Hindawi
Mobile Information Systems
Volume 2021, Article ID 6629520, 12 pages
https://doi.org/10.1155/2021/6629520

Research Article

Interinstitutional Research Team Formation Based on Bibliographic Network Embedding

O-Joun Lee,[1] **Seungha Hong**,[2] **and Jin-Taek Kim**[3]

[1]*Future IT Innovation Laboratory, Pohang University of Science and Technology, 77, Cheongam-ro, Nam-gu, Pohang-si 37673, Gyeongsangbuk-do, Republic of Korea*
[2]*Department of Computer Science and Engineering, Pohang University of Science and Technology, 77, Cheongam-ro, Nam-gu, Pohang-si 37673, Gyeongsangbuk-do, Republic of Korea*
[3]*Department of Creative IT Engineering, Pohang University of Science and Technology, 77, Cheongam-ro, Nam-gu, Pohang-si 37673, Gyeongsangbuk-do, Republic of Korea*

Correspondence should be addressed to Jin-Taek Kim; jintaek@postech.ac.kr

Received 3 December 2020; Revised 14 January 2021; Accepted 1 February 2021; Published 11 February 2021

Academic Editor: Jong M. Choi

This study aims at forming research teams for interinstitutional collaborations. Research institutes have their own purposes and topics of interest. Thus, supporting joint research between multiple institutes, we have to consider not only synergies between scholars but also purposes of the institutes. To solve this problem, we propose a bibliographic network embedding method that can learn characteristics of institutes, not only of each scholar. First, we compose a bibliographic network that consists of scholars, publications, venues, research projects, and institutes. Collaboration styles and research topics of institutes and scholars are extracted by mining subgraphs from the bibliographic network. Then, vector representations of network nodes are learned based on occurrences of subgraphs on the nodes and neighborhoods of the nodes. Based on the vector representations, we train multilayer perceptrons (MLP) to assess collaboration probability between scholars affiliated in different institutes. For training the MLP, we suggest three strategies: (i) considering every collaboration, (ii) focusing on interinstitutional collaborations, and (iii) focusing on collaboration outcomes. To evaluate the proposed methods, we have analyzed research collaborations of POSTECH (Pohang University of Science and Technology) and RIST (Research Institute of Industrial Science and Technology) from 2011 to 2020. Then, we conducted the research team formation for joint research of the two institutes according to two purposes: pure research and commercialization research.

1. Introduction

Research collaborations are one of primary features that affect performances of the research [1–6]. The existing studies for the research team formation concentrated on synergies between scholars [7–10]. To predict the synergies, analyzing or embedding bibliographic networks has recently been the most popular approach [11, 12]. These studies searched for adequate collaboration partners of each scholar by analyzing his/her research history. They supposed that structures of bibliographic networks reflect reputations (e.g., the number of citations), research topics (e.g., preferred venues), and even working styles (e.g., sustainability of collaborations) of scholars [1, 5, 6].

However, this approach does not consider that scholars are not the sole stakeholder of research. As employers and funding sources, research institutes influence research directions and outcomes of scholars. For interinstitutional research projects, the institutes evaluate team members and counterparts of their joint projects according to individual research interests and purposes, as we carefully choose our collaboration partners. For example, POSTECH (Pohang University of Science and Technology) is a research-oriented university. This institute encourages its members to publish research articles with scientific impact. On the other hand, RIST (Research Institute of Industrial Science and Technology) aims to develop practical technology and prefers patents rather than papers. Thus, when members of

POSTECH find their collaborators, they may prefer scholars who published many high-impact papers in other research-oriented institutes. However, if scholars in POSTECH want to commercialize their research outcomes, scholars in RIST can be a collaboration partner. Individual expertise and interest of the institutes can be discovered from bibliographic networks. Scholars in POSTECH will focus on papers rather than patents, and RIST might be contrary to POSTECH. Also, as a university, POSTECH covers much broader research areas than RIST. Thus, contributions from POSTECH will be published at more various venues compared to those from RIST.

