.28	**61.18**	86.24	69.61	67.71
LSA	51.05	52.10	51.57	51.84	51.42	82.19	85.88	85.62	85.07	84.54	53.54	86.02	68.10	67.04
DW+LSA	41.11	48.78	50.72	51.98	53.89	70.68	76.90	81.23	81.34	82.60	-	-	98.00	-
TADW	**54.50**	**55.11**	**55.95**	**55.19**	54.88	82.09	85.17	86.39	85.18	85.69	44.79	89.88	60.46	62.59
G2G	29.58	32.11	33.70	35.06	35.71	83.59	86.87	87.85	88.41	89.16	32.23	82.69	97.79	68.78
MATAN	34.38	36.67	37.32	39.00	38.46	87.26	89.15	90.24	**90.58**	91.09	41.17	**93.22**	80.28	**72.57**
GVNR-t	19.51	23.15	25.55	28.04	29.54	68.53	70.72	75.49	73.65	77.37	32.89	81.08	**98.94**	69.18
IDNE	43.26	43.96	44.85	44.57	45.03	**89.20**	**90.09**	**90.73**	**90.58**	**91.16**	44.49	92.29	74.72	70.80

Visualisation avec l'algorithme UMAP [McInnes et al., 2018]



Thématiques apprises par IDNE

Thématique	Mots les plus proches
1	*pattern*, *features*, *classification*, *regulate*, *machine*, *limitations*, *domains*, *datasets*, methods, patterndirected, dataset, connectionist, backpropagation, pythia, accuracy, classifier, realworld, exemplar, symbolic, links
2	tree, decision, selection, trees, markov, radar, subsurface, moisture, cband, ers, cedar, sar, lband, polarized, minnesota, lter, ersjers, creek, raco, seasonality
3	*saturation*, *separation*, *stabilization*, *gradient*, *blind*, *asymptotic*, *universal*, *matrix*, *square*, *signals*, necessarily, sign, projection, approximated, bandpass, descent, norm, cubic, subproblems, speakers
4	design, casebased, pac, sme, analogical, mcmc, structuremapping, planning, wellunderstood, retrieval, reasoning, mistakes, case, reuse, retrieving, instance, chain, analogy, convolution, solving
5	mcmc, execution, speculation, ga, parallelism, genetic, instruction, aimed, dirichlet, instructions, chain, consensus, substantive, sequences, macroscopic, issue, processor, recipient, analogy, nerve
6	accuracy, learning, experiments, machine, ilp, datasets, comprehensibility, sufficed, sbc, inductive, accurate, decompose, warehouses, mining, induction, dataset, gratefully, attribute, assign, challenges
7	*reinforcement*, *mdp*, *mdps*, *qlearning*, *policy*, *crosses*, *actions*, *value*, *reward*, *policies*, brigade, rl, macros, relax, hinders, dynamic, satisficingoptimizing, action, functionings, barto
Plus élevées	genetic (13.72), reinforcement (13.32), evolutionary (11.08), mdp (10.36), bayesian (10.17), mcmc (9.87)
Plus basses	counterexamples (0.19), lecun (0.19), it (0.19), epochs (0.19), shavlik (0.19), cdc (0.18)

Résumé des résultats

- GVNR-t constitue une approche efficace de plongement de réseau de documents, sauf pour la classification inductive.
- IDNE est efficace sur toutes les tâches et sur tous les jeux de données.
- IDNE produit des représentations interprétables des mots, thématiques et documents qui corrèlent avec les thématiques vérités terrains sur Cora.

Apprentissage de Représentation de Documents en Réseau

Conclusions et perspectives

Conclusion et perspectives

- GVNR-t, IDNE et MATAN permettent d'aborder des tâches transductives et inductives de classification et de prédiction.
- IDNE peut être utilisé comme outil d'exploration de réseau de documents.
- Peut-on utiliser ce mécanisme d'attention sur des données textuelles uniquement ?
- Quel est l'impact de la construction de S sur les thématiques apprises ?

