# Representation Learning for Recommender Systems with Application to the Scientific Literature

Robin Brochier (PhD Symposium)

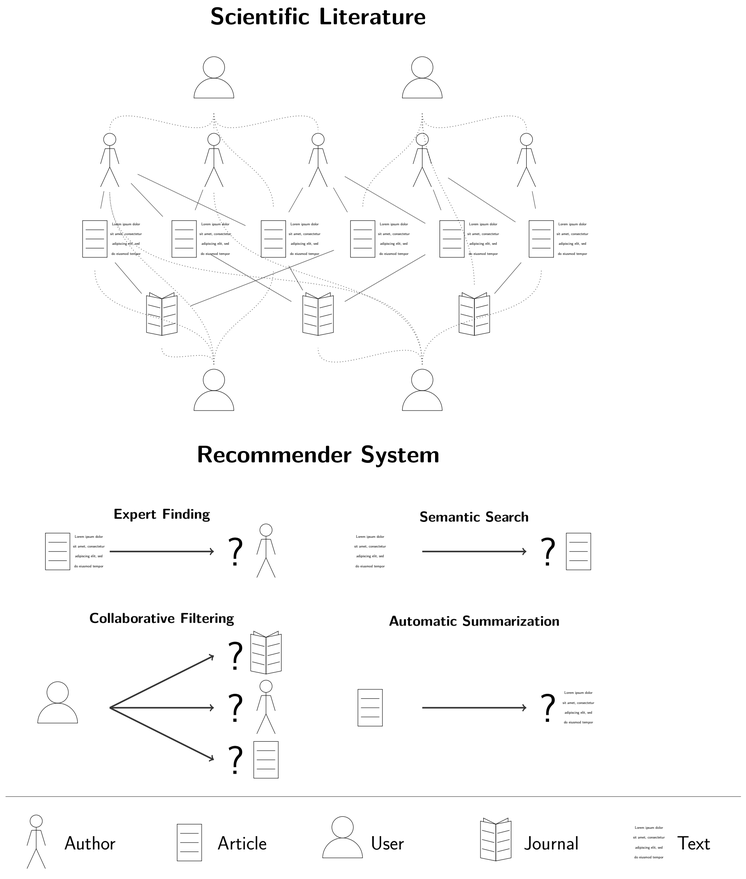
Université de Lyon, Lyon 2, ERIC EA3083
Peerus, DSRT F-69003

## Introduction

The ongoing research presented in this poster, done in partnership with an industrial player *Digital Scientific Research Technology* (*DSRT*) for its web application *Peerus* (`https://peer.us`), deals with the problem of learning representations in heterogeneous networks of documents applied to the recommendation of scientific literature in real time.

## Research Objectives

The scientific literature is a large information network linking various actors (laboratories, companies, institutions, etc.). The objective of this thesis is to build strong representations of both these actors and their scientific publications to solve various tasks of recommendation, such as expert finding [1].




## Embedding networks of documents

*GVNR-t* [2] embeds nodes, words and documents in the same vector space by (1) formulating a factorization problem on a thresholded co-occurrence matrix obtained by random walks and (2) expressing a document representation by the average of its bag of word embeddings:

$$\underset{U,V,b^U,b^V}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{n} s(x_{ij})\big(u_i \cdot v_{Dj} + b_i^U + b_j^V - \log(c + x_{ij})\big)^2$$

$$v_{Dj} = \frac{\delta_j\ W}{|\delta_j|_1}$$

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

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

## Application to keyword recommendation

Table 1: Keyword recommendation by selecting the closest word embeddings $w_k$ to both embeddings $u$ (node) and $\frac{\delta_j\ W}{|\delta_j|_1}$ (content) of an input document.

| | |
|---|---|
| Document | ***Wide coverage natural language processing using kernel methods and neural networks for structured data*** *Convolution kernels and recursive neural networks are both suitable approaches for supervised learning when the input is a discrete structure like a labeled tree or graph. We compare these techniques in two natural language problems...* |
| Closest words to $u$ | *successfully generally best outperforms predicting gene quantity genes tuning biology protein various competitive samples test established nips applied naive score* |
| Closest words to $\frac{\delta_j\ W}{\|\delta_j\|_1}$ | *complementary parsing discriminative generative parse deep shallow feed art pos ner predicting treebank trained natural models established combining classes latent* |

