Scientific Article Recommendation by using Distributed Representations of Text and Graph

Shashank Gupta
International Institute of Information Technology
Hyderabad
India
shashank.gupta@research.iiit.ac.in

Vasudeva Varma
International Institute of Information Technology
Hyderabad
India
vv@iiit.ac.in

ABSTRACT

Scientific article recommendation problem deals with recommending similar scientific articles given a query article. It can be categorized as a content based similarity system. Recent advancements in representation learning methods have proven to be effective in modeling distributed representations in different modalities like images, languages, speech, networks etc. The distributed representations obtained using such techniques in turn can be used to calculate similarities. In this paper, we address the problem of scientific paper recommendation through a novel method which aims to combine multimodal distributed representations, which in this case are: 1. distributed representations of paper's content, and 2. distributed representation of the graph constructed from the bibliographic network. Through experiments we demonstrate that our method outperforms the state-of-the-art distributed representation methods in text and graph, by 29.6% and 20.4%, both in terms of precision and mean-average-precision respectively.

1. INTRODUCTION

With the rapid increase in the number of scientific articles published each year, it becomes difficult to keep track of relevant research papers which are of interest to an user. Development of a good content based recommendation system, which can find relevant articles given an user's interest is important to deal with this problem.

The task of content based matching is challenging, mainly due to the problem of vocabulary mismatch. Methods like Latent Semantic Indexing (LSI) [2] and Latent Dirichlet Allocation (LDA) [1] attempt to solve this problem through latent space models. In these methods documents are represented in a latent space with the assumption that similar documents are mapped closer to each other in this space. [9] attempt to use these latent variable models for scientific article recommendation. While it can be argued that these methods bridge the vocabulary mismatch between query and candidate document to some extent, and yet they fail to capture the semantic relationship between them. Recent advances in representation learning methods are an attempt to address the aforementioned problem. Methods like word2vec [5] and doc2vec [4] capture semantic information at word level and document level respectively. Methods like DeepWalk [6] capture semantic information in a network setting. Due to the citations present in the scientific articles, a bibliographic network can be formed and used for enhancing scientific article recommendation's performance. Cluscite [7] proposes an approach to use heterogeneous bibliographic network for the task of scientific article recommendation.

In this paper, we consider content information and network information as two different modalities and propose a novel method to combine the two modalities using Canonical Correlation Analysis (CCA) [3]. Our main contributions are as follows: (1) We investigate the application of recently proposed representation learning methods for graphs in the context of scientific article recommendation.(2) We propose a novel method to combine the network representation from the bibliographic network with the content information. (3) We demonstrate applicability of our method on a large scale scholarly dataset. Our method outperforms the state-of-the-art method by 29.6% in precision@10.

2. PROPOSED APPROACH

We first discuss baseline methods used in our experiments and then our proposed approach.

Text Similarity Methods: For baseline, we consider TF-IDF, BM25 ranking based methods. We use lucene[1] to conduct experiments with these methods.

Representation Learning Methods: For baseline we use LSI, LDA, word2vec and doc2vec, which are popular representation learning methods for text. We use gensim[2] to conduct experiments with these models. For representation learning on graphs we use DeepWalk [6], which learns distributed representation of nodes in a graph by optimizing skip-gram [5] like objective on random walk sequences of nodes in a graph.

Proposed Method: We propose a model which combines network embeddings with the semantic embeddings of the text using CCA. CCA is a popular method to combine two different modalities by projecting both in a common representation space with the objective of maximizing the correlation between both modalities in the common space. Formally, given two views of data, $Q = \{q_1, q_2, ..., q_n\}^T \in \mathbb{R}^{nxd_q}$

[1] https://lucene.apache.org/

[2] https://radimrehurek.com/gensim/

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04..
http://dx.doi.org/10.1145/3041021.3053062

Table 1: Comparison of Various Methods (Dimension=64 for DeepWalk and 300 for all other methods. CCA(α) indicates value of α.)