Cas d'Application : la Recherche d'Experts

Cas d'Application : la Recherche d'Experts

Introduction

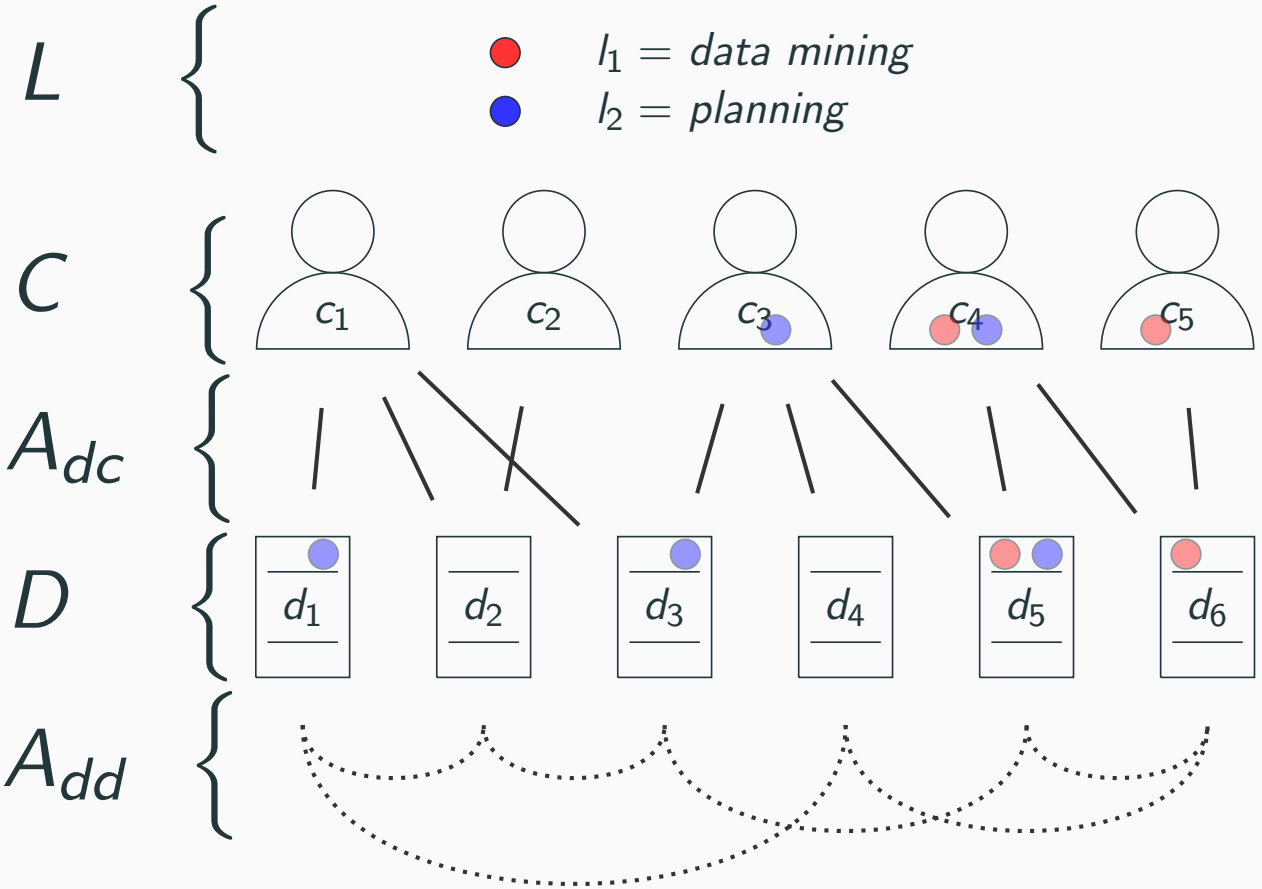
Motivations

- *Peerus Review* : un outil d'aide à la recherche de relecteurs scientifiques pour les maisons d'édition [Brochier et al., 2018b].
- Mettre en place un protocole d'évaluation adapté à « notre » tâche de recherche d'experts [Brochier et al., 2020a].
- Appliquer le plongement de réseau à une problématique de recommandation plus complexe que les tâches de classification et de prédiction usuelles.

Cas d'Application : la Recherche d'Experts

Contributions

Une nouvelle évaluation



De nouveaux jeux de données

Propriétés générales des jeux de données pour la recherche d'experts.

	# candidats	# documents	# expertises	# experts	# requêtes	exemple d'expertise
DBLP	707	1641	7	199	114	'information_extraction'
Stats SE	5765	14834	59	5765	3966	'maximum-likelihood'
Academia SE	6030	20799	55	6030	4214	'recommendation-letter'
Math Overflow	7382	38532	98	7382	10614	'galois-representations'

Cas d'Application : la Recherche d'Experts

Expérimentations

Résultats expérimentaux

Scores moyens et leurs écarts-types sur Math Overflow suivant la méthode d'évaluation avec requêtes documents.

	AUC	P@10	AP
Aléatoire	49.98 (01.62)	06.44 (08.28)	06.53 (03.06)
P@noptic (TF-IDF)	81.87 (04.46)	21.95 (19.15)	22.95 (07.54)
Vote (TF-IDF)	86.80 (03.23)	61.11 (18.68)	40.10 (08.27)
Propagation (TF-IDF)	88.08 (03.38)	**93.68** (12.16)	**49.58** (08.90)
Pré-agr TADW	-	-	-
Pré-agr GVNR-t	65.34 (09.22)	44.02 (28.31)	16.88 (08.55)
Pré-agr G2G	66.84 (08.99)	22.95 (17.81)	12.49 (05.70)
Pré-agr IDNE	67.01 (09.26)	22.96 (17.84)	13.40 (06.02)
Post-agr TADW	-	-	-
Post-agr GVNR-t	63.84 (07.59)	41.81 (22.68)	14.96 (06.25)
Post-agr G2G	65.06 (09.09)	22.43 (16.94)	11.78 (05.51)
Post-agr IDNE	66.74 (09.10)	21.92 (17.21)	13.11 (05.87)
Vote (IDNE)	**88.71** (03.76)	68.46 (18.53)	43.53 (09.90)
Propagation (IDNE)	69.38 (09.65)	92.35 (13.88)	39.62 (09.89)

