[Practitioners paper] A Novel Researcher Search System Based on Research Content Similarity and Geographic Information*

Tetsuya Takahashi, Koya Tango, Yuto Chikazawa, and
Marie Katsurai[0000−0003−4899−2427]

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

Abstract. Collaborative research is becoming increasingly important because it yields effective results and helps difficult research projects run smoothly. Previous studies have proposed many kinds of collaborator recommendation methods based on research features, such as specialty fields. However, few studies have constructed systems in which users can discover experts who have similar research interests using recommendation techniques. This paper proposes a novel researcher search system where users can efficiently discover potential candidates whose work locations are near theirs. Researchers are visualized on a map by our proposed system and users can use researcher's names and research keywords to narrow down the search. Specifically, given a researcher's name as a query, the system displays its relevant individuals based on either one of the following measures among researchers: research content similarity or collaborative relationship similarity. Our experiments demonstrated that recommendation results of these two similarity measures are minimally overlapped one another, indicating that our system could potentially help researchers discover collaborator candidates.

Keywords: researcher search system · collaborator recommendation · researcher similarity · academic database analysis

1 Introduction

Complex research project can be conducted effectively through collaborative research. Several studies have explored the relationship between collaboration and productivity. For example, Lee and Bozeman [11] investigated how the number of collaborators has influenced journal publication. Abramo et al. [1] analyzed the correlation among several types of collaborations, including interdisciplinary research, extramural collaborative research, industry-academia collaborative research, and their achievements, to assess the correlation between scientific productivity and collaboration intensity. Lopes et al. [13] proposed a method for

* This research was partly supported by JST ACT-I (Grant number: JPMJPR18UC), JST ACT-X (Grant number: JPMJAX1909), and JSPS KAKENHI (Grant number: 20H04484).

ranking research quality and found that authors of high-ranking research collaborated more. In the field of scholarly data mining, several collaborator recommendation methods have been proposed [14, 3, 10, 15, 5, 12, 2, 9, 16]. Most conventional methods defined the researcher similarity using bibliographic analysis, such as the closeness of existing relationships and correlation between research fields. However, few studies constructed systems in which users could discover researchers based on the existing recommendation methods.

Here, we focus on the recent work [16], which demonstrated that including the locations of researchers' affiliations improved research candidate recommender's performance. Inspired from this work, this study proposes a novel researcher search system based on research content similarity and geographic information to promote domestic collaboration opportunities. The proposed system allows users to search for potential collaborator candidates using researchers' names and research keywords. Subsequently, users can then filter results based on (i) researchers whose published works feature at least one of the keywords, (ii) existing collaborative partnerships among researchers when searching a researcher's name, and (iii) researchers whose interests or collaborative partnerships similar to the query researcher's those. In particular, in the third function, users can use the research content or existing collaborative relationships as features to calculate the similarity among researchers. Search results are displayed on a map of Japan using yellow pins. When a user clicks a pin, our system displays the researcher's information, including their specializations and past research projects. Our system enables users to discover researchers with desired specializations from their neighboring area, which helps encourage collaboration and discussion between researchers and local research institutions. We constructed the system using the Database of Grants-in-Aid for Scientific Research (KAKEN)[1]. Our experiments showed that there is little overlap between researchers found using content similarity and those in existing collaboration relationships, indicating that the proposed method can effectively help users to find potential collaborator candidates using these two different similarity measures.

2 Proposed System

2.1 System overview

The proposed system's architecture is shown in Fig. 1. Researcher similarity is regularly analyzed on a local computer, and the results are stored in a database on the cloud. To fulfill an API request from a client, the server cuts down a large network to extract a small network. By processing response sent from the server to the client, users can see results interactively. The client's home screen is shown in Fig. 2. When users click the search button, a search query field appears, as shown in Fig. 3. Users can use search queries appearing as candidates for the query. Clicking a yellow pin displays that researcher's details (Fig. 4). Users have the option to have the pins displayed on an aerial photo (as in Fig. 4) or on a

