How Can Graph Neural Networks Help Document Retrieval: A Case Study on CORD19 with Concept Map Generation

Hejie Cui, Jiaying Lu, Yao Ge, and Carl Yang(✉)

Department of Computer Science, Emory University, Atlanta, Georgia
{hejie.cui,jiaying.lu,yao.ge,j.carlyang}@emory.edu

Abstract. Graph neural networks (GNNs), as a group of powerful tools for representation learning on irregular data, have manifested superiority in various downstream tasks. With unstructured texts represented as concept maps, GNNs can be exploited for tasks like document retrieval. Intrigued by how can GNNs help document retrieval, we conduct an empirical study on a large-scale multi-discipline dataset CORD-19. Results show that instead of the complex structure-oriented GNNs such as GINs and GATs, our proposed semantics-oriented graph functions achieve better and more stable performance based on the BM25 retrieved candidates. Our insights in this case study can serve as a guideline for future work to develop effective GNNs with appropriate semantics-oriented inductive biases for textual reasoning tasks like document retrieval and classification. All code for this case study is available at https://github.com/HennyJie/GNN-DocRetrieval.

Keywords: Document retrieval · Graph neural networks · Concept maps · Graph representation learning · Textual reasoning.

1 Introduction

Concept map, which models texts as a graph with words/phrases as vertices and relations between them as edges, has been studied to improve information retrieval tasks previously [10,14,46]. Recently, graph neural networks (GNNs) attract tremendous attention due to their superior power established both in theory and through experiments [6,12,16,20,32]. Empowered by the structured document representation of concept maps, it is intriguing to apply powerful GNNs for tasks like document classification [38] and retrieval [45]. Take Fig. 1 as an example. Towards the query about "violent crimes in society", a proper GNN might be able to highlight query-relevant concept of "crime" and its connection to "robbery" and "citizen", thus ranking the document as highly relevant. On the other hand, for another document about precaution, the GNN can capture concepts like "n95 mask" and "vaccine", together with their connections to "prevention", thus ranking it as not so relevant.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 75–83, 2022.
https://doi.org/10.1007/978-3-030-99739-7_9

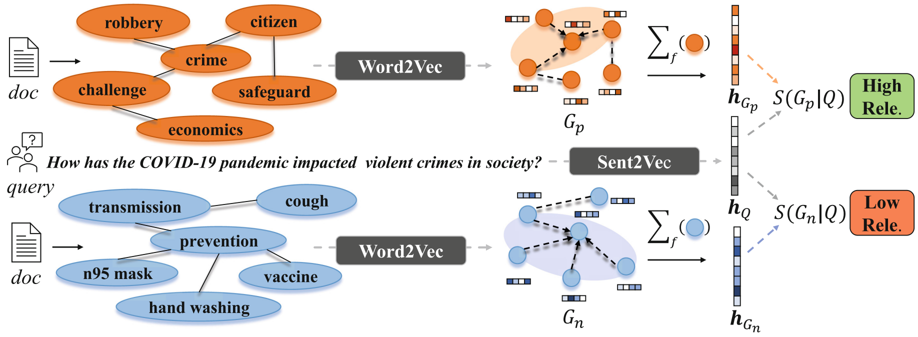


Fig. 1. An overview of GNN-based document retrieval.

Present work. In this work, we explore how GNNs can help document retrieval with generated concept maps. The core contributions are three-fold:

- We use constituency parsing to construct semantically rich concept maps from documents and design quality evaluation for them towards document retrieval.
- We investigate two types of graph models for document retrieval: the structure-oriented complex GNNs and our proposed semantics-oriented graph functions.
- By comparing the retrieval results from different graph models, we provide insights towards GNN model design for textual retrieval, with the hope to prompt more discussions on the emerging areas such as IR with GNNs.

2 GNNs for Document Retrieval

2.1 Overview

In this section, we describe the process of GNN-based document retrieval. As is shown in Fig. 1, concept maps $G = \{V, E\}$ are first constructed for documents. Each node $v_i \in V$ is a concept (usually a word or phrase) in the document, associated with a frequency f_i and an initial feature vector $\boldsymbol{a}_i$ from the pretrained model. The edges in E denote the interactions between concepts. GNNs are then applied to each individual concept map, where node representation $\boldsymbol{h}_i \in \mathbb{R}^d$ is updated through neighborhood transformation and aggregation. The graph-level embedding $\boldsymbol{h}_G \in \mathbb{R}^d$ is summarized over all nodes with a read-out function.

