Comparative Study of Document Embedding for Link Prediction

Abstract

Document embedding is general method to preprocess document in language model and directly influence further performance. Previous studies on this task are based on either text of documents or citation network structure. In this paper, we proposed to leverage citation context to generate document embedding. We then propose to combine text-based and network structure characteristics by using initialization and re-training method. Our proposed method outperforms both popular text-based and network-based embedding on AAN dataset.

1 Introduction

Document Embedding is a general method to preprocessing documents in Machine Learning, Text Mining and Natural Language Processing. Mapping variable-length, sparse word sequences into fixed-length vectors is necessary for many Language model and Neural Network. (e.g. CNN). Document embedding generation aims to contain more information and reflect potential relation between documents. Therefore, Link Prediction can be effective to measure Document Embedding methods.

Document Embedding is an important task because quality of embedding can directly affect further results. Most previous works embedded documents by either text or citation network. Text embedding is the most popular method to map word sequences into vectors and one of the most common embedding methods is Bag-of-Word. Despite its popularity, BOW casts off the order of words and thus cannot reflect the genuine semantics of word sequences. Moreover, sparsity of its embedding space brings huge noise into analysis. In recent years, a variety of neural network methods (Mikolov, 2013; Le and Mikolov, 2014) have been proposed to map words, paragraphs or documents to dense vectors which are considered to embody their latent semantics.

Leveraging Citation Network (Lim and Buntine, 2016; Yang et al., 2016) is also an effective way to reflect relationship between documents. In many datasets, like web pages, papers and articles, links/citations between documents can directly express relationship between documents. Many general network embedding methods, including BPRMF (Rendle et al., 2009), LINE (Tang, 2015), are widely used for document embedding. However, few Document Embedding methods try to combine text embedding and network structure between documents while both of them can reflect important information of documents. What's more, as directional text, citation context is seldom used for document embedding generation. In this study, we proposed to leverage citation context to generate document embedding by using different models. Also, we proposed to combine text features and network structure to generate document embedding. After that, we compare different embedding methods to assess contribution of different network structure and text-based embedding. Comparison results for link prediction on ACL Anthology Network (AAN) datasets show that our proposed method can significantly improve the predication performance.

The contributions of this paper are summarized as follows:

1) We proposed new methods to leverage citation context.
2) We proposed to generate document embedding by leveraging both text embedding and network structure.
3) We compared various network generation and text-based embedding methods in our link prediction task of AAN dataset.

2 Models

Two sets of embedding methods will be introduced in this section: (1) Text-based Embedding (2) Network Generation. Also, our proposed retraining method will be introduced.

2.1 Text-based Embedding

Text-based Embedding is most common method to map word sequences into fixed-length vectors. Many popular works can be used for document embedding, including Bag-of-Word (BOW), Word2Vec and Paragraph2vec (Mikolov, 2013; Le and Mikolov, 2014).

However, how to select embedding section is also considerable. Some works embed full text while some works focus on specific section, like <*abstract*>. Generally, when document is cited by another document, context of this citation highly summarize target document in an objective way. Therefore, we proposed to embed citation context of document. Formally, for documents set $\boldsymbol{D} = \{d_i\}$, $text_i$ means embedded full text of d_i.

Also, we have citation network:

$$\boldsymbol{E} = \left\{e_j =< o_j, t_j > \mid d_{o_j}, d_{t_j} \in \boldsymbol{D}\right\}$$

and $context_j$ means embedded context of citation e_j. We have two sets of text-based embedding:

Document2vec:

$$d_i = text_i$$

CitationContext2Document:

$$d_i = average(context_j, t_j = d_i)$$

In our study, the open source toolkit word2vec[1] is used to generate paragraph vectors. This toolkit was firstly developed to implement the word vector model and then modified to obtain paragraph vector. The vector dimension is simply set to 100 and the window size of words is set to 10. The negative sampling method is used and the number of negative samples is set to 5 and the sampling threshold is set to 1e-3. The iteration time is set to 20. All the above parameter values were suggested by Quoc Le[2].

2.2 Network Generation

In this section, several methods of network generation will be introduced.