[1] https://kaken.nii.ac.jp/

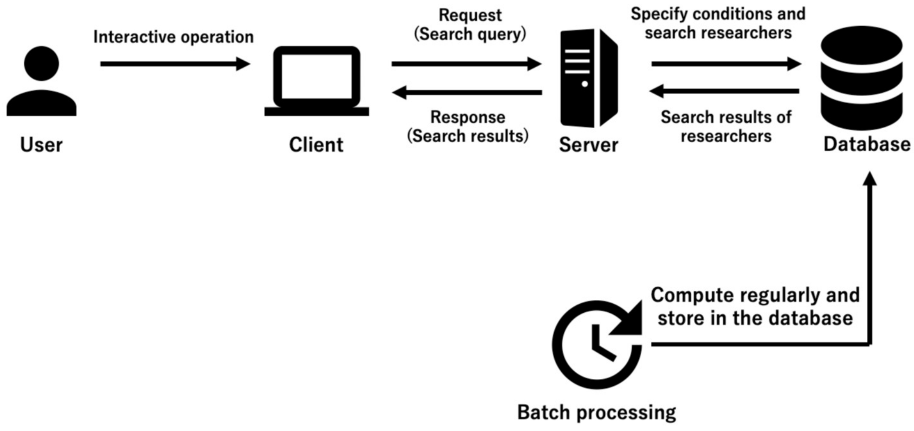


Fig. 1. The proposed system's architecture.

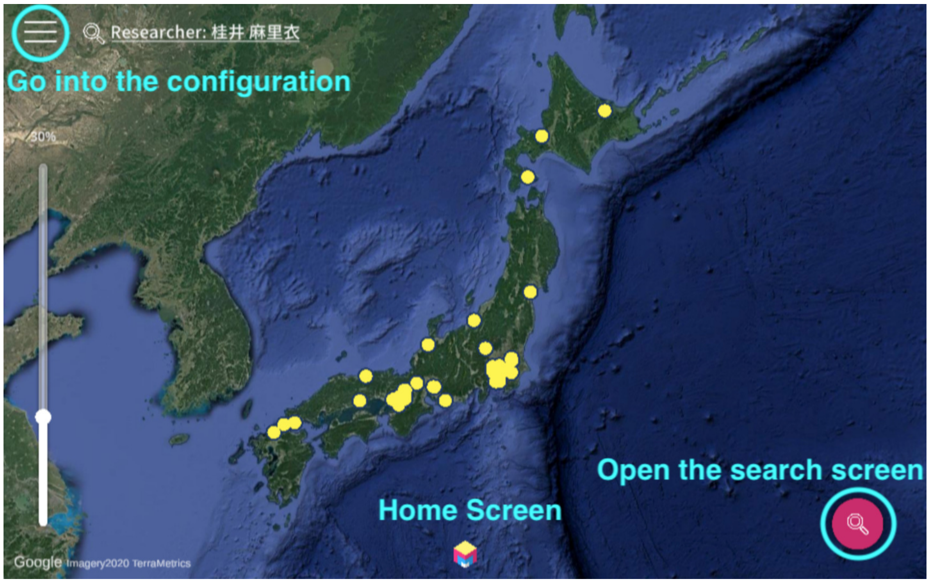
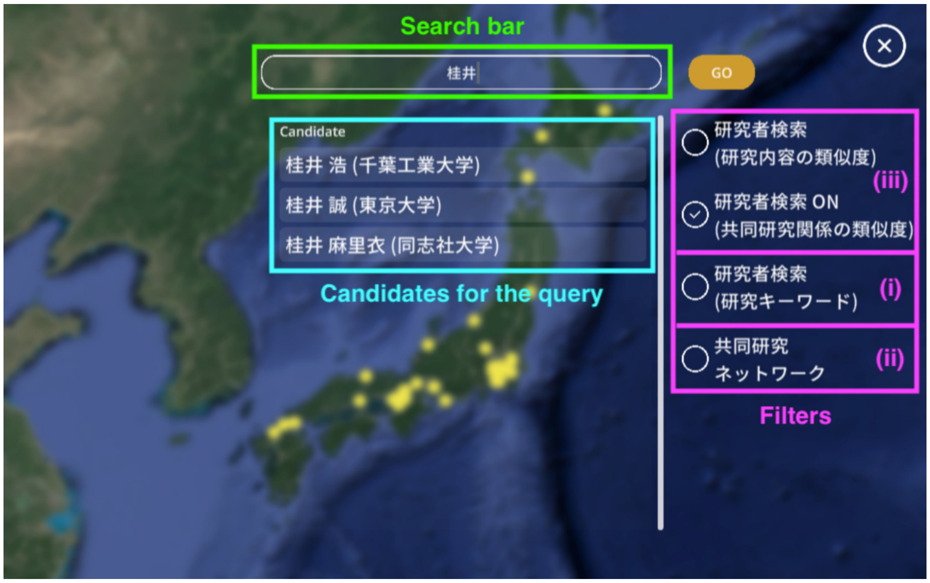