A comparison of POSTECH with RIST shows differences caused by types of research institutes. However, within the same type, research institutes have individual characteristics according to their research interests. Figure 1 shows topic distributions of papers published by scholars in three major research-oriented universities in Korea. Although the three institutes share common research topics, their priorities for the topics are different. Also, the priorities can be correlated with infrastructures for each research field. In team formation for a project, we should match the project's research fields with participating institutes' expertise.

To conduct the research team formation, we should consider both characteristics of each of the scholars and their affiliation. There can be scholars who prefer intrainstitutional collaborations or are not familiar with collaborations. Scholars can also prefer particular types of institutes as collaboration partners (e.g., companies or universities). We can extract collaboration styles of both stakeholders (scholars and institutes) from the bibliographic networks. First, affiliations of collaborators of each scholar reveal what kinds of institutes are preferred by the scholar as collaboration partners. Second, venues of publications written by the scholars show their research interests. Finally, structures of the bibliographic networks represent more detailed research styles of the scholars, such as whether they focus on a few high-quality papers or write prolifically [5, 6]. Also, the structures can reveal working styles of research groups; for example, all group members focus on a research topic, the group leader manages multiple independent projects, or plural middle managers lead individual projects [1].

Therefore, in this study, we propose a method for forming research teams that can consider both collaboration styles of individual scholars and aims of research institutes by embedding bibliographic networks. First, this study suggests the interinstitutional collaboration network (Figure 2). This network includes information for research institutes and projects funded by the institutes, which are barely dealt with by the existing studies. Then, we apply the substructure-based graph embedding methods [1, 13–16] for representing scholars, publications, venues, research institutes, and projects with a fixed-size vector. Collaboration probabilities between scholars are estimated based on the vector representations. To consider individual characteristics of research institutes, we have the following assumptions:

(i) RQ 1. Interinstitutional collaborations have distinct characteristics from other types of collaborations.

(ii) RQ 2. Characteristics of research institutes affect collaborations between the institutes.

(iii) RQ 3. Research institutes have individual interests in topics, types of publications, and so on, and the interests affect employees of the institutes.

Based on the assumptions, we propose three approaches for the interinstitutional team formation: (i) considering every collaboration, (ii) focusing on collaborations between target institutes (based on RQ 1 and RQ 2), and (iii) focusing on collaboration outcomes preferred by the target institutes (based on RQ 1, RQ 2, and RQ 3). The three approaches were evaluated based on research outcomes of POSTECH and RIST from 2011 to 2020. By comparing (i) with the other two approaches, we can validate RQ 1. A comparison of (i) with (ii) can verify RQ 2. Finally, RQ 3 can be validated by comparing (iii) with the others and examining performances of the proposed methods for different types of publications (e.g., papers and patents). Contributions of this study can be categorized as follows:

(i) Modeling and embedding the interinstitutional collaboration network: This study proposes a novel bibliographic model representing interinstitutional collaborations and a model for embedding the proposed network. Finally, we propose three approaches for predicting collaboration probabilities by using the embedding vectors.

(ii) Discovering features of the interinstitutional team formation: The three approaches for team formation are based on individual features. The first one focuses on collaboration styles of each scholar. The second and third approaches consider collaboration styles of both research institutes and scholars and research interests of the institutes, respectively. Thus, experimental results for the approaches can exhibit these features' significance for the interinstitutional team formation.

(iii) Validating distinctiveness of interinstitutional collaborations: The comparisons between the three approaches also validate the fundamental assumptions of this study. The validation assures that we need specialized methods for composing interinstitutional research teams. Our findings can also be applied to other bibliography analysis tasks, such as predicting research institutes' performances and matching employers (institutes) and employees (scholars).

The remainder of this paper is organized as follows. Section 2 introduces the existing studies for the research team formation. In Section 3, we introduce the interinstitutional collaboration network, and we propose methods for embedding the network and for composing interinstitutional research teams. Section 4 explains experimental procedures for evaluating the proposed methods and validates their effectiveness based on the experimental results. Section 5 presents concluding remarks and future research directions.