Résumé des résultats

- Notre méthodologie d'évaluation pour la recherche d'experts montre une plus grande stabilité que l'approche reposant sur les requêtes thématiques.
- Les algorithmes DNE ne peuvent être adaptés de manière triviale pour obtenir des scores suffisants.
- L'intégration de représentations apprises par un algorithme de DNE améliorent un modèle de vote mais peuvent altérer un modèle de propagation.

Cas d'Application : la Recherche d'Experts

Conclusions et perspectives

Conclusions et perspectives

- Les algorithmes de plongement de réseau ne capturent pas la notion de centralité, caractéristique importante en recherche d'experts.
- Nos méthodes de plongement de réseau de documents fonctionnent sur des réseaux homogènes. La prise en compte de l'hétérogénéité devrait pallier les scores décevants obtenus par nos méthodes d'aggrégation.

Conclusions et Perspectives

Conclusions et Perspectives

Résumé

Résumé

- GVNR, une nouvelle approche de plongement de réseau et son extension GVNR-t pour les réseaux de documents.
- IDNE et MATAN permettent de :
  - aborder des tâches transductives et inductives ;
  - interpréter en langage naturel les représentations construites ;
  - expliquer les liens structurant un corpus.
- Recherche d'experts :
  - nouveaux jeux de données ;
  - nouveau protocole d'évaluation ;
  - benchmark comprenant des algorithmes de plongement.

Conclusions et Perspectives

Perspectives

Perspectives

- Approfondir la notion d'apprentissage autosupervisé (*self-supervised learning*) dans les réseaux [Hu et al., 2019].
- Étudier plus en détail les liens entre les méthodes traditionnelles employées dans les systèmes de recommandation et les récentes avancées en plongement de réseau [Ying et al., 2018].
- Aborder l'encodage de la dynamique des réseaux (travaux en cours) [Bamler and Mandt, 2017, Rudolph and Blei, 2018].

Conclusions et Perspectives

Questions

Merci !

Tous les codes, jeux de données et protocoles d'évaluation présentés lors de cette soutenance sont publiquement accessibles sur `https://github.com/brochier`.

- **Brochier, R., Guille, A., and Velcin, J. (2020b).** Inductive document network embedding with topic-word attention.
  In ***Proceedings of the 42nd European Conference on Information Retrieval Research*****. Springer**