Fig. 2. The proposed system's home screen.

street map (as in Fig. 5). Hence, local information, such as train stations near the researcher's office, can be confirmed.

Users can filter search results by checking corresponding boxes, allowing them to tailor results based on (i) researchers whose published works feature at least one of the keywords used, (ii) existing collaborative partnerships among researchers when searching a researcher's name, and (iii) researchers whose in-



桂井

Candidate

桂井 浩 (千葉工業大学)

桂井 誠 (東京大学)

桂井 麻里衣 (同志社大学)

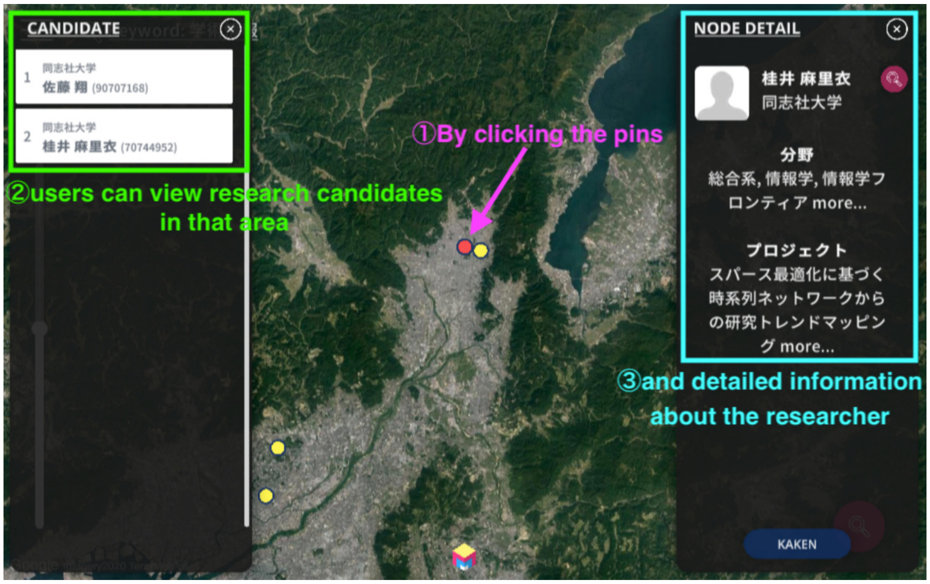
○ 研究者検索 (研究内容の類似度)

✓ 研究者検索 ON (共同研究関係の類似度)

○ 研究者検索 (研究キーワード)

○ 共同研究 ネットワーク

Fig. 3. The proposed system's search interface.


①By clicking the pins
②users can view research candidates in that area
③and detailed information about the researcher
KAKEN

CANDIDATE

1 同志社大学 佐藤 翔 (90707168)

2 同志社大学 桂井 麻里衣 (70744952)

