Fouille de la littérature médicale à l'aide de graphes

Elise Bigeard[1] Aman Sinha[1, 2] Marianne Clausel[1] Mathieu Constant[3]
(1) Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France
(2) Indian Institute of Technology, Dhanbad, Jharkhand 826004, India
(3) Université de Lorraine, CNRS, ATILF UMR 7118, 44 Avenue de la Libération, 54000 Nancy
elise.bigeard@univ-lorraine.fr, marianne.clausel@univ-lorraine.fr

MOTS-CLÉS : Fouille de texte, littérature médicale, graphes, embeddings.

La littérature scientifique est abondante, au point que l'explorer efficacement est maintenant une tâche majeure. Il existe de nombreuses plateformes répertoriant des publications : Arxiv, DBLP... et bien entendu Pubmed pour la littérature médicale. Ces plateformes sont une ressource clé, mais contiennent un très large contenu, et peuvent être difficiles à explorer.

Nous proposons d'analyser la littérature scientifique, et en particulier médicale, à l'aide de graphes de connaissance (Ji *et al.*, 2020). L'objectif est double : d'une part, proposer un outil graphique permettant d'explorer plus facilement et intuitivement la littérature. D'autre part, d'utiliser des plongements (embeddings) de graphes pour réaliser diverses tâches d'apprentissage automatique telles que la recommandation : étant donné par example une publication, suggérer des publications similaires, ou suggérer des co-auteurs potentiels.

Nous présentons une méthode full stack, allant d'une collection de publications au format PDF jusqu'à une représentation graphique accessible en ligne sur notre démonstrateur : https://gremie-demonstrator.atilf.fr

Nous testons notre méthode sur un ensemble de publications médicales au format PDF provenant d'un même établissement de recherche. Nous utilisons également des corpus non médicaux : ACM, ACL Anthlogy et DBLP.

En ce qui concerne la partie représentation des connaissances : nous créons un graphe de notre corpus, où sont présents les types de nœuds suivants : publication, auteur et mot-clé. Les mots-clé sont détectés automatiquement dans le texte de la publication, auxquels s'ajoutent les mots-clé indiqués explicitement par les auteurs. Nous nous basons sur la terminologie MESH et sur des synonymes issus de Wikidata pour détecter les mots-clé. Gephi est utilisé pour la représentation graphique.

En ce qui concerne les plongements, nous utilisons conjointement deux sources de données : le graphe, et le texte des articles.

Le graphe nous permet de comparer des nœuds en fonction de leurs liens. Ainsi, des auteurs peuvent être considérés comme similaires s'ils sont liés aux mêmes mot-clé, ou s'ils ont tous les deux un grand nombre de publications. Nous nous basons sur Deepwalk (Perozzi *et al.*, 2014), GraphSage (Hamilton *et al.*, 2017) et GCN Kipf2017.

Le texte de l'article nous permet de rapprocher des articles sur un sujet commun, ou utilisant des méthodes similaires. Nous utilisons plusieurs représentations de texte classiques : TF-IDF, plongements de mots (Mikolov *et al.*, 2013) et plongements de documents (Le & Mikolov, 2014).

Remerciement

Nous remercions Cancéropôle Est et l'INIST pour leur contribution, notamment leur corpus de publications.

Nous remercions le Proket Olki et l'agence AMIES pour leur financement.

Ce travail a bénéficié d'une aide de l'État, gérée par l'Agence Nationale de la Recherche, au titre du projet Investissements d'Avenir Lorraine Université d'Excellence, portant la référence ANR-15-IDEX-04-LUE.

Références

HAMILTON W. L., YING R. & LESKOVEC J. (2017). Inductive representation learning on large graphs. *CoRR*, **abs/1706.02216**.

JI S., PAN S., CAMBRIA E., MARTTINEN P. & YU P. S. (2020). A survey on knowledge graphs : Representation, acquisition and applications.

LE Q. & MIKOLOV T. (2014). Distributed representations of sentences and documents. In E. P. XING & T. JEBARA, Éds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 de *Proceedings of Machine Learning Research*, p. 1188–1196, Bejing, China : PMLR.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. Q. WEINBERGER, Éds., *Advances in Neural Information Processing Systems*, volume 26, p. 3111–3119 : Curran Associates, Inc.

PEROZZI B., AL-RFOU R. & SKIENA S. (2014). Deepwalk. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. DOI : 10.1145/2623330.2623732.