For the training of GNN models, the widely-used triplet loss in retrieval tasks [22,37,42] is adopted. Given a triplet (Q, G_p, G_n) composed by a relevant document G_p (denoted as positive) and an irrelevant document G_n (denoted as negative) to the query Q, the loss function is defined as:

$$L(Q, G_p, G_n) = \max\{S(G_n \mid Q) - S(G_p \mid Q) + margin, 0\}. \tag{1}$$

The relevance score $S\left(G \mid Q\right)$ is calculated as $\frac{\boldsymbol{h}_G \cdot \boldsymbol{h}_Q}{\|\boldsymbol{h}_G\| \|\boldsymbol{h}_Q\|}$, where $\boldsymbol{h}_G$ is the learned graph representation from GNN models and $\boldsymbol{h}_Q$ is the query representation from a pretrained model. In the training process, the embeddings of relevant documents are pulled towards the query representation, whereas those of the irrelevant ones are pushed away. For retrieval in the testing phrase, documents are ranked according to the learned relevance score $S(G \mid Q)$.

2.2 Concept Maps and Their Generation

Concept map generation, which aims to distill structured information hidden under unstructured text and represent it with a graph, has been studied extensively in literature [3,39,40,45]. Since entities and events often convey rich semantics, they are widely used to represent core information of documents [5,18,21]. However, according to our pilot trials, existing concept map construction methods based on name entity recognition (NER) or relation extraction (RE) often suffer from limited nodes and sparse edges. Moreover, these techniques rely on significant amounts of training data and predefined entities and relation types, which restricts the semantic richness of the generated concept maps [34].

To increase node/edge coverage, we propose to identify entities and events by POS-tagging and constituency parsing [23]. Compared to concept maps derived from NER or RE, our graphs can identify more sufficient phrases as nodes and connect them with denser edges, since pos-tagging and parsing are robust to domain shift [26,43]. The identified phrases are filtered via articles removing and lemmas replacing, and then merged by the same mentions. To capture the interactions (edges in graphs) among extracted nodes, we follow the common practice in phrase graph construction [17,27,31] that uses the sliding window technique to capture node co-occurrence. The window size is selected through grid search. Our proposed constituency parsing approach for concept map generation alleviates the limited vocabulary problem of existing NER-based methods, thus bolstering the semantic richness of the concept maps for retrieval.

2.3 GNN-based Concept Map Representation Learning

Structure-Oriented Complex GNNs. Various GNNs have been proposed for graph representation learning [12,16,32,36]. The discriminative power of complex GNNs mainly stems from the 1-WL test for graph isomorphism, which exhaustively capture possible graph structures so as to differentiate non-isomorphic graphs [36]. To investigate the effectiveness of structured-oriented GNNs towards document retrieval, we adopt two state-of-the-art ones, Graph isomorphism network (GIN) [36] and Graph attention network (GAT) [32], as representatives.

Semantics-Oriented Permutation-Invariant Graph Functions. The advantage of complex GNNs in modelling interactions may become insignificant for semantically important task. In contrast, we propose the following series of graph functions oriented from semantics perspectives.

Table 1. The similarity of different concept map pairs.

Pair Type	# Pairs	NCR (%)	NCR+ (%)	ECR (%)	ECR+ (%)
Pos-Pos	762,084	4.96	19.19	0.60	0.78
Pos-Neg	1,518,617	4.12	11.75	0.39	0.52
(*t-score*)	–	(*187.041*)	(*487.078*)	(*83.569*)	(*105.034*)
Pos-BM	140,640	3.80	14.98	0.37	0.43
(*t-score*)	–	(*126.977*)	(*108.808*)	(*35.870*)	(*56.981*)

- **N-Pool**: independently process each single node v_i in the concept map by multi-layer perceptions and then apply a read-out function to aggregate all node embeddings $\boldsymbol{a}_i$ into the graph embedding $\boldsymbol{h}_G$, i.e.,

$$\boldsymbol{h}_G = \text{READOUT}\Big(\{\text{MLP}(\boldsymbol{a}_i) \mid v_i \in V\}\Big). \tag{2}$$

- **E-Pool**: for each edge $e_{ij} = (v_i, v_j)$ in the concept map, the edge embedding is obtained by concatenating the projected node embedding $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ on its two ends to encode first-order interactions, i.e.,

$$\boldsymbol{h}_G = \text{READOUT}\Big(\ \{\mathit{cat}\,(\text{MLP}(\boldsymbol{a}_i), \text{MLP}(\boldsymbol{a}_j)) \mid e_{ij} \in E\}\ \Big). \tag{3}$$