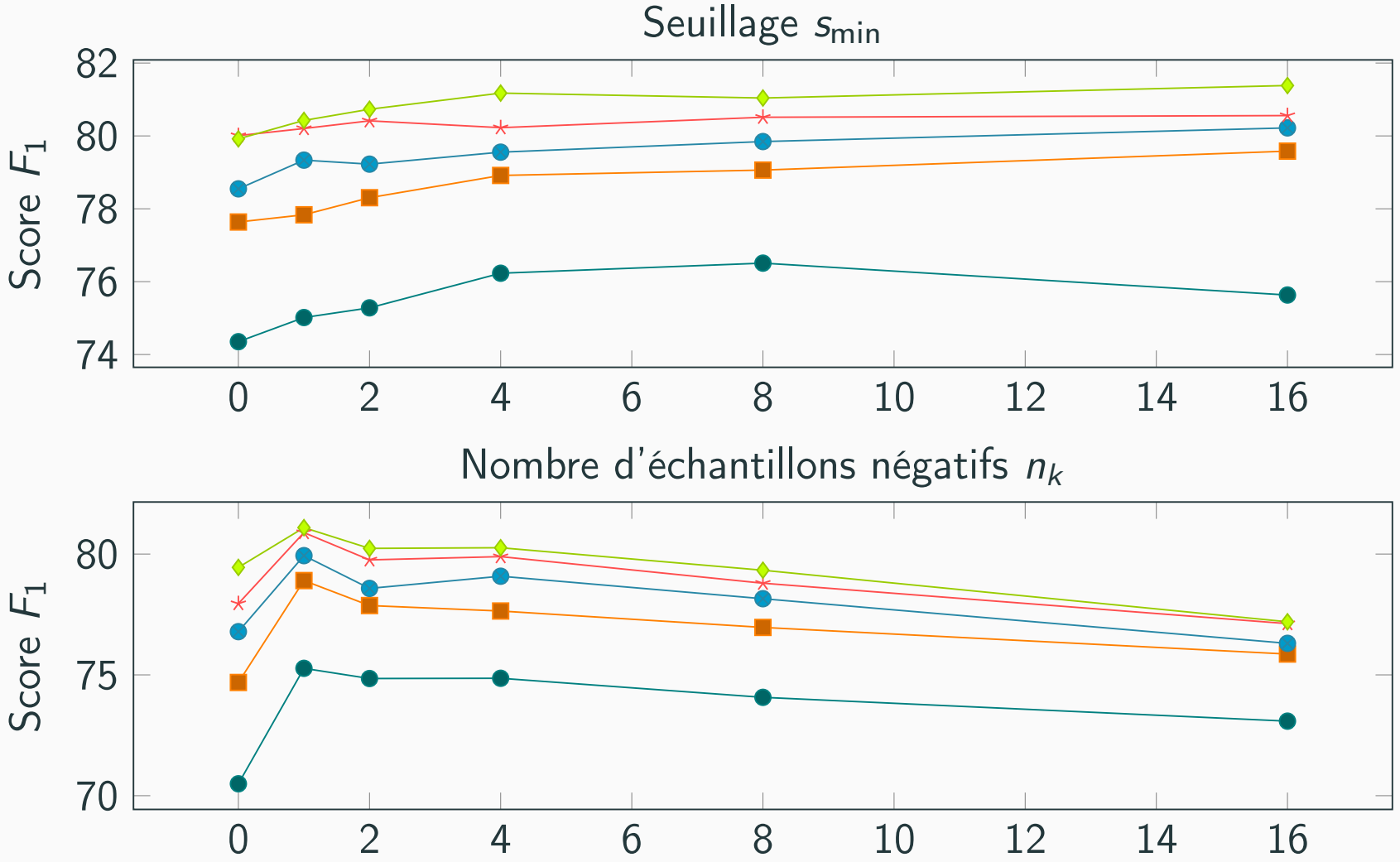
- **Brochier, R., Gourru, A., Guille, A., and Velcin, J. (2020a).** New datasets and a benchmark of document network embedding methods for scientific expert finding.
  In ***Bibliometric-enhanced Information Retrieval: 10th International BIR Workshop at ECIR***

- **Brochier, R., Guille, A., and Velcin, J. (2019a).** Global vectors for node representations.
  In ***Proceedings of The 2019 World Wide Web Conference*****. International World Wide Web Conferences Steering Committee**

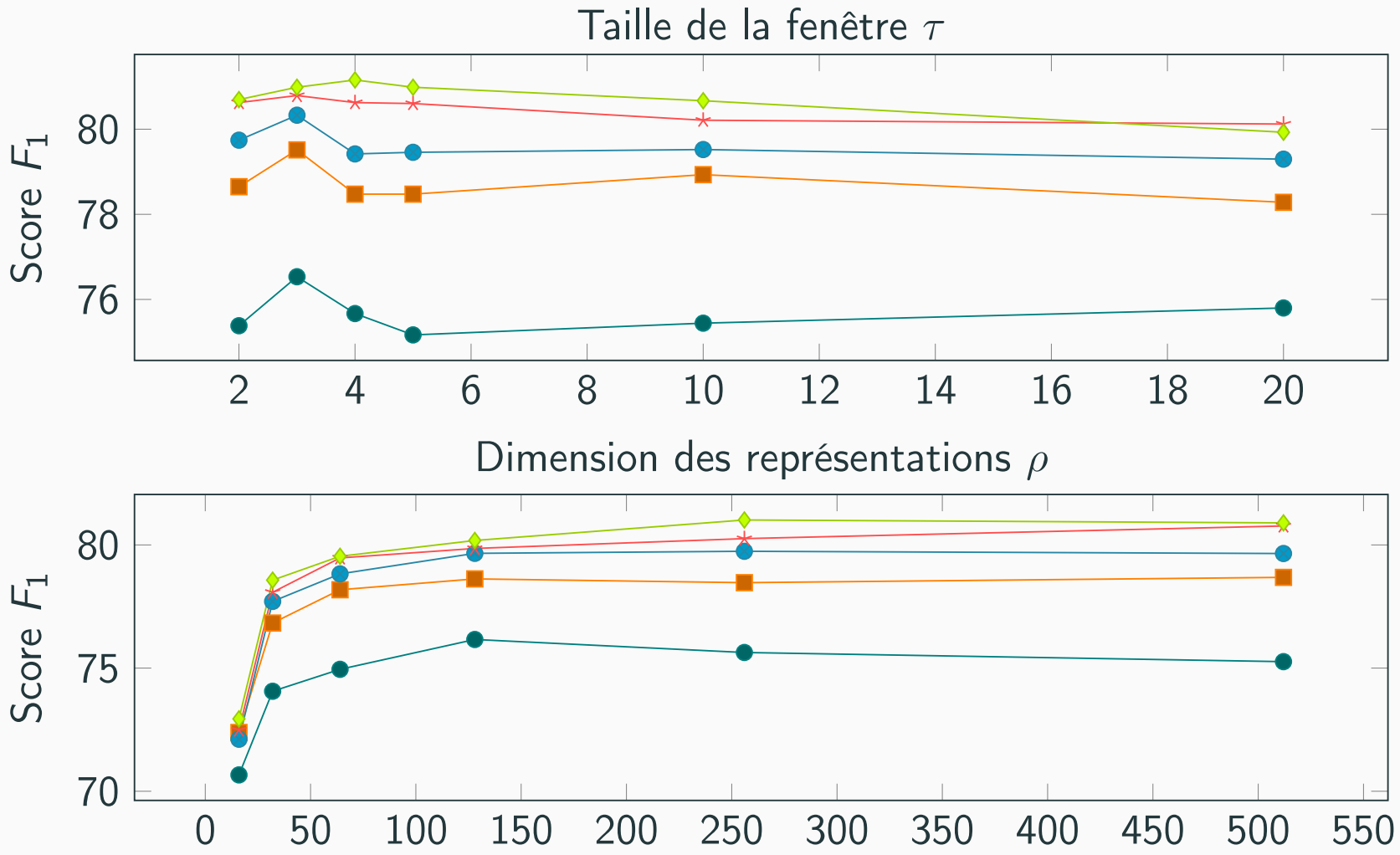
- **Brochier, R., Guille, A., and Velcin, J. (2019b).** Link prediction with mutual attention for text-attributed networks.
  In ***Companion Proceedings of The 2019 World Wide Web Conference*****. International World Wide Web Conferences Steering Committee**

