


Concept-Based Approach for Research Paper Recommendation

Ritu Sharma^(✉), Dinesh Gopalani, and Yogesh Meena

Malaviya National Institute of Technology, Jaipur, India
ritu.sharma3@hotmail.com, {dgopalani.cse,ymeena.cse}@mnit.ac.in
<http://www.mnit.ac.in>

Abstract. Research Paper Recommender Systems are developed to deal with the increasing amount of published information over web and provide recommendations for research articles based on the user preferences. Researchers invest their huge time in literature search to carry out the research work. To provide ease in building literature and finding useful research articles in less time, a novel concept-based recommendation approach is proposed that represents research article in terms of its concept or semantics, used to recommend conceptually related papers (based on the higher relevance of concepts) to researchers. This paper provides a brief overview of popular algorithms and previous systems developed to solve the problem of information explosion. Then, discuss the proposed approach with implementation details and a comparative analysis is presented between the proposed approach and baseline method.

Keywords: Research paper recommendation · Recommender system · Distributed representation · Concept-based approach · Semantics · Paragraph vector

1 Introduction

Due to the increasing number of articles, researchers find it difficult to search for the suitable paper. Recommender systems were developed to solve the problem of information overload and provide suggestions to choose appropriate article from a large set of available articles. Research paper recommender system finds relevant papers based on the users current requirements which can be gathered explicitly or implicitly through ratings, user profile, and text reviews.

Renown algorithms for paper recommendation are content-based filtering, collaborative filtering, co-occurrence based approaches and citation-based algorithms [1]. Content-based approaches works on the document text to find similarity between articles. For this purpose, most commonly used technique is bag-of-words (BOW) model. As this model works on word-matching principle, it do not consider the natural language ambiguities like synonym, polysemy, homonym etc. To overcome the drawbacks of existing approaches, we introduce a novel approach for recommendation which captures the semantic meaning of documents

via distributed representation and is used to recommend conceptually relevant papers. This paper organizes as follows: Sect. 2 presents a short survey on previous recommendation system, its approaches and their shortcomings. Section 3 discusses the proposed recommendation approach in detail. Section 4 examines the training parameter, evaluation scheme and implementation results. Finally, Sect. 5 concludes this work with future directions. In this paper, we will use the terms research paper, article and document interchangeably.

2 Related Work

The very first initiative in this field was the development of CiteSeer autonomous indexing system [6,8] which helps in developing the background knowledge by providing research articles that cites a given paper and it also displays the context of citation. Open archives were developed to offer storage and sharing of information along with recommendation services [9]. Various literature management tools were developed like Papits [15], an academic literature suite, Docear [2,3] to facilitate writing, sharing, retrieving, classifying, annotating and recommending research articles.

Most of the research paper recommender systems use content-based approaches based on key-phrase searching [7,10,13] and document content analysis [14]. Most commonly used technique for natural language processing is Bag-of-Words model and its variants like Term Frequency-Inverse Document Frequency (TF-IDF), bag-of-n-grams etc. The basic model represents text as set of words and forms a matrix where each column represents a unique term from vocabulary and row represents the document vectors. These vectors are the sequence of 0s and 1s, 0 indicates the absence of words and 1 is used to show that the word is present in the document. Other variants this is widely used approach due to its simplicity but it has some limitations as it is completely based on the syntactic representation of the document, it is unable to capture word order and context. However, word order is considered by bag-of-n-grams in short context but strives against the curse of dimensionality. Both algorithms have little knowledge about the word semantics.

To bridge this gap, distributed representations came into existence which was first used by Bengio et al. for statistical language modeling [4]. These neural nets are used to learn a vector representation for each term, called word embedding. Later, concept of deep learning is applied to neural networks for developing deep architectures that outperform state-of-the-art in several applications [5].

3 Proposed Approach

This approach is based on the idea of representing every document in terms of its concept or semantics by constructing distributed representation in high dimensional space. This unique representation is used to find articles in accordance with the user requirements. The whole process of recommendation is bifurcated into two stages as vector generation for candidate papers and recommendation algorithm.

3.1 Vector Generation for Candidate Papers

To visualize documents (candidate papers) in high dimensional space where similar documents sharing related concepts appear in the same area of space, we employed an unsupervised algorithm called Paragraph Vector [12] which extend the methods for learning the word vectors and term ‘paragraph’ is used to refer text of variable length which is research document in our case. In this framework, every column of matrix D represents a unique vector for every document. Similarly, every word is mapped to a unique vector represented by a column in matrix W. Paragraph vectors are asked to predict the next word given a set of contexts, sampled from a sliding window that runs over a document. Stochastic gradient descent is used to train word vectors and paragraph vectors. At every step of training, fixed-length context is drawn from a random paragraph to figure out the gradient error and is used to update the model parameters. At the time of prediction, an inference step is performed to calculate vector for an unseen paragraph. Total words present in all papers forms the training vocabulary V which is used to train the given model.