- **RW-Pool**: for each sampled random walk $p_i = (v_1, v_2, \ldots, v_m)$ that encode higher-order interactions among concepts ($m = 2, 3, 4$ in our experiments), the embedding is computed by the sum of all node embeddings on it, i.e.,

$$\boldsymbol{h}_G = \text{READOUT}\Big(\{\mathit{sum}\,(\text{MLP}(\boldsymbol{a}_1), \text{MLP}(\boldsymbol{a}_2), \ldots, \text{MLP}(\boldsymbol{a}_m)) \mid p_i \in P\}\Big). \tag{4}$$

All of the three proposed graph functions are easier to train and generalize. They preserve the *message passing* mechanism of complex GNNs [11], which is essentially *permutation invariant* [15,24,25], meaning that the results of GNNs are not influenced by the orders of nodes or edges in the graph; while focusing on the basic semantic units and different level of interactions between them.

3 Experiments

3.1 Experimental Setup

Dataset. We adopt a large scale multi-discipline dataset from the TREC-COVID[1] challenge [29] based on the CORD-19[2] collection [33]. The raw data includes a corpus of 192,509 documents from broad research areas, 50 queries about the pandemic that interest people, and 46,167 query-document relevance labels.

[1] https://ir.nist.gov/covidSubmit/.
[2] https://github.com/allenai/cord19.

Experimental Settings and Metrics. We follow the common two-step practice for the large-scale document retrieval task [7,19,28]. The initial retrieval is performed on the whole corpus with full texts through BM25 [30], a traditional yet widely-used baseline. In the second stage, we further conduct re-ranking on the top 100 candidates using different graph models. The node features and query embeddings are initialized with pretrained models from [4,44]. NDCG@20 is adopted as the main evaluation metric for retrieval, which is used for the competition leader board. Besides NDCG@K, we also provide Precision@K and Recall@K (K=10, 20 for all metrics).

3.2 Evaluation of Concept Maps

We empirically evaluate the quality of concept maps generated from Sect. 2.2. The purpose is to validate that information in concept maps can indicate query-document relevance, and provide additional discriminative signals based on the initial candidates. Three types of pairs are constructed: a Pos-Pos pair consists of two documents both relevant to a query; a Pos-Neg pair consists of a relevant and an irrelevant one; and a Pos-BM pair consists of a relevant one and a top-20 one from BM25. Given a graph pair G_i and G_j, their similarity is calculated via four measures: the node coincidence rate (NCR) defined as $\frac{|V_i \cap V_j|}{|V_i \cup V_j|}$; NCR+ defined as NCR weighted by the tf-idf score [1] of each node; the edge coincidence rate (ECR) where an edge is coincident when its two ends are contained in both graphs; and ECR+ defined as ECR weighted by the tf-idf scores of both ends.

It is shown in Table 1 that Pos-Neg pairs are less similar than Pos-Pos under all measures, indicating that concept maps can effectively reflect document semantics. Moreover, Pos-BM pairs are not close to Pos-Pos and even further away than Pos-Neg. This is because the labeled "irrelevant" documents are actually hard negative ones difficult to distinguish. Such results indicate the potential for improving sketchy candidates with concept maps. Besides, student's t-Test [13] is performed, where standard critical values of (Pos-Pos, Pos-Neg) and (Pos-Pos, Pos-BM) under 95% confidence are 1.6440 and 1.6450, respectively. The calculated *t-scores* shown in Table 1 strongly support the significance of differences.

3.3 Retrieval Performance Results

In this study, we focus on the performance improvement of GNN models based on sketchy candidates. Therefore, two widely-used and simple models, the forementioned BM25 and Anserini[3], are adopted as baselines, instead of the heavier language models such as BERT-based [8,9,41] and learning to rank (LTR)-based [2,35] ones. The retrieval performance are shown in Table 2. All the values are reported as the averaged results of five runs under the best settings.

[3] https://git.uwaterloo.ca/jimmylin/covidex-trec-covid-runs/-/tree/master/round5, whichisrecognizedbythecompetitionorganizersasabaselineresult.

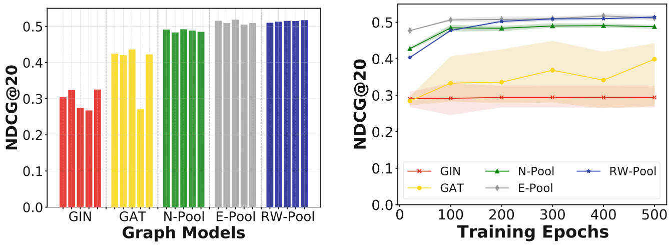
Table 2. The retrieval performance results of different models.