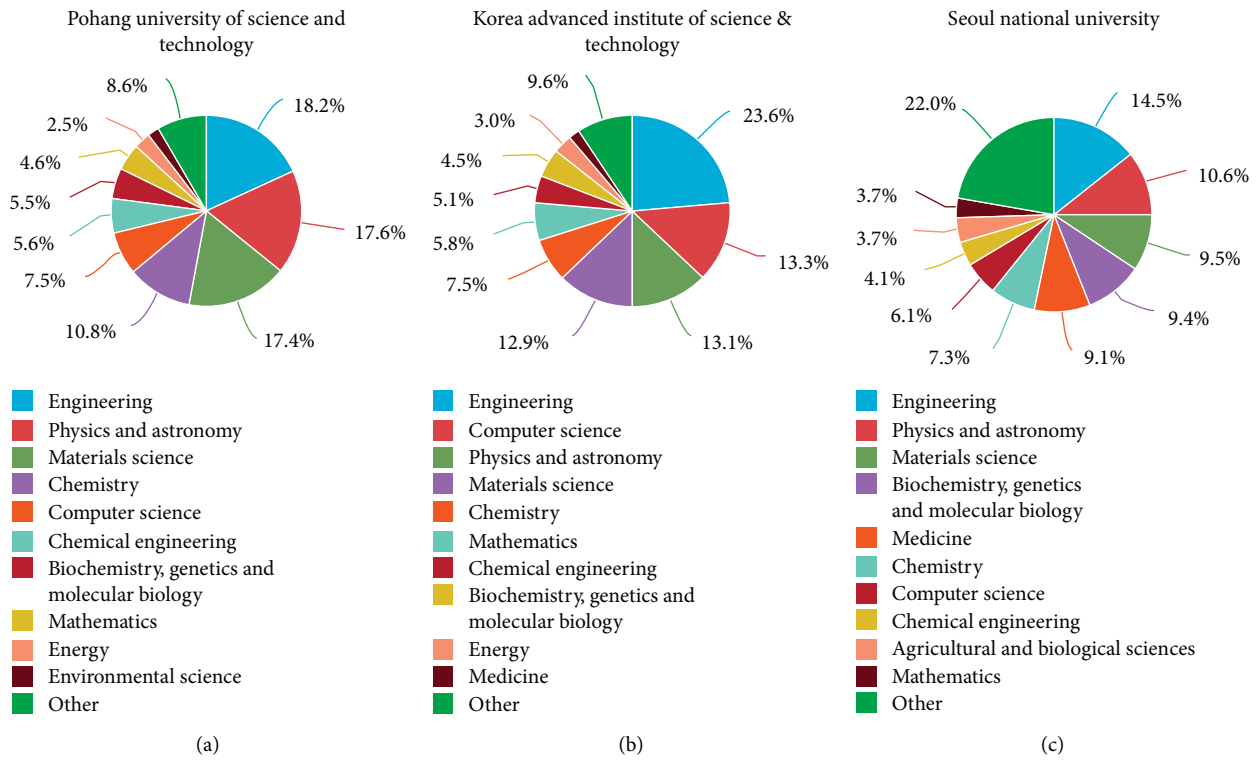

Figure 1: Topic distributions of publication records of three major research-oriented universities in Korea. These data and pie charts were acquired from affiliation profile pages on Scopus. Although they are the same type of research institutes with similar reputations and the same nationality, these institutes focus on individual research topics.