2.2.1 Citation Network

For citation, there are some popular modified networks:

Citation Network:

$$\boldsymbol{E_{CN}} = \boldsymbol{E}$$

Co-Cited Network:

$$\boldsymbol{E_{CoCitedN}} = \begin{Bmatrix} e =< o, t > | d_o, d_t \in \boldsymbol{D}, \\ \exists d_x \in \boldsymbol{D}, \\ < x, o >, < x, t > \in \boldsymbol{E} \end{Bmatrix}$$

Co-Citing Network:

$$\boldsymbol{E_{CoCitingN}} = \begin{Bmatrix} e =< o, t > | d_o, d_t \in \boldsymbol{D}, \\ \exists d_x \in \boldsymbol{D}, \\ \exists < o, x >, < t, x > \in \boldsymbol{E} \end{Bmatrix}$$

Citation Network + Co-Cited Network:

$$\boldsymbol{E_{CN+CCitedN}} = \boldsymbol{E} \cup \boldsymbol{E_{CCitedN}}$$

Citation Network + Co-Citing Network:

$$\boldsymbol{E_{CN+CCitingN}} = \boldsymbol{E} \cup \boldsymbol{E_{CCitingN}}$$

2.2.2 Weighted Citation Network

Empirically, we proposed to weight edges of network:

Weight Co-Cited Network by Context Similarity:

For $e =< o, t > \in \boldsymbol{E_{CoCitedN}}$,

$$weight(e) = \sum \begin{Bmatrix} Similarity(context_i, context_j) | \\ \exists x, e_i =< x, o >, e_j =< x, t > \in \boldsymbol{E} \end{Bmatrix}$$

Weight Citation Network by OwnerText-Context Similarity:

For $e_j =< o_j, t_j > \in \boldsymbol{E_{CN}}$,

$$weight(e_j) = \sum Similarity\left(text_{o_j}, context_j\right)$$

Weight Citation Network by TargetText-Context Similarity:

For $e_j =< o_j, t_j > \in \boldsymbol{E_{CN}}$,

$$weight(e_j) = \sum Similarity\left(text_{t_j}, context_j\right)$$

2.2.3 Good-Customer Model

In this section, we proposed to analyze relationship between documents as goods and customers. That is, we regarded citation as purchase behavior, context of citation as comment. For example, if document o cites document t, we regarded o as customer and t as good. Context of this citation can be regarded as a comment from customer o to good t. There are two assumptions: (1) for customers, similarity of their purchases and

[1] https://groups.google.com/d/msg/word2vec/toolkit-Q49FIrNOQ Ro/J6KG8mUj45sJ

[2] https://groups.google.com/d/msg/word2vec-toolkit/Q49FIrNOQ Ro/CJLWzmr0LaUJ

similarity of their comments towards same good can be regarded as their similarity (2) for two goods, similarity of their purchasers and similarity of their comments from same customer can be regarded as their similarity. Thus, we proposed to weight co-cited network and co-citing network:

Good-Based Quantity-Focused Network:

For $e =< o,t >\in \boldsymbol{E_{CoCitedN}}$,

$$weight(e) = \frac{count(x)}{\log\left(indegree(o) + indegree(t)\right)},$$

$$< x,o >, < x,t >\in \boldsymbol{E}$$

Good-Based Quality-Focused Network:

For $e =< o,t >\in \boldsymbol{E_{CoCitedN}}$,

$$weight(e) = \frac{\sum Similarity\left(context_i, context_j\right)}{count(x)},$$

$$e_i =< x,o >, e_j =< x,t >\in \boldsymbol{E}$$

Good-Based Balance-Focused Network:

For $e =< o,t >\in \boldsymbol{E_{CoCitedN}}$,

$$weight(e) = \frac{\sum Similarity\left(context_i, context_j\right)}{\log\left(indegree(o) + indegree(t)\right)},$$

$$e_i =< x,o >, e_j =< x,t >\in \boldsymbol{E}$$

Customer-Based Quantity-Focused Network:

For $e =< o,t >\in \boldsymbol{E_{CoCitingN}}$,

$$weight(e) = \frac{count(x)}{\log\left(outdegree(o) + outdegree(t)\right)},$$

$$< o,x >, < t,x >\in \boldsymbol{E}$$

Customer -Based Quality-Focused Network:

For $e =< o,t >\in \boldsymbol{E_{CoCitingN}}$,

$$weight(e) = \frac{\sum Similarity\left(context_i, context_j\right)}{count(x)},$$

$$e_i =< o,x >, e_j =< t,x >\in \boldsymbol{E}$$

Customer -Based Balance-Focused Network:

For $e =< o,t >\in \boldsymbol{E_{CoCitingN}}$,

$$weight(e) = \frac{\sum Similarity\left(context_i, context_j\right)}{\log\left(outdegree(o) + outdegree(t)\right)},$$

$$e_i =< o,x >, e_j =< t,x >\in \boldsymbol{E}$$

Here, $indegree(i)$ and $outdegree(i)$ represents in-degree and out-degree of i in Citation Network $\boldsymbol{E}$.

2.3 Network Embedding

After generating network, we embed these networks to generate our document embedding. Recently, many frameworks are proposed to embed network, including BPR Matrix Factorization (Rendle et al., 2009), and Large-scale information network embedding (LINE) (Tang, 2015). In this study, we use LINE[3] as network embedding toolkit.

2.3.1 Large-scale information network embedding

In LINE, for each edge $< o,t >\in \boldsymbol{E}$, the joint probability between d_o and d_t is defined as :

$$p(d_o, d_t) = \frac{1}{1 + \exp(-u_o^T . u_t)},$$

where u_i is the embedding of vertex i.

And empirical probability between d_o and d_t is defined as:

$$\hat{p}(d_o, d_t) = \frac{w_{ot}}{W},$$

where $W = \sum w_{ij}, < i,j >\in \boldsymbol{E}$.

The aim of embedding process is to minimize the following objective function:

$$O = d\left(p(.,.), p(.,.)\right),$$

where $d(.,.)$ is the distance between two distribution.

In this study, we used open source toolkit *LINE* to generate network embedding and all parameters are set default.

2.3.2 Initialization

Intuitively, we tended to combine text-based embedding and network-based embedding in order to capture more information. In LINE work, embedding of document is initialized randomly. In this paper, we proposed to use normalized text-based embedding as initialization of network embedding and then generate network embedding with a relative low rho. In this way, we used network structure to re-train our text-based embedding. By using this method, we obtain the re-training embedding by considering both text-based and network-based characteristics.

3 Evaluation

3.1 Data Set and Evaluation Metrics

In this study, we used ACL Anthology Network (AAN)[4] 2014 as our dataset. AAN contains 22,486 papers and we extracted 99,376 edges with citation context. There are 12,4812 edges in AAN's citation network. We randomly select 1061 edges as testing set and remove these edges from citation network. We used recall@N as metrics. That is, for each document, we calculate top N documents as our predicted links and then

[3] https://github.com/tangjianpku/LINE

[4] http://clair.eecs.umich.edu/aan/index.php

evaluate embedding by recall of testing edges in these predict links.

3.2 Evaluation Result

Table 1 shows comparison results of all document embedding introduced in section 2. Results are divided into 6 groups by dotted lines.

Group 1 and 2 show our text-based embedding. We can see that documents embedded by citation context outperform than full-text embedding in most models. It proves our assumption that context of citation can highly summarize cited article. Also, in these embedding methods, Para2vec and TFIDF achieve best recall.

Group 3 shows comparison results between different modified citation networks. Undirected citation network is denoted as "CitationNetworkBi". We experiment citation network embedding in 1st order proximity (denoted as suffix "-1st" or no suffix) and 2nd order proximity (denoted as suffix "-2nd") of LINE respectively. We can see Citation Network in 1st order proximity performs best. Moreover, we can also see that Co-Cited network performs better than Co-Citing network.

Group 4 shows comparison results of different weighted modified citation network. We can see that using similarity (TFIDF) of citation context as weight of Co-Cited network outperforms original Co-Cited network in recall@10 and recall @20. Also, we can hardly use similarity between citation context and full text of document to weight citation.

In Group 5, we analyze relationship between documents in Good-Customer model. We can see Good model perform better than Customer model. It shows that citation context tends to describe cited documents.