More formally, consider a set of words $w_1, w_2, w_3, \dots, w_{|V|}$, the paragraph vector aims to maximize the below mentioned average log probability [12]

$$\frac{1}{|V|} \sum_{v=i}^{|V|-i} \text{log}p(w_v|w_{v-i}, \dots, w_{v+i}). \tag{1}$$

The model is trained for prediction tasks which is carried out using multi-class classifier. So, we have given equation [12]

$$p(w_v|w_{v-i}, \dots, w_{v+i}) = \frac{e^{y_{w_v}}}{\sum_j e^{y_j}}. \tag{2}$$

Here, y_j is un-normalized log probability for output word j which can be evaluated using following equation [12]

$$y = a + Kh(w_{v-i}, \dots, w_{v+i}; V). \tag{3}$$

where a, K denotes the softmax parameters and h is constructed using matrix D and concatenation of word vectors from W. When the training converges, every document is represented by a unique n-dimension vector that captures the semantic meaning of the document and these vectors are stored in database.

3.2 Recommendation Algorithm

Now, every research article is mapped to a unique n-dimension vector which is used to find research papers that interests to the target user. Preferences of researcher are recorded by collecting papers of their interest which are then transformed to n-dimension vector using the aforementioned algorithm. Later, a similarity measure is applied to derive likeness between the concepts of input

paper and the candidate papers. For this, we have used the cosine similarity as follows

$$S(I, C) = \frac{\sum_{j=1}^n I_j C_j}{\sqrt{\sum_{j=1}^n I_j^2} \sqrt{\sum_{j=1}^n C_j^2}} \tag{4}$$

where, I_j and C_j denotes the distributed representation for input paper and candidate paper respectively. Here, C_j belongs to C and C denotes the set of candidate papers. $S(I, C)$ varies between -1 and 1 . Value tending towards positive one shows higher relevance in concepts and is negative for distinct concepts. Based on the higher similarity, N-most relevant research papers are fetched and given as recommendation.

4 Results and Discussion

We have developed a set of candidate research papers for recommendation. These are available online in PDF format, transformed to text format for further processing. Vector generation module which is based on unsupervised algorithm, distributed-memory paragraph vector [12] takes text document as input.

We have prepared our own dataset by downloading research articles freely available over internet which are categorized into six specialized fields of computer Science. To train the model for generating feature vector, we have set the value of certain parameters and default value is used for the rest. Size of feature vector is set equal to 300 i.e. each document is represented by a unique vector of size 300. Initial learning rate is set to 0.025 which is nearly drop to 0 with a step size of 0.002. Minimum count parameter is kept equal to 5 which ignores all words with frequency less than 5 and context window size as 10 which denotes the number of words taken into account to predict the next word.

To evaluate the proposed algorithm, Normalized Discounted Cumulative Gain (NDCG) [11] is used which measures accuracy of the recommendation algorithm by assigning more weight to top-ranked documents and considers two relevance levels (relevant and irrelevant) through different gain values.

$$DCG(r) = \begin{cases} G(1) & \text{if } r = 1, \\ DCG(r - 1) + \frac{G(r)}{\log(r)} & \text{otherwise} \end{cases}$$

where, r specifies the document position in recommendation list. Here, binary notion denotes relevance level (0 and 1), depending on whether recommended articles are relevant or not. $G(r)$ is equal to 1 if research document is relevant to the user and is 0 for irrelevant papers.

After the realization of proposed approach and to validate the results, we used a set of 30 users for evaluation. For every user, 10-most relevant articles were retrieved and $NDCG@10$ is calculated. This measure was averaged over the entire set of users to determine the overall accuracy of our recommendation algorithm. We have also implemented the baseline model of content-based technique (BOW model) for the same set of data and results were compared with the proposed approach. Figure 1 shows the normalized discounted cumulative gain value