NODE DETAIL

桂井 麻里衣

同志社大学

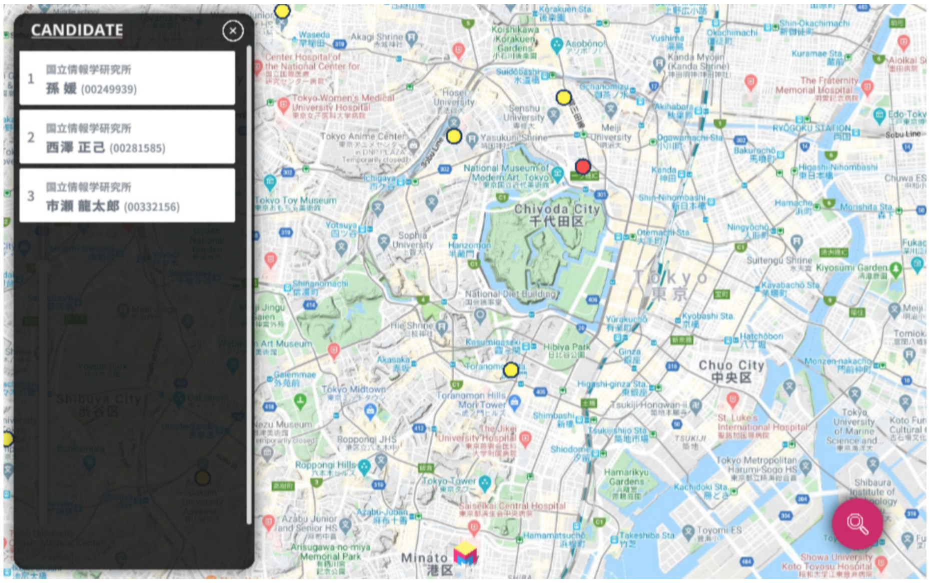
分野

総合系, 情報学, 情報学フロンティア more...

プロジェクト

スパース最適化に基づく時系列ネットワークからの研究トレンドマッピング more...

Fig. 4. Detail screen of a researcher.



CANDIDATE

1 国立情報学研究所
孫 媛 (00249939)

2 国立情報学研究所
西澤 正己 (00281585)

3 国立情報学研究所
市瀬 龍太郎 (00332156)

Fig. 5. Visualization by the street map mode.

terests or collaborative partnerships similar to the query researcher's those. Section 2.2 describes how to construct a dataset to implement the first and the second functions, while Section 2.3 describes two similarity measures to realize the third function.

2.2 Dataset construction and basic search

This subsection describes our system's dataset that consists of researcher information (i.e., research projects, work location information, and collaborative relationships). In this study, we used KAKEN to construct the researcher search system. KAKEN is the Database of Grants-in-Aid for Scientific Research projects granted by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Society for the Promotion of Science (JSPS). All of the database's research reports cover all research fields. Compared to other academic databases, it is better because it has a lower field deviation and all data can be searched using the same form. KAKEN assigns unique numbers to researchers, which are linked to their research projects, which display project names, principal investigators, co-investigators, research institutions, research fields, research keywords, summaries, and corresponding registered researchers' achievements. In this paper, we constructed the dataset using 911,724 research projects and 259,509 registered researchers.

Geographic information: The proposed system displays institutions each researcher belong to using information based on latest research projects. Insti-

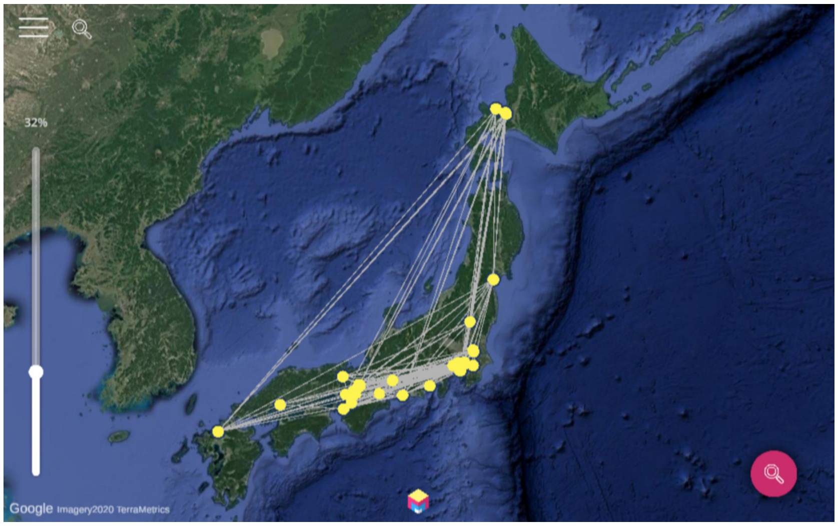


Fig. 6. A screenshot of the window for visualizing existing collaborative relationships among researchers.

tution name registered in KAKEN's projects had spelling inconsistencies, which we corrected manually. Work location information (i.e., latitude and longitude) was obtained using Google Maps API [2], which is visualized on a map of Japan.