- **Brochier, R., Guille, A., Rothan, B., and Velcin, J. (2018a).** Impact of the query set on the evaluation of expert finding systems.
  In ***3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) at SIGIR***

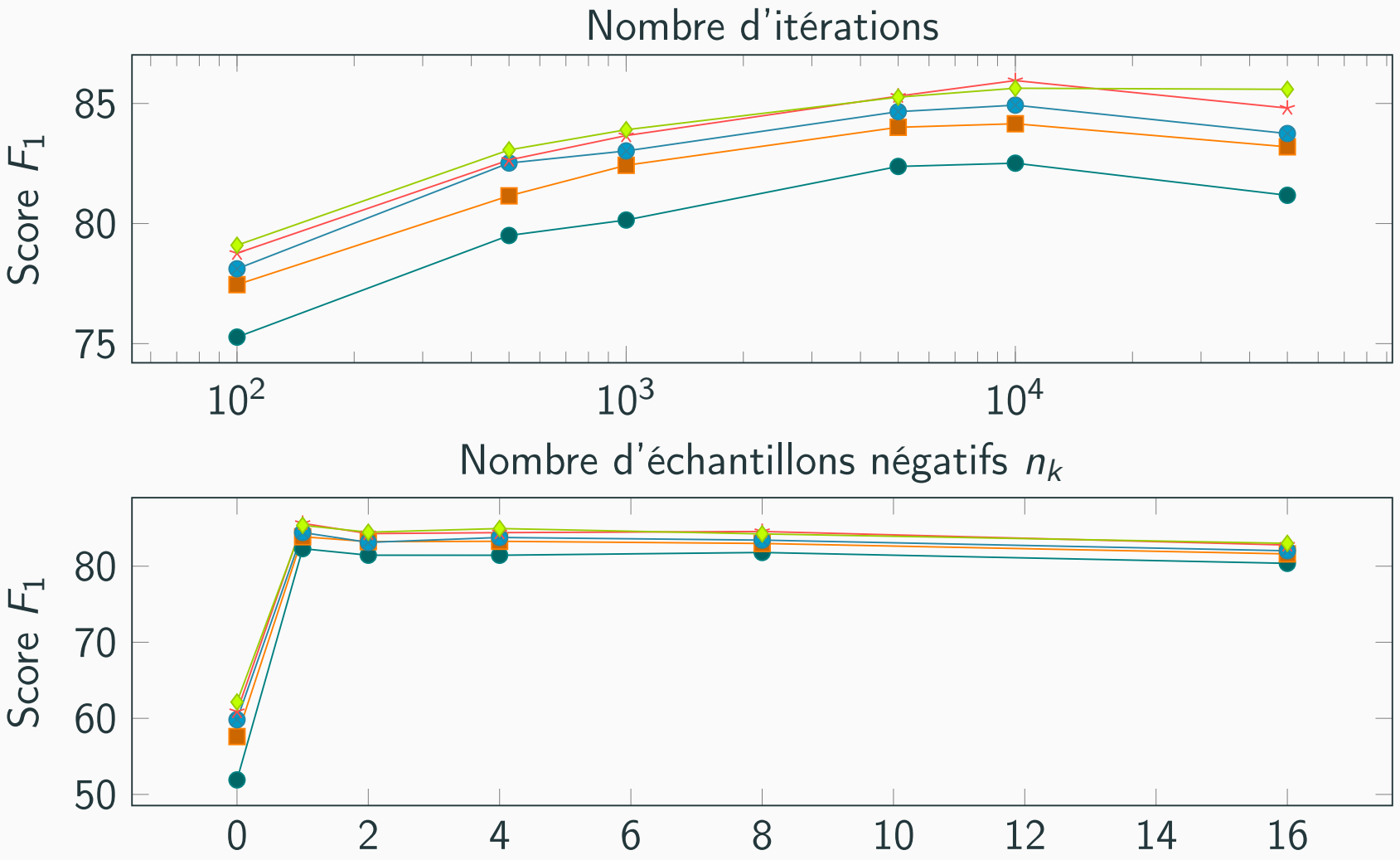
Étude des hyperparamètres de GVNR (1)