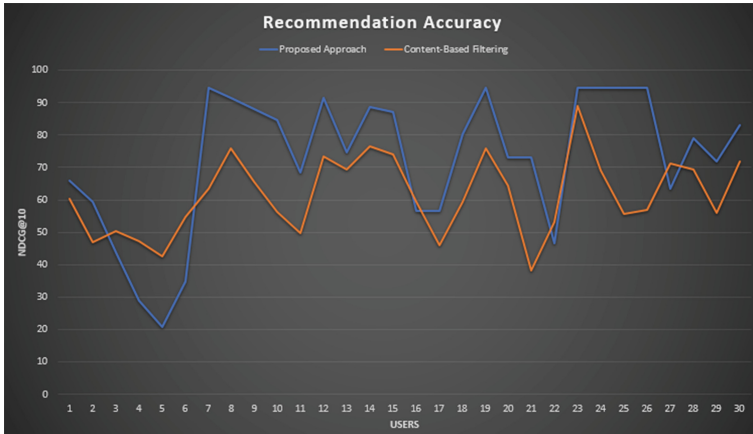


Fig. 1. Recommendation accuracy evaluated with $NDCG@10$

for all 30 users. These values are averaged over all the users and it is noted that our approach outperforms the baseline method. We have achieved recommendation accuracy ($NDCG@10$) of 61.38% for content-based filtering and 74.09% is recorded for the concept-based approach. Content-based algorithm achieved higher accuracy for papers where the same terminologies were used to represent similar concepts and in some cases, it is found that low similarity was predicted for related papers because authors have used different terms for indicating similar ideas. On the other hand, proposed method performs well and is able to recommend related documents to most of the users.

5 Conclusion and Future Work

Researchers have to spend a lot of time in searching for research papers of their interest. Content-based filtering is one of the most popular and widely used algorithms for research paper recommendation which is based on syntactic representation of document, so it is unable to capture word ordering and semantics. To overcome this limitation, we have proposed a novel concept-based approach that recommends research articles based on their semantic relatedness. Recommendation process is divided into two phases, first is vector generation which assigns a unique vector to every document and other is recommendation algorithm which make use of these vectors to recommend useful research articles. In future, this algorithm can be tested for finding the optimal value of parameters. Secondly, distributed representation of words can be combined to determine a unique vector for candidate documents which is further utilize to recommend papers, one can also compare it with the proposed algorithm.

References

1. Beel, J., Gipp, B., Langer, S., Breitingner, C.: Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.* **17**(4), 305–338 (2016)
2. Beel, J., Gipp, B., Langer, S., Genzmehr, M.: Docear: an academic literature suite for searching, organizing and creating academic literature. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, pp. 465–466. ACM (2011)
3. Beel, J., Langer, S., Genzmehr, M., Nürnberger, A.: Introducing Docear's research paper recommender system. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 459–460. ACM (2013)
4. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003)
5. Bengio, Y., et al.: Learning deep architectures for AI. *Found. Trends® Mach. Learn.* **2**(1), 1–127 (2009)
6. Bollacker, K.D., Lawrence, S., Giles, C.L.: Citeseer: an autonomous web agent for automatic retrieval and identification of interesting publications. In: Proceedings of the Second International Conference on Autonomous Agents, pp. 116–123. ACM (1998)
7. Ferrara, F., Pudota, N., Tasso, C.: A keyphrase-based paper recommender system. In: Agosti, M., Esposito, F., Meghini, C., Orio, N. (eds.) *IRCDL 2011. CCIS*, vol. 249, pp. 14–25. Springer, Heidelberg (2011). doi:[10.1007/978-3-642-27302-5_2](https://doi.org/10.1007/978-3-642-27302-5_2)
8. Giles, C.L., Bollacker, K.D., Lawrence, S.: Citeseer: an automatic citation indexing system. In: Proceedings of the Third ACM Conference on Digital Libraries, pp. 89–98. ACM (1998)
9. Gross, T.: Cyclades: a distributed system for virtual community support based on open archives. In: Eleventh Euromicro Conference on Parallel, Distributed and Network-Based Processing, 2003. Proceedings, pp. 484–491. IEEE (2003)
10. Hong, K., Jeon, H., Jeon, C.: Userprofile-based personalized research paper recommendation system. In: 2012 8th International Conference on Computing and Networking Technology (ICCNT), pp. 134–138. IEEE (2012)
11. Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 41–48. ACM (2000)
12. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: *ICML*, vol. 14, pp. 1188–1196 (2014)
13. Le Anh, V., Hoang Hai, V., Tran, H.N., Jung, J.J.: SciRecSys: a recommendation system for scientific publication by discovering keyword relationships. In: Hwang, D., Jung, J.J., Nguyen, N.T. (eds.) *ICCCI 2014. LNCS*, vol. 8733, pp. 72–82. Springer, Cham (2014). doi:[10.1007/978-3-319-11289-3_8](https://doi.org/10.1007/978-3-319-11289-3_8)
14. Philip, S., Shola, P., Ovyte, A.: Application of content-based approach in research paper recommendation system for a digital library. *Int. J. Adv. Comput. Sci. Appl.* **5**(10), 37–40 (2014)
15. Watanabe, S., Ito, T., Ozono, T., Shintani, T.: A paper recommendation mechanism for the research support system papits. In: International Workshop on Data Engineering Issues in E-Commerce, 2005, Proceedings, pp. 71–80. IEEE (2005)