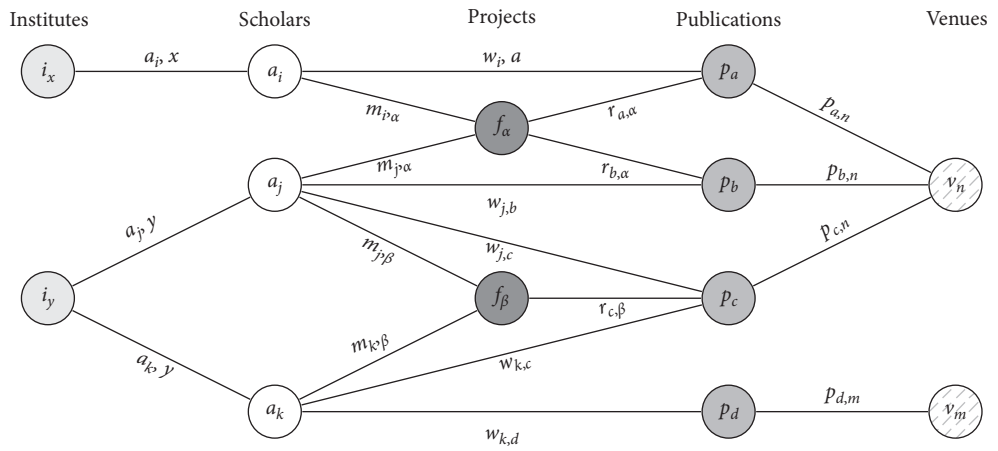

Figure 2: An example of the interinstitutional collaboration network. Circular nodes and edges indicate entities and relations in the proposed bibliographic networks, respectively. Also, colors of the nodes correspond to types of the nodes. Scholars (a_i) are affiliated in $(a_{i,x})$ research institutes (i_x), write $(w_{i,a})$ publications (p_a), and participate in $(m_{i,\alpha})$ projects (f_α). Publications (p_a) can be outcomes of $(r_{a,\alpha})$ the projects (f_α) and are published in $(p_{a,n})$ venues (v_n).

2. Related Work

There have not been studies for forming interinstitutional research teams. Although Purwitasari et al. [17] have proposed a team formation method for interdepartmental research collaboration, this method considers only topics of publications and does not consider departments/institutes as one of the stakeholders of research. Hernandez-Gress et al. [18] analyzed bibliographic data to recommend collaborations between universities by using only research topics of each scholar. Additionally, Guerrero-Sosa et al. [19] analyzed internal and external research collaborations of Universidad Autónoma de Yucatán, but their analysis results were limited in the data statistics. Therefore, this study

validates whether research institutes are significant stakeholders of research and proposes team formation methods that can consider the interests of both institutes and scholars.

Looking up from the interinstitutional research, there have been numerous studies for recommending research collaborators. Most of the existing studies applied link prediction techniques on bibliographic networks. They extracted various features from research publications or bibliographic networks by searching for scholars who can potentially (or sustainably [20, 21]) collaborate. Structures of bibliographic networks provide various information for bibliographic entities (e.g., scholars, publications, and venues) [1, 5, 6]. Regarding scholars, coauthorship relations show which types of collaborators are preferred by each scholar [1]. Temporal changes in coauthorship relations also reveal the sustainability of collaborations [6]. By analyzing structures of citation networks, we can extract publications' scientific impact and topical relevancy between publications [20]. Even without citations, relations between scholars and venues partially represent research topics of scholars [5]. Therefore, various studies [8–10, 12, 20–24] attempted to extract structural features of the bibliographic networks and to apply to predicting future coauthorship. To deal with the structural features, affinity propagation based on random walks was the most popular [9, 10, 20, 25]. However, recently, network embedding models enable us to represent the structural features by using low-dimensional fixed-length vectors [1, 5, 6, 12, 26]. Due to the vector representations, we can use conventional machine learning techniques to predict the collaboration probability without much modification.

There have been mainly two kinds of embedding models: proximity-based and structure-based models. If we employ proximity-based models [26], the obtained vector representations will have high similarity for scholars in the same community. However, we can search for collaborator candidates in a circle of acquaintance by ourselves. Also, some scholars prefer collaborators who come from diverse research groups [1]. Therefore, for the practicality of team formation methods, we have to provide unexpected collaborator candidates that are similar to previous collaborators of users. This study employs a structure-based network embedding model and modifies it to apply to the proposed bibliographic network.