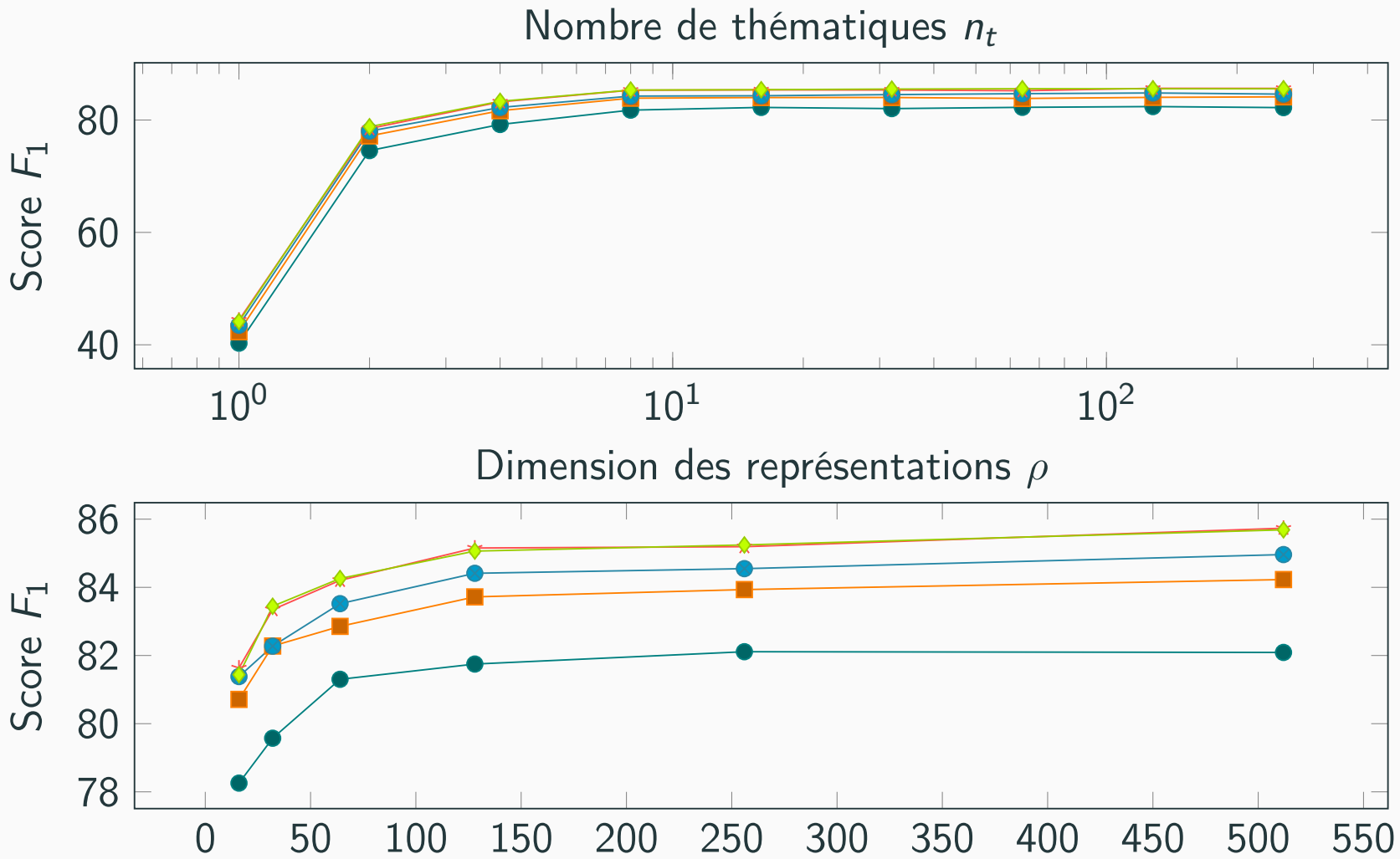
Étude des hyperparamètres de GVNR (2)