Method	p @10	MAP @10	NDCG @10	p @100	NDCG @100	MAP @100
TF-IDF	0.1966	0.3341	0.5277	0.0542	0.4790	0.2402
BM25	0.1942	0.3305	0.5230	0.0515	0.4744	0.2384
LSI	0.1210	0.2257	0.3801	0.0402	0.3650	0.1609
LDA	0.0710	0.1503	0.2649	0.0235	0.2798	0.1179
D2V	0.0930	0.2052	0.3611	0.0256	0.3500	0.1565
D2V+DW	0.3497	0.4544	0.6408	0.0986	0.6201	0.3576
CCA(0.1)	0.4445	0.5399	0.7134	0.1080	0.6897	0.4431
CCA(0.05)	**0.4532**	**0.5472**	**0.7193**	**0.1086**	**0.6961**	**0.4510**

and $V = \{v_1, v_2, ..., v_n\}^T \in \mathbb{R}^{nxd_v}$, where d_q is the dimensionality of the first view, and d_v is the dimensionality of the second view. The objective of CCA is to find a pair of linear transforms defined as:

$$\begin{aligned} f(q_i) &= q_i W^0 \\ f(v_j) &= v_j W^1 \end{aligned} \tag{1}$$

such that the correlation between the transforms is maximized, which is defined as:

$$(W^0, W^1) = \underset{(W^0, W^1)}{\operatorname{argmax}} \operatorname{corr}(QW^0, VW^1) \tag{2}$$

where d is the dimensionality of the common space and $W^0 \in \mathbb{R}^{d_q xd}$ and $W^1 \in \mathbb{R}^{d_v xd}$ are linear transform matrices of the two views respectively. For our experiments we transform distributed representation from doc2vec and distributed representation from DeepWalk using eq. 2.

We consider distributed representations of textual content W_t, and graph W_g, as two different views of an user, and we use CCA to find a common representation space. We then use a linear combination of the resulting projections as the final representation, defined as:

$$W = \alpha W_t + (1 - \alpha) W_g \tag{3}$$

where W_g is the projection from DeepWalk, and W_t is the projection from Doc2vec, and α controls the relative contribution of each projection.

3. EXPERIMENTS

3.1 Dataset and Experimental Settings

For our experiments we use a large scale scholarly article dataset from Arnetminer [8]. It is a collection of research articles from the period of 1958-2014. As part of preprocessing, we filtered out incomplete records and records with less than 10 citation counts. For pre-processing text, we lower-cased the text and removed stop-words.

We consider a document as relevant to a research article if it is cited by that article, otherwise it is considered as irrelevant. We use Precision, MAP and NDCG as evaluation metrics. For training the model, we consider articles from the period of 1958-2011. For testing, we consider articles from the period of 2012-2015. There are 74097 articles for training, with 489828 number of citation pairs, 618 articles for testing, with 8532 number of citation pairs. We use dimension of 64 for DeepWalk, and a dimension of 300 for all distributed representation methods for text. We use a dimension of 64 for the projection from CCA.

3.2 Results and Analysis

We present the results of our method in Table 1. Among the traditional lexical matching methods, TF-IDF performs better than BM25 ranking method. Surprisingly, LSI performed better than all representation learning methods in text, including LDA and doc2vec. The concatenation of representations of doc2vec and DeepWalk performed better than all the methods mentioned above.

The representation learnt from the combination of DeepWalk and doc2vec from eq.3, outperforms all the methods described above. α is the controlling parameter which indicates the contribution of each representation in the final representation. The experimental results from simple concatenation of DeepWalk and doc2vec indicates a stronger influence of DeepWalk on the performance. Keeping this observation in mind, we keep the contribution of DeepWalk stronger in the final representation. The final representation with the value of alpha set to 0.05 performed best across all measures.

4. CONCLUSIONS

In this paper, we present a novel method to combine distributed representations of scientific article's content and graph from the bibliographical network using CCA. We investigate its application in similar article recommendation and demonstrate that our method outperforms traditional content based methods and simple combination strategy using concatenation. It would be interesting to experiment with non-linear methods of combination instead of a linear one.

5. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3(Jan):993–1022, 2003.

[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[3] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[4] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[6] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *SIGKDD*, pages 701–710. ACM, 2014.

[7] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. Cluscite: Effective citation recommendation by information network-based clustering. In *SIGKDD*, pages 821–830. ACM, 2014.

[8] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *SIGKDD*, pages 990–998. ACM, 2008.

[9] C. Wang and D. Blei. Collaborative topic modeling for recommending scientific articles. In *SIGKDD*, pages 448–456. ACM, 2011.