SolutionTailor: Scientific paper recommendation based on fine-grained abstract analysis

Tetsuya Takahashi and Marie Katsurai[0000−0003−4899−2427]

Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto, Japan,
{takahashi, katsurai}@mm.doshisha.ac.jp

Abstract. Locating specific scientific content from a large corpora is crucial to researchers. This paper presents SolutionTailor[1], a novel system that recommends papers that provide diverse solutions for a specific research objective. The proposed system does not require any prior information from a user; it only requires the user to specify the target research field and enter a research abstract representing the user's interests. Our approach uses a neural language model to divide abstract sentences into "Background/Objective" and "Methodologies" and defines a new similarity measure between papers. Our current experiments indicate that the proposed system can recommend literature in a specific objective beyond a query paper's citations compared with a baseline system.

Keywords: paper recommendation system, sentence classification, sentence embedding

1 Introduction

Researchers usually spend a great deal of time searching for useful scientific papers for their continued research from a constantly growing number of publications in various academic fields. To assist academic search, various methods have been presented to recommend papers that approximate the user's interests and expertise. For example, content-based approaches often calculate sentence similarity between papers using natural language processing techniques, such as TF-IDF [9] and BERT [6]. Another line of research focuses on the user's past research history alongside co-authorship and citations and uses collaborative filtering or graph-based algorithms [3]. However, these methods usually focus on the semantic similarity of the overall content. Hence, the recommendation results often list papers that the user can easily access using a combination of research term queries or citation information in search engines. When considering the practicality of the usual literature survey, the search system must explain "why" and "how similar" the recommendated papers are to the user's interests.

To clarify the recommendation intention, this paper presents SolutionTailor, a novel system that recommends literature on diverse research methodologies in a specific research objective. Given a research abstract as a query representing the

[1] The demo video is available at: https://mm.doshisha.ac.jp/sci2/SolutionTailor.html

user's interests, the proposed system searches for papers whose research problems are similar but whose solution strategies are significantly different. To achieve this, we divide abstract sentences into background and methodology parts and propose a novel scoring function to answer the reason for the similarity. Users can specify a target research field for the search, so the system provides insights beyond the users' expertise.

The main contributions of this paper are twofold. First, we apply fine-grained analysis to abstracts to clarify the recommendation intentions. Second, we present evaluation measures that characterize the proposed system compared with a baseline system that uses full abstract sentences.

2 SolutionTailor Framework

2.1 Dataset construction

First, we prepared a list of research fields and their typical publication venues by referring to the rankings of the h5-index in Google Scholar Metrics[2]. Our current study uses the field "Engineering and Computer Science," comprising 56 subcategories, such as artificial intelligence, robotics, and sustainable energy. Then, from the Semantic Scholar corpus [2], we extracted papers whose publication venues were listed in the top-20 journals in each subcategory. The resulting dataset comprised 805,063 papers, all of which had English abstracts.

2.2 Sentence labeling and score calculation

Owing to the recent advances in neural language models, there are several studies on fine-grained scientific text analyses, such as classifying sentences into problems and solution parts [8] and classifying citation intentions [7]. Following this line of research, our study uses a BERT-based pretrained model [5] to classify abstract sentences into categories of "Background," "Objective," and "Method." If no sentence is clearly assigned to these categories, we choose the sentence having the highest probability of the corresponding labels. Then, we concatenate the sentences labeled with Background and Objective into a single sentence, from which we extract the embedded context vector using SciBERT, a BERT model pretrained using scientific text [4]. Using the resulting vectors, we compute the cosine similarity, cos_{BO}, between the Background/Objective sentences of the query abstract and the abstract of each paper in the database. SolutionTailor extracts the top-100 abstracts having the highest cos_{BO} from the target research field. These abstracts are recommendation candidates whose backgrounds are similar to the focus of the query abstract. Finally, for each recommendation candidate, we compute the final similarity score with the query abstract as follows:

$$score = cos_{BO} - cos_{M}, \tag{1}$$

[2] https://scholar.google.co.jp/citations?view_op=top_venues

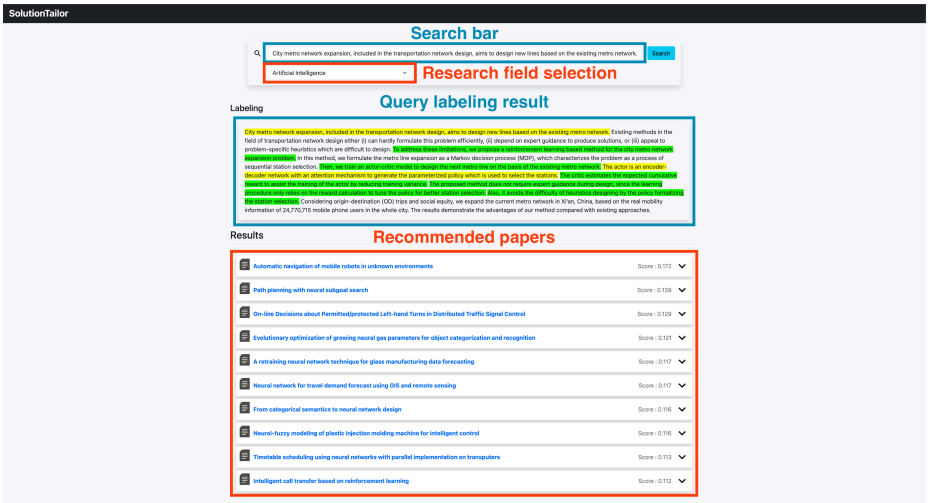

Fig. 1. Interface of the proposed system.

where cos_M denotes the cosine similarity between the method sentence vectors. A high score implies that the two abstracts have similar backgrounds, but presented different methodologies. Our system recommends the top-10 papers having the highest scores. This two-step search filters papers with irrelevant backgrounds and re-ranks the remaining ones in terms of methodological differences.