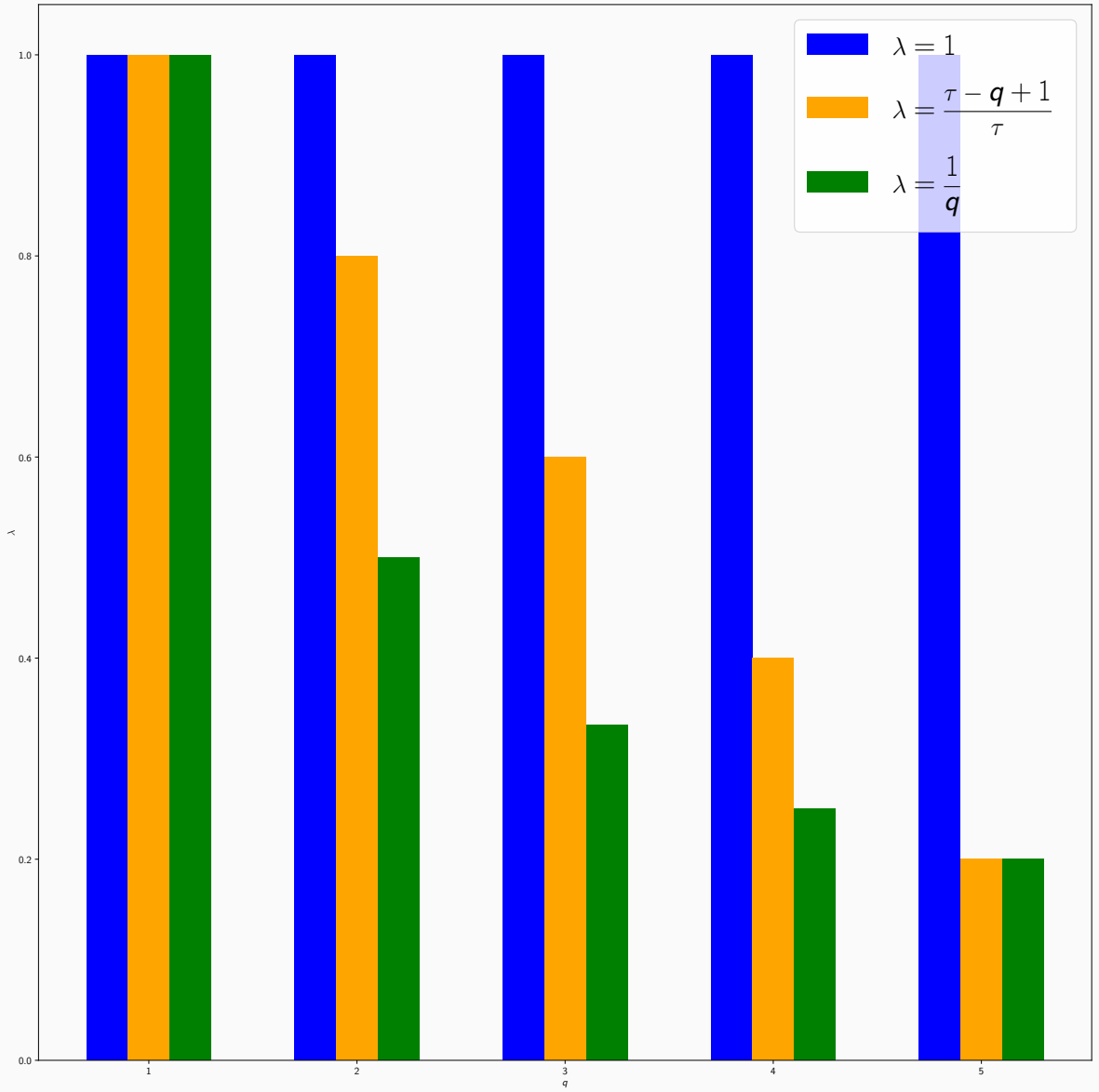
Étude des hyperparamètres de IDNE (1)



Étude des hyperparamètres de IDNE (2)



Pondérations expérimentées



Effet de la pondération et du seuillage sur GVNR

Cora		F1					AUC					Temps
λ	$s_{\min}$	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	
1	0	75.04	78.65	79.46	79.92	80.69	94.89	95.81	96.20	96.36	96.75	772s
$\frac{\tau-q+1}{\tau}$	0	74.84	78.11	79.55	79.77	81.14	94.39	95.28	96.04	96.09	96.57	702s
$\frac{1}{q}$	0	74.34	78.18	79.70	79.85	80.94	94.04	95.24	95.84	96.01	96.48	718s
1	1	75.48	**78.98**	**80.01**	80.14	**81.46**	**94.99**	**96.06**	**96.32**	96.47	**96.82**	701s
$\frac{\tau-q+1}{\tau}$	1	75.69	78.47	79.57	79.92	80.94	94.75	95.63	96.01	96.36	96.70	642s
$\frac{1}{q}$	1	**76.01**	78.75	79.74	**80.26**	80.81	94.90	95.71	96.13	**96.49**	96.75	**625s**