## Mutual attention mechanism

*MATAN* [3] adapts the smallest unit of the Transformer, the *Scaled Dot-Product Attention* (*SDPA*) to learn a semantic similarity from a network of documents. Given pre-trained bags of word embeddings $W^t$ and the multiset $C$ of pairs of connected nodes in the network, the representation of node $u$ given node $v$ is computed as follow:

- words embeddings are linearly projected to generate queries $Q_u = W^{t_u}P^Q$, keys $K_v = W^{t_v}P^K$ and values $V_v = W^{t_v}P^V$.
- the mutual representation of $u$ given $v$ is computed following the *SDPA* formula and averaging: $e_u^v = \sum_{i=0}^{L}\left[\text{softmax}(\frac{Q_u K_v^T}{\sqrt{D}})V_v\right]_i$.

The projection matrices are learned by gradient descent:

$$\underset{P^Q,P^K,P^V}{\operatorname{argmin}} \sum_{(u,v)\in C} \Bigg( \log\sigma(e_u^v \cdot e_v^u) + \sum_{i=1}^{k} \mathbb{E}_{z\sim q}\big[\log\sigma(-e_u^z \cdot e_z^u)\big] \Bigg)$$

## Explainability

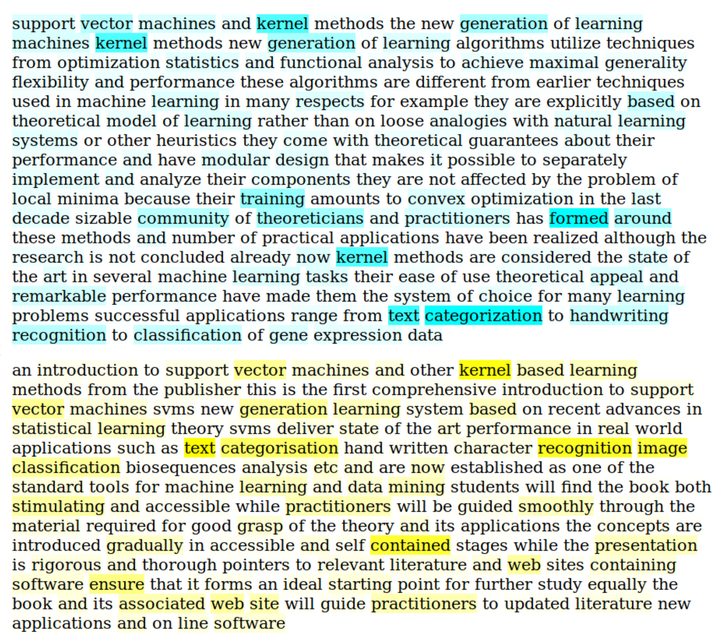



Figure 1: Average contribution of the mutual attention weights of the words of two connected documents in a citation network.

## Experiment Results

Table 2: Edges-hidden classification accuracy on a citation network.

| % 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** |

Table 3: Edges-hidden classification accuracy on a citation network.

| % of training data | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|
| *LSA* | 52.0 | 54.7 | 54.7 | 58.4 | 65.7 |
| *DeepWalk+LSA* | 58.3 | 60.7 | 61.1 | 60.0 | 61.2 |
| *TADW* | 60.6 | 60.1 | 60.1 | 66.2 | 69.3 |
| *GVNR-t* | **63.3** | **62.5** | **64.9** | **68.6** | **70.4** |

Table 4: Edges-hidden link prediction ROC AUC on a citation network.

| % of training data | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|
| *NetMF* | 59.0 | 67.2 | 77.5 | 83.2 | 87.2 |
| *TADW* | 68.0 | 82.0 | 87.1 | **93.2** | **94.5** |
| *MATAN* | **82.3** | **87.1** | **88.6** | 90.9 | 91.0 |

Table 5: Nodes-hidden link prediction ROC AUC on a citation network.

| % of training data | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|
| *TADW* | 64.2 | **75.8** | **80.3** | **81.9** | **82.3** |
| *MATAN* | **69.4** | 73.0 | 75.4 | 77.9 | 78.6 |

## Future work

- Improving the natural-language understanding component.
- Handling the heterogeneity of types of nodes and links in networks.
- Evaluation on real-world recommendation tasks.

## References

[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

- GitHub: https://github.com/brochier
- Web: http://www.robinbrochier.com
- Email: robin.brochier@univ-lyon2.fr