Although bibliographic network structures reflect research topics of publications and scholars, they are difficult to be as accurate as analyzing the publications' content. Therefore, various studies applied topic modeling [7] and word/document embedding [9, 26] techniques to textual data in publications with an assumption that the scholars who deal with similar research fields can collaborate together [7–10, 12, 26, 27]. Obviously, information for research topics is valuable for team formation. If we make matches between two scholars in irrelevant domains, they are difficult to collaborate however talented they are. Nevertheless, this assumption cannot deal with forming interdisciplinary research teams, despite its significance for pioneering new research areas and providing practical experiences to scholars [28, 29]. We can also analyze probabilities of interdisciplinary research by combining the research topic information with bibliographic network structures. However, analyzing academic publications' content is out of coverage of this study. Our further research will attempt to cover the combination of two kinds of information.

Additionally, a few studies used statistical features extracted from bibliographic data. Bibliometrics (e.g., h-index) are effective to represent performance of scholars (and other kinds of bibliographic entities) with a single value [21, 27]. However, each of the bibliometrics reflects only fragmentary aspects of research. When a scholar wrote a few high-impact publications, another scholar published numerous intermediate publications, and they have the same h-index, it is not difficult to say which scholar has a better performance than the other. Even a few existing studies validated that network embedding models can reflect features represented by the bibliometrics [1, 5, 6]. Also, career ages of scholars were used in several existing studies [10, 21, 27, 30]. Nevertheless, this information is already included in bibliographic networks, and we do not always require collaborators who have similar career ages with us.

In summary, the existing methods have mainly two limitations. First, the existing studies suppose that scholars are the only stakeholder of research. However, as discussed in Section 1, research institutes have their own research interests and purposes. Also, scholars are influenced by the interest and purposes, as employees of the institutes. Second, sharing research topics or being active in the same research communities is not always good for research collaborations. To conduct research, which is a cooperative task, we need team members who can serve individual parts. Thus, a method that can consider both scholars' diverse roles and research institutes' purposes is required.

3. Interinstitutional Research Collaboration Prediction

This study aims at composing interinstitutional research teams by considering characteristics of both research institutes and their members. We have improved the conventional methods in terms of the three following points:

(i) The proposed bibliographic network model covers information for research institutes and projects.

(ii) Substructure-based graph embedding methods enable us to reveal research interests and expertise of institutes/scholars.

(iii) We propose the three approaches for learning collaboration history of target institutes. The approaches were evaluated and compared with each other in Section 4.

3.1. Interinstitutional Collaboration Network. Most of the existing studies only use coauthorship relations for analyzing/predicting collaborations. However, using solely coauthorship has difficulties for discovering characteristics of scholars and institutes in collaborations, such as research

interests, roles in research groups, and expertise. Therefore, we extend the conventional bibliographic network, which consists of scholars, publications, and venues, to cover research institutes and projects. The proposed network model is defined as follows.

Definition 1 (Interinstitutional Collaboration Network). This study defines the bibliographic network as a heterogeneous network, which has multiple kinds of nodes and relations. The bibliographic network ($\mathcal{N}$) contains five kinds of nodes: scholars ($\mathbb{A}$), publications ($\mathbb{P}$), venues ($\mathbb{V}$), institutes ($\mathbb{I}$), and projects ($\mathbb{F}$). Between these nodes, there are five kinds of relations: a scholar "writes" an academic publication ($\mathcal{W} \in \mathbb{R}^{|\mathbb{A}|\times|\mathbb{P}|}$), an academic publication is published in a venue ($\mathcal{P} \in \mathbb{R}^{|\mathbb{P}|\times|\mathbb{V}|}$), a scholar "is affiliated in" an institute ($\mathcal{A} \in \mathbb{R}^{|\mathbb{I}|\times|\mathbb{A}|}$), a scholar can "participate in" a project ($\mathcal{M} \in \mathbb{R}^{|\mathbb{A}|\times|\mathbb{F}|}$), and an academic publication can "be a result of" a project ($\mathcal{R} \in \mathbb{R}^{|\mathbb{F}|\times|\mathbb{P}|}$).