Wikipedia		F1					AUC					Temps
λ	$s_{\min}$	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%	
1	0	22.88	25.69	28.22	28.86	28.77	67.15	67.08	70.61	69.81	67.73	6031s
$\frac{\tau-q+1}{\tau}$	0	24.26	28.02	29.17	30.60	30.25	67.91	69.49	71.47	69.16	69.41	6048s
$\frac{1}{q}$	0	25.38	28.49	30.16	30.05	30.09	63.63	69.84	71.16	70.12	69.15	6165s
1	1	27.47	32.28	30.87	32.98	31.75	67.64	70.94	72.03	70.92	71.14	3768s
$\frac{\tau-q+1}{\tau}$	1	**31.95**	**34.79**	34.99	35.37	35.62	**72.75**	75.49	74.11	74.57	**74.01**	3409s
$\frac{1}{q}$	1	31.36	34.64	**35.71**	**36.51**	**35.64**	70.92	**76.06**	**75.24**	**74.91**	73.81	**2669s**

Biblio i

Andersen, R., Chung, F., and Lang, K. (2006).
Local graph partitioning using pagerank vectors.
In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486. IEEE.

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L., et al. (2012).
Expertise retrieval.
Foundations and Trends® in Information Retrieval, 6(2–3) :127–256.

Bamler, R. and Mandt, S. (2017).
Dynamic word embeddings.
In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org.

Bengio, Y., Courville, A., and Vincent, P. (2013).
Representation learning : A review and new perspectives.
IEEE transactions on pattern analysis and machine intelligence, 35(8) :1798–1828.

Brochier, R., Gourru, A., Guille, A., and Velcin, J. (2020a).
New datasets and a benchmark of document network embedding methods for scientific expert finding.
In *Bibliometric-enhanced Information Retrieval : 10th International BIR Workshop at ECIR*.

Biblio ii

Brochier, R., Guille, A., Rothan, B., and Velcin, J. (2018a).
Impact of the query set on the evaluation of expert finding systems.
In *3rd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) at SIGIR*.

Brochier, R., Guille, A., and Velcin, J. (2019a).
Global vectors for node representations.
In *Proceedings of The 2019 World Wide Web Conference*. International World Wide Web Conferences Steering Committee.

Brochier, R., Guille, A., and Velcin, J. (2019b).
Link prediction with mutual attention for text-attributed networks.
In *Companion Proceedings of The 2019 World Wide Web Conference*. International World Wide Web Conferences Steering Committee.

Brochier, R., Guille, A., and Velcin, J. (2020b).
Inductive document network embedding with topic-word attention.
In *Proceedings of the 42nd European Conference on Information Retrieval Research*. Springer.

Brochier, R., Guille, A., Velcin, J., Rothan, B., and Cioccio, D. (2018b).
Peerus review : a tool for scientific experts finding.
In *Extraction et Gestion des Connaissances (EGC)*.

Biblio iii

Fouss, F., Pirotte, A., Renders, J.-M., and Saerens, M. (2007).
Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation.
IEEE Transactions on knowledge and data engineering, 19(3) :355–369.

Fruchterman, T. M. and Reingold, E. M. (1991).
Graph drawing by force-directed placement.
Software : Practice and experience, 21(11) :1129–1164.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., and Holzinger, A. (2018).
Explainable ai : the new 42 ?
In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 295–303. Springer.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. (2019).
Strategies for pre-training graph neural networks.
In *International Conference on Learning Representations*.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. (2019).
Set transformer : A framework for attention-based permutation-invariant neural networks.
In *International Conference on Machine Learning*, pages 3744–3753.

Biblio iv


Levy, O. and Goldberg, Y. (2014).
Neural word embedding as implicit matrix factorization.
In *Advances in neural information processing systems*, pages 2177–2185.


McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018).
Umap : Uniform manifold approximation and projection.
Journal of Open Source Software, 3(29).


Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).
Efficient estimation of word representations in vector space.
arXiv preprint arXiv :1301.3781.


Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b).
Distributed representations of words and phrases and their compositionality.
In *Advances in neural information processing systems*, pages 3111–3119.


Page, L., Brin, S., Motwani, R., and Winograd, T. (1999).
The pagerank citation ranking : Bringing order to the web.
Technical report, Stanford InfoLab.


Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove : Global vectors for word representation.
In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Biblio v

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014).
Deepwalk : Online learning of social representations.
In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Rudolph, M. and Blei, D. (2018).
Dynamic embeddings for language evolution.
In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011.

Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002).
Methods and metrics for cold-start recommendations.
In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
In *Advances in neural information processing systems*, pages 5998–6008.

Yang, C., Liu, Z., Zhao, D., Sun, M., and Chang, E. (2015).
Network representation learning with rich text information.
In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Biblio vi

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., and Leskovec, J. (2018).
Graph convolutional neural networks for web-scale recommender systems.
In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983.