2.3 Recommendation interface

Figure 1 shows the interface of the proposed system. When a user inputs a research abstract as a query to the text box and selects a target category from the pull-down menu, SolutionTailor displays a list of top-ranked papers in terms of the similarity measure. By clicking on the title, the user can jump to the paper page in Semantic Scholar. The results of abstract sentence labels are highlighted in yellow and green, corresponding to Background/Objective and Method, respectively, to facilitate the interpretation of the recommendation results. The system shows detailed bibliographic information of each recommended paper as well as its abstract labeling result by clicking the toggle to the right of the score.

3 Evaluation

SolutionTailor uses the original similarity measure, whereas we can construct a baseline system that uses the embedded context vectors extracted from the full abstracts and their cosine similarities. We quantified the characteristics of the proposed system compared with the baseline system using two types of quantitative measures.

i) Overlap with citations: The first experiment used an abstract of an existing paper as a query and investigated whether the query paper itself cited the papers in the recommendation results. If the overlap between the query's citations and the recommended papers is low, it implies that the system provides new insights for the user into a specific research background. We selected "artificial intelligence" from the categories and calculated MAP@10 by querying 100 papers that contained at least five citations in the category. The total number of papers to be searched was 36,355. The MAP@10 scores of the proposed and baseline systems were 0.003 and 0.076, respectively.We should emphasize that the MAP may not be necessarily be high because we do not aim to predict the citations; this performance measure just characterizes the recommendation results. The lower MAP score implied that our system provides literature beyond the query paper's knowledge by using fine-grained sentence analysis.
ii) Similarity of objectives: The second experiment evaluated whether the similarity measure, cos_{BO}, used in the proposed system could actually find the same objective papers. Focusing on the fact that papers reporting results at a conference competition generally target the same research objective, we used the proceedings of SemEval-2021 [1], a workshop for the evaluation of computational semantic analysis systems. Each of the 11 tasks at SemEval-2021 had a single task description paper, and the papers submitted to a task were called "system description" papers. We used 11 task description papers as queries and evaluated whether the similarity measure, cos_{BO}, could appropriately recommend their system description papers.[3] The testing dataset for each query comprised $36,455$ $(36,355 + 100)$ papers used in the first experiment in addition to the system description papers of the target task. Each task had 15.91 system description papers on average, and a system should rank these higher than other papers. We removed the target task name and the competition name SemEval from all abstracts for fair experimental settings. As a result, the MAP scores obtained by our similarity measure, cos_{BO}, and the baseline system were 0.141 and 0.115, respectively, which demonstrated that our fine-grained sentence analysis could find the similarity of the objective more effectively than embedding the full abstracts.

4 Conclusion

This paper presented SolutionTailor, a novel system that recommends papers that provide diverse solutions to the same research background/objective. The system only requires text that summarizes the user's interests as a query, providing simple utility. The two types of the evaluation showed that our similarity measure provides insight beyond the user's cited papers and identifies the same-objective papers more effectively compared with the whole-text embedding. In future work, we will extend the dataset from engineering and computer science by adding more journals to be covered and conduct a subjective evaluation.

[3] We evaluated not the final similarity score but only cos_{BO} because the competition papers do not always have significantly different solutions.

References

1. Semeval-2021 tasks. https://semeval.github.io/SemEval2021/tasks.html (2020). Online; accessed 20 October 2021
2. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H.H., Peters, M., Power, J., Skjonsberg, S., Wang, L., Wilhelm, C., Yuan, Z., van Zuylen, M., Etzioni, O.: Construction of the literature graph in Semantic Scholar. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), pp. 84–91 (2018)
3. Asabere, N.Y., Xia, F., Meng, Q., Li, F., Liu, H.: Scholarly paper recommendation based on social awareness and folksonomy. International Journal of Parallel, Emergent and Distributed Systems **30**(3), 211–232 (2015)
4. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620 (2019)
5. Cohan, A., Beltagy, I., King, D., Dalvi, B., Weld, D.S.: Pretrained language models for sequential sentence classification. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
7. Ferrod, R., Di Caro, L., Schifanella, C.: Structured semantic modeling of scientific citation intents. In: R. Verborgh, K. Hose, H. Paulheim, P.A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, M. Alam (eds.) The Semantic Web, pp. 461–476. Springer International Publishing, Cham (2021)
8. Heffernan, K., Teufel, S.: Identifying problems and solutions in scientific text. Scientometrics **116**(2), 1367–1382 (2018)
9. Jomsri, P., Sanguansintukul, S., Choochaiwattana, W.: A framework for tag-based research paper recommender system: an IR approach. In: 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, pp. 103–108 (2010)