This can be formulated as

$$\mathcal{N} = \langle \mathbb{A}, \mathbb{P}, \mathbb{V}, \mathbb{I}, \mathbb{F}, \mathcal{W}, \mathcal{P}, \mathcal{A}, \mathcal{M}, \mathcal{R} \rangle. \quad (1)$$

Edges in the network represent only existence of the relations, and the edges connect only heterogeneous nodes (not necessary to annotate edge directions). Thus, the interinstitutional collaboration network is undirected and unweighted. Figure 2 illustrates an example of the bibliographic network, where $a_i \in \mathbb{A}$, $p_a \in \mathbb{P}$, $v_n \in \mathbb{V}$, $i_x \in \mathbb{I}$, $f_\alpha \in \mathbb{F}$, $w_{a,i} \in \mathcal{W}$, $p_{a,n} \in \mathcal{P}$, $a_{i,x} \in \mathcal{A}$, $m_{i,\alpha} \in \mathcal{M}$, and $r_{a,\alpha} \in \mathcal{R}$.

As shown in Figure 1, we can reveal characteristics of research institutes, such as their preferences for research fields, using only publication records. However, this information does not include collaboration styles of the institutes. Also, we assume that scholars' choices for their collaborators are different according to their and their collaborators' affiliations. This point can be revealed by a metapath, $\mathbb{I}$-$\mathbb{A}$-$\mathbb{P}$-$\mathbb{A}$-$\mathbb{I}$. This metapath represents preferences of research institutes for partner institutes. Research interests and aims of the institutes will also be reflected by $\mathbb{I}$-$\mathbb{A}$-$\mathbb{P}$-$\mathbb{V}$. Projects nodes enable us to know whether joint projects between target institutes have been successful or not ($\mathbb{I}$-$\mathbb{A}$-$\mathbb{F}$-($-\mathbb{P}$)-$\mathbb{A}$-$\mathbb{I}$). We can also analyze the sustainability of interinstitutional teams after the joint projects are finished. The sustainable teams will be benchmarks for composing productive research teams.

3.2. Bibliographic Network Embedding. Adjacency-based graph embedding methods (e.g., LINE [31]) can be effective for revealing preferences of scholars and research institutes. If a_i and a_j collaborate frequently and a_j wrote a number of publications with a_k, these methods will assign close vector representations to the three scholars. Then, a_k will be one of collaborator candidates of a_i with high priority. When all the three scholars have similar roles in their collaboration, this recommendation is reasonable. However, scholars with the same expertise will not have much motivation for collaboration. If a_j has been advising a_i and a_k as a domain expert, a_i and a_k will not have much reason to work with each other.

Our previous study [1] showed that substructure-based graph embedding methods can resolve this issue. These methods assign similar vector representations on nodes that have similar substructures. In the above example, if a_j prefers applying his/her own expertise to various domains, substructures rooted in a_j will have the star topology. The various domains will also be revealed by diversity of scholars and venues connected with a_j. Otherwise, a_i and a_k will be connected with less diverse venues than a_j.

This point is the same for discovering characteristics of research institutes and projects. Universities will have connections with more various venues than nonuniversity research institutes, which mostly have particular research fields. Also, participants of pure research projects will be members of universities rather than of companies. On the other hand, both universities and companies will participate in projects for technology commercialization.

Therefore, this study applies Subgraph2Vec [13], which aims at embedding subgraphs rooted in each node, on the bibliographic network. Subgraph2Vec consists of WL (Weisfeiler-Lehman) relabeling process [32] and Word2Vec [33]. This model assigns close vectors on subgraphs rooted in the same (or adjacent) nodes.

First, WL relabeling is a method for describing subgraphs rooted in each node exactly. This method assigns new labels on each node by using labels of itself and its adjacent nodes, iteratively. For example, a_i on Figure 2 has $\mathbb{A}$, which is its node type, as an initial label. At the first iteration, we check labels of neighborhoods of a_i, for example, $\mathbb{I}$ of i_x, $\mathbb{F}$ of f_α, and $\mathbb{P}$ of p_a. Then, a_i gets a new label, $\mathbb{A}$: $\mathbb{I}, \mathbb{F}, \mathbb{P}$. By iterating this process, scales of subgraphs represented by the labels become wider. To observe network structures with multiple scales, we call labels generated at the d-th iteration "subgraphs on degree d" and describe substructures rooted in a node as a set of the subgraphs. In practice, we sort the labels of neighborhoods and apply the hash function on the new label to avoid making redundant labels. Algorithm 1 presents procedures of the WL relabeling on our bibliographic network model, where $a_i^{(d)}$ indicates the subgraph rooted in a_i on degree d, $\mathcal{S}$ denotes a subgraph dictionary, and D refers to the maximum degree.