According to previous results, we proposed to combine text-based embedding and citation network through our initialization method (denoted as "CNITBE"). We use "Context2Document-Para2vec" as initialization and citation network to re-train embedding. In order to investigate the influence of rho parameter, we try different values of rho, which can stand for level of retraining. Group 6 shows comparison results of our proposed method in different rho. We can see from results that both test-based embedding and network contribute to the improvement of the results, and our proposed method can achieve very good performance. We can see that all results outperform previous results, which demonstrates the robustness of our proposed method.

Embedding Method	R@10	R@20	R@50	R@100
document2vec-TF	0.078	0.111	0.169	0.232
document2vec-TFIDF	**0.160**	**0.233**	**0.335**	**0.423**
document2vec-Word2vec	0.097	0.144	0.221	0.299
document2vec-Para2vec	0.129	0.196	0.303	0.405
Context2Document-TF	0.119	0.179	0.263	0.333
Context2Document-TFIDF	0.119	0.180	0.304	0.401
Context2Document-Word2vec	0.107	0.156	0.234	0.299
Context2Document-Para2vec	**0.159**	**0.223**	**0.322**	**0.423**
CitationNetwork-1st	**0.100**	**0.169**	**0.303**	**0.419**
CitationNetwork-2nd	0.074	0.118	0.204	0.283
CitationNetworkBi-2nd	0.074	0.110	0.205	0.285
CoCitedNetwork	0.088	0.149	0.251	0.332
CoCitingNetwork	0.035	0.059	0.114	0.181
CN+CCitedN	0.094	0.164	0.291	0.399
CN+CCitingN	0.034	0.069	0.134	0.211
ContextSim-Para2vec	0.075	0.132	0.217	0.299
ContextSim-TFIDF	**0.098**	**0.150**	**0.232**	**0.310**
ContextSim-Word2vec	0.071	0.125	0.207	0.279
OwnerText-ContextSim	0.018	0.026	0.040	0.064
TargetText-ContextSim	0.011	0.020	0.034	0.043
Goods-Based-Balance	0.079	0.138	0.230	0.304
Goods-Based-Quality	0.065	0.109	0.197	0.279
Goods-Based-Quantity	**0.087**	**0.145**	**0.241**	**0.341**
Customer-Based-Balance	0.044	0.074	0.130	0.194
Customer-Based-Quality	0.038	0.062	0.109	0.167
Customer-Based-Quantity	0.039	0.064	0.118	0.177
CNITBE(rho = 0.01)	0.162	0.256	0.395	0.507
CNITBE(rho = 0.02)	**0.171**	**0.278**	0.414	0.527
CNITBE(rho = 0.1)	0.166	0.270	**0.417**	**0.533**
CNITBE(rho = 0.25)	0.131	0.218	0.380	0.496

Table 1: Comparison Results between different document embedding generation methods.

4 Conclusion

In this paper, we proposed to combine the test-based embedding and citation network structure to generate document embedding. Our proposed retraining method can outperform both popular test-based and network-based document embedding generation methods. In future work, we will explore to make use of more advanced deep learning techniques (e.g. LSTMs) for learning semantic and network representations and further improve the document embedding performance.

References

Doslu M, Bingol H O. 2016. *Context sensitive article ranking with citation context analysis*. Scientometrics, 2016, 108(2):653-671.

Le Quoc and Mikolov Tomas. 2014. *Distributed Representations of Sentences and Documents.* In Proceedings of the 31st International Conference on Machine Learning, Beijing, China.

Lim K W, Buntine W. 2016. *Bibliographic analysis on research publications using authors, categorical labels and the citation network.* Machine Learning, 2016, 103(2):185-213.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg, and Dean Jeffrey. 2013. *Distributed representations of phrases and their compositionality.* In Advances on Neural Information Processing Systems.

Rendle Steffen, Freudenthaler Christoph, Gantner Zeno, Schmidt-Thieme Lars. 2009. *BPR: Bayesian Personalized Ranking from Implicit Feedback.* Appears in Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence.

Tang Jian, Qu Meng, Wang Mingzhe, Zhang Ming, Yan Jun, Mei Qiaozhu. 2015. *LINE: Large-scale Information Network Embedding.* WWW 2015, Florence, Italy.

Yang Weiwei, Boyd-Graber Jordan, Resnik Philip. 2016. *A Discriminative Topic Model using Document Network Structure.* Meeting of the Association for Computational Linguistics. 2016:686-696.