.5Type	.5Methods	**Precision (%)**		**Recall (%)**		**NDCG (%)**	
		k = 10	k = 20	k = 10	k = 20	k = 10	k = 20
Traditional	BM25	55.20	49.00	1.36	2.39	51.37	45.91
	Anserini	54.00	49.60	1.22	2.25	47.09	43.82
Structure-Oriented	GIN	35.24	34.36	0.77	1.50	30.59	29.91
	GAT	46.48	43.26	1.08	2.00	42.24	39.49
Semantics-Oriented	N-Pool	58.24	52.20	1.38	2.41	53.38	48.80
	E-Pool	59.60	53.88	1.40	2.49	56.11	51.16
	RW-Pool	**59.84**	**53.92**	**1.42**	**2.53**	**56.19**	**51.41**

For the structure-oriented GIN and GAT, different read-out functions including mean, sum, max and a novel proposed tf-idf (i.e., weight the nodes using the tf-idf scores) are experimented, and tf-idf achieves the best performance. It is shown that GIN constantly fails to distinguish relevant documents while GAT is relatively better. However, they both fail to improve the baselines. This performance deviation may arise from the major inductive bias on complex structures, which makes limited contribution to document retrieval and is easily misled by noises. In contrast, our three proposed semantics-oriented graph functions yield significant and consistent improvements over both baselines and structure-oriented GNNs. Notably, E-Pool and RW-Pool improve the document retrieval from the initial candidates of BM25 by 11.4% and 12.0% on NDCG@20, respectively. Such results demonstrate the potential of designing semantics-oriented GNNs for textual reasoning tasks such as classification, retrieval, etc.

3.4 Stability and Efficiency

We further examine the stability and efficiency of different models across runs. As is shown in Fig. 2(a), GIN and GAT are less consistent, indicating the diffi-



(a) Stability comparison **(b)** Efficiency comparison

Fig. 2. Stability and efficiency comparison of different graph models.

culty in training over-complex models. The training efficiency in Fig. 2(b) shows that GIN can hardly improve during training, while GAT fluctuates a lot and suffers from overfitting. In contrast, our proposed semantics-oriented functions perform more stable in Fig. 2(a), and improve efficiently during training in Fig. 2(b), demonstrating their abilities to model the concepts and interactions important for the retrieval task. Among the three graph functions, E-Pool and RW-Pool are consistently better than N-Pool, revealing the utility of simple graph structures. Moreover, RW-Pool converges slower but achieves better and more stable results in the end, indicating the potential advantage of higher-order interactions.

4 Conclusion

In this paper, we investigate how can GNNs help document retrieval through a case study. Concept maps with rich semantics are generated from unstructured texts with constituency parsing. Two types of GNNs, structure-oriented complex models and our proposed semantics-oriented graph functions are experimented and the latter achieves consistently better and stable results, demonstrating the importance of semantic units as well as their simple interactions in GNN design for textual reasoning tasks like retrieval. In the future, more textual datasets such as news, journalism and downstream tasks can be included for validation. Other types of semantics-oriented graph functions can also be designed based on our permutation-invariant schema, such as graphlet based-pooling.

References

1. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern Information Retrieval, vol. 463 (1999)
2. Burges, C.J.C., et al.: Learning to rank using gradient descent. In: ICML (2005)
3. Chen, N., Kinshuk, Wei, C., Chen, H.: Mining e-learning domain concept map from academic articles. Comput. Educ. **50**(5), 1009–1021 (2008)
4. Chen, Q., Peng, Y., Lu, Z.: Biosentvec: creating sentence embeddings for biomedical texts. In: ICHI, pp. 1–5 (2019)
5. Christensen, J., Mausam, Soderland, S., Etzioni, O.: Towards coherent multi-document summarization. In: NAACL, pp. 1163–1173 (2013)
6. Cui, H., Lu, Z., Li, P., Yang, C.: On positional and structural node features for graph neural networks on non-attributed graphs. CoRR abs/2107.01495 (2021)
7. Dang, V., Bendersky, M., Croft, W.B.: Two-stage learning to rank for information retrieval. In: Serdyukov, P., et al. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 423–434. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36973-5_36
8. Deshmukh, A.A., Sethi, U.: IR-BERT: leveraging BERT for semantic search in background linking for news articles. CoRR abs/2007.12603 (2020)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
10. Farhi, S.H., Boughaci, D.: Graph based model for information retrieval using a stochastic local search. Pattern Recognit. Lett. **105**, 234–239 (2018)
11. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: ICML, pp. 1263–1272 (2017)

12. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: NeurIPS (2017)
13. Hogg, R.V., McKean, J., et al.: Introduction to Mathematical Statistics (2005)
14. Kamphuis, C.: Graph databases for information retrieval. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 608–612. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_79
15. Keriven, N., Peyré, G.: Universal invariant and equivariant graph neural networks. In: NeurIPS (2019)
16. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
17. Krallinger, M., Padron, M., Valencia, A.: A sentence sliding window approach to extract protein annotations from biomedical articles. BMC Bioinform. **6**, 1–12 (2005)
18. Li, M., et al.: Connecting the dots: event graph schema induction with path language modeling. In: EMNLP, pp. 684–695 (2020)
19. Liu, T.Y.: Learning to Rank for Information Retrieval, pp. 181–191 (2011). https://doi.org/10.1007/978-3-642-14267-3_14
20. Liu, Z., et al.: Geniepath: graph neural networks with adaptive receptive paths. In: AAAI, vol. 33, no. 1, pp. 4424–4431 (2019)
21. Lu, J., Choi, J.D.: Evaluation of unsupervised entity and event salience estimation. In: FLAIRS (2021)
22. Manmatha, R., Wu, C., Smola, A.J., Krähenbühl, P.: Sampling matters in deep embedding learning. In: ICCV, pp. 2840–2848 (2017)
23. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: ACL, pp. 55–60 (2014)
24. Maron, H., Ben-Hamu, H., Shamir, N., Lipman, Y.: Invariant and equivariant graph networks. In: ICLR (2019)
25. Maron, H., Fetaya, E., Segol, N., Lipman, Y.: On the universality of invariant networks. In: ICML, pp. 4363–4371 (2019)
26. McClosky, D., Charniak, E., Johnson, M.: Automatic domain adaptation for parsing. In: NAACL Linguistics, pp. 28–36 (2010)
27. Mihalcea, R., Tarau, P.: Textrank: bringing order into text. In: EMNLP, pp. 404–411 (2004)
28. Nogueira, R., Cho, K.: Passage re-ranking with bert. arXiv preprint arXiv:1901.04085 (2019)
29. Roberts, K., et al.: Searching for scientific evidence in a pandemic: an overview of TREC-COVID. J. Biomed. Inform. **121**, 103865 (2021)
30. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at trec-3. In: TREC (1994)
31. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Min. Appl. Theory **1**, 1–20 (2010)
32. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: ICLR (2018)
33. Wang, L.L., Lo, K., Chandrasekhar, Y., et al.: CORD-19: the COVID-19 open research dataset. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL (2020)
34. Wang, X., Yang, C., Guan, R.: A comparative study for biomedical named entity recognition. Int. J. Mach. Learn. Cybern. **9**(3), 373–382 (2015). https://doi.org/10.1007/s13042-015-0426-6
35. Wu, Q., Burges, C.J.C., Svore, K.M., Gao, J.: Adapting boosting for information retrieval measures. Inf. Retr. **13**, 254–270 (2010)

36. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: ICLR (2019)
37. Yang, C., et al.: Multisage: empowering GCN with contextualized multi-embeddings on web-scale multipartite networks. In: KDD, pp. 2434–2443 (2020)
38. Yang, C., Zhang, J., Wang, H., Li, B., Han, J.: Neural concept map generation for effective document classification with interpretable structured summarization. In: SIGIR, pp. 1629–1632 (2020)
39. Yang, C., et al.: Relation learning on social networks with multi-modal graph edge variational autoencoders. In: WSDM, pp. 699–707 (2020)
40. Yang, C., Zhuang, P., Shi, W., Luu, A., Li, P.: Conditional structure generation through graph variational generative adversarial nets. In: NeurIPS (2019)
41. Yilmaz, Z.A., Wang, S., Yang, W., Zhang, H., Lin, J.: Applying BERT to document retrieval with birch. In: EMNLP, pp. 19–24 (2019)
42. Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L., Leskovec, J.: Graph convolutional neural networks for web-scale recommender systems. In: KDD, pp. 974–983 (2018)
43. Yu, J., El-karef, M., Bohnet, B.: Domain adaptation for dependency parsing via self-training. In: Proceedings of the 14th International Conference on Parsing Technologies, pp. 1–10 (2015)
44. Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z.: Biowordvec, improving biomedical word embeddings with subword information and mesh. Sci. Data **6**, 1–9 (2019)
45. Zhang, Y., Zhang, J., Cui, Z., Wu, S., Wang, L.: A graph-based relevance matching model for ad-hoc retrieval. In: AAAI (2021)
46. Zhang, Z., Wang, L., Xie, X., Pan, H.: A graph based document retrieval method. In: CSCWD, pp. 426–432 (2018)