Research keywords: In KAKEN, each researcher is assigned keywords based on the contents of their research. Users can use these keywords when performing searches. Researchers containing keywords corresponding to the search query are displayed on the map. Through this, users can find researchers who work in specific fields.

Existing collaborative partnerships: We constructed a network whose nodes were represented by researchers and whose edges were represented by collaborative relationships in KAKEN projects. When users search a researcher's name, the researcher's collaborative relationships are also displayed on the map. Specifically, the system obtains the network associated with a researcher ID from the server and displays the collaborative relationships, which enables users to find experts who actually work with the query researcher. Figure 6 shows an example of the relationships between researchers.

2.3 Computing researcher similarity and displaying potential collaborators

To search potential collaborators, users can choose the similarity of collaborative relationships or the similarity based on the contents of researcher's projects.

[2] https://cloud.google.com/maps-platform/

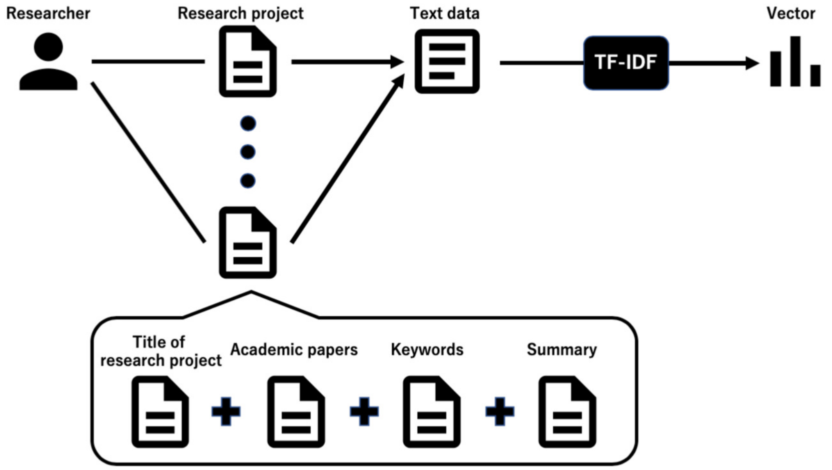


Fig. 7. The process of creating a vector that represents research content.

First, let R_{r_1}, R_{r_2} be two sets of research collaborators with each researcher r_1, r_2. Let the similarity of collaborative relationships between two researchers be represented by the Jaccard index as follows:

$$sim1(r_1, r_2) = \frac{|R_{r_1} \cap R_{r_2}|}{|R_{r_1} \cup R_{r_2}|}. \tag{1}$$

When there is no overlap between research collaborators, $sim1$ is 0.

Next, we compute the similarity between researchers based on the contents of their research. Each research project is considered to be a text data that reflects researcher's interests because each of them is influenced by each contributing researcher's achievements. Therefore, we computed the TF-IDF vector using the titles of research projects, titles of academic papers registered as achievement, research keywords, and the summary, as shown in Fig. 7.

Although the word embedding can be used to vectorize documents (as which was used in the related work [8]), we decided to use the TF-IDF vector to clearly understand coincidence of research interest in the proposed system. How to effectively combine these different word features will be discussed in our future work. Let $vec1$ and $vec2$ be the TF-IDF vectors calculated for researchers r_1 and r_2. We calculated the similarity based on the contents of their research using the cosine similarity:

$$sim2 = cosine(vec1, vec2). \tag{2}$$

The *sim*2 of 1 means the contents of their research is exactly the same.

In general, it is not realistic to compute similarity between over 100,000 researchers in any combination because it requires much calculation cost. Therefore, we used FAISS [7], which includes the nearest-neighbor search, to perform quick computation. FAISS reduces the dimensionality of a vector through product quantization [6, 4] and performs an approximate nearest-neighbor search. FAISS previously divided a set of vectors for search into Voronoï regions to improve search speed and specify search scope. In this study, there were 100 Voronoï, and the search scope was 10. Thus, 500 candidates for collaborative research were computed and saved in the database.

3 Evaluation Experiment

To show that the two methods for computing similarity described in 2.3 were useful for discovering potential collaborators, we determined the number of researchers yielded in the search results based on research contents. Specifically, we computed the overlap ratio of search results *overlap* as follows:

$$overlap(sim1, sim2) = \frac{|S_{sim1} \cap S_{sim2}|}{|S_{sim1}|}, \tag{3}$$

where S_{sim} denotes a set of researchers based on similarity *sim*. The lower $overlap(sim1, sim2)$ is, the more novel researchers excluded based on similarity *sim*2 are discovered using similarity *sim*1. In this experiment, 100 researchers were chosen at random and the overlap ratio was computed for all pairs. As a result, the mean value of *overlap*(sim1, sim2) is 0.199. Because the value is small, we demonstrated that it is possible to present potential collaborators that users cannot discover through existing relationships via switching these two similarity functions.

4 Conclusions and Future Work

This paper presented a novel researcher search system based on research content similarity and geographic information. In the proposed system, users can freely search for researchers in Japan using their names and research keywords. Specifically, we implemented three filters: researchers whose published works feature at least one of the keywords used, existing collaborative partnerships among researchers when searching a researcher's name, and researchers whose work or collaborative partnerships similar to the query researcher's those. It is expected that the proposed system will facilitate research collaboration and discussion among researchers and users that work near one another. In our future work, we will qualitatively evaluate the system's usability and consider more methods to compute similarity.

References

1. Abramo, G., D'Angelo, C.A., Di Costa, F.: Research collaboration and productivity: is there correlation? Higher Education **57**(2), 155–171 (2009)
2. Araki, M., Katsurai, M., Ohmukai, I., Takeda, H.: Interdisciplinary Collaborator Recommendation Based on Research Content Similarity. IEICE Transactions on Information and Systems **100**(4), 785–792 (2017)
3. Chen, H.H., Gou, L., Zhang, X., Giles, C.L.: CollabSeer: A Search Engine for Collaboration Discovery. In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. pp. 231–240. ACM (2011)
4. Ge, T., He, K., Ke, Q., Sun, J.: Optimized Product Quantization for Approximate Nearest Neighbor Search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2946–2953 (2013)
5. Guo, Y., Chen, X.: Cross-domain Scientific Collaborations Prediction with Citation Information. In: Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International. pp. 229–233. IEEE (2014)
6. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1), 117–128 (2010)
7. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. arXiv preprint arXiv:1702.08734 (2017)
8. Kawamura, T., Watanabe, K., Matsumoto, N., Egami, S.: Mapping Science Based on Research Content Similarity (07 2018). https://doi.org/10.5772/intechopen.77067
9. Kong, X., Jiang, H., Wang, W., Bekele, T.M., Xu, Z., Wang, M.: Exploring dynamic research interest and academic influence for scientific collaborator recommendation. Scientometrics **113**(1), 369–385 (2017)
10. Lee, D.H., Brusilovsky, P., Schleyer, T.: Recommending Collaborators using Social Features and MeSH Terms. Proceedings of the Association for Information Science and Technology **48**(1), 1–10 (2011)
11. Lee, S., Bozeman, B.: The Impact of Research Collaboration on Scientific Productivity. Social Studies of Science **35**(5), 673–702 (2005)
12. Li, J., Xia, F., Wang, W., Chen, Z., Asabere, N.Y., Jiang, H.: ACRec: A Co-authorship based Random Walk Model for Academic Collaboration Recommendation. In: Proceedings of the 23rd International Conference on World Wide Web. pp. 1209–1214. ACM (2014)
13. Lopes, G., Moro, M., Silva, R., Barbosa, E., Palazzo Moreira de Oliveira, J.: Ranking Strategy for Graduate Programs Evaluation (01 2011)
14. Lopes, G.R., Moro, M.M., Wives, L.K., de Oliveira, J.P.M.: Collaboration Recommendation on Academic Social Networks. In: Advances in Conceptual Modeling – Applications and Challenges. pp. 190–199. Springer Berlin Heidelberg (2010)
15. Tang, J., Wu, S., Sun, J., Su, H.: Cross-domain Collaboration Recommendation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1285–1293. ACM (2012)
16. Zhang, Q., Mao, R., Li, R.: Spatial–temporal restricted supervised learning for collaboration recommendation. Scientometrics **119**(3), 1497–1517 (04 2019)