To apply Word2Vec on subgraphs, we have to define ranges of their neighborhoods. In texts, sentences are sequences of words, and neighboring words can easily be extracted using sliding windows. However, nodes in networks are not sequential. Therefore, we define neighborhoods based on adjacency of nodes and degrees as with the previous study [1]. Neighborhoods of $a_i^{(d)}$ can be formulated as

$$\begin{aligned}\mathcal{N}\left(a_i^{(d)}\right) = &\left\{p_a^{(d+\Delta d)} \mid w_{i,a} \neq 0, |\Delta d| \leq \mathcal{W}_D, \forall p_a \in \mathbb{P}\right\} \\ &\cup \left\{i_x^{(d+\Delta d)} \mid a_{i,x} \neq 0, |\Delta d| \leq \mathcal{W}_D, \forall i_x \in \mathbb{I}\right\} \\ &\cup \left\{f_\alpha^{(d+\Delta d)} \mid m_{i,\alpha} \neq 0, |\Delta d| \leq \mathcal{W}_D, \forall f_\alpha \in \mathbb{F}\right\},\end{aligned} \quad (2)$$

where $\mathcal{W}_D$ is a widow size for the degree. The same way is used to compose neighborhoods for other node types.

(1) **procedure** WLRELABELLING $(\mathcal{N}, \mathcal{S})$
(2) **for** d: $1 \longrightarrow D$ **do**
(3) **for** $a_i \in \mathbb{A}$ **do**
(4) $\mathcal{S}_i^{(d-1)} \longleftarrow \left\{p_a^{(d-1)} | w_{i,a} \neq 0, \forall p_a \in \mathbb{P}\right\} \cup \left\{i_x^{(d-1)} | a_{i,x} \neq 0, \forall i_x \in \mathbb{I}\right\}$
(5) $\cup \left\{f_\alpha^{(d-1)} | m_{i,\alpha} \neq 0, \forall f_\alpha \in \mathbb{F}\right\}$
(6) $a_i^{(d)} \longleftarrow \langle a_i^{(d-1)}, \mathcal{S}_i^{(d-1)} \rangle$
(7) Put $\left\{\text{HASH}(a_i^{(d)})\text{: } a_i^{(d)}\right\}$ into $\mathcal{S}$
(8) **for** $p_a \in \mathbb{P}$ **do**
(9) $\mathcal{S}_a^{(d-1)} \longleftarrow \left\{a_i^{(d-1)} | w_{i,a} \neq 0, \forall a_i \in \mathbb{A}\right\} \cup \left\{v_n^{(d-1)} | p_{a,n} \neq 0, \forall v_n \in \mathbb{V}\right\}$
(10) $\cup \left\{f_\alpha^{(d-1)} | r_{a,\alpha} \neq 0, \forall f_\alpha \in \mathbb{F}\right\}$
(11) $p_a^{(d)} \longleftarrow \langle p_a^{(d-1)}, \mathcal{S}_a^{(d-1)} \rangle$
(12) Put $\left\{\text{HASH}(p_a^{(d)})\text{: } p_a^{(d)}\right\}$ into $\mathcal{S}$
(13) **for** $v_n \in \mathbb{V}$ **do**
(14) $\mathcal{S}_n^{(d-1)} \longleftarrow \left\{p_a^{(d-1)} | p_{a,n} \neq 0, \forall p_a \in \mathbb{P}\right\}$
(15) $v_n^{(d)} \longleftarrow \langle v_n^{(d-1)}, \mathcal{S}_n^{(d-1)} \rangle$
(16) Put $\left\{\text{HASH}(v_n^{(d)})\text{: } v_n^{(d)}\right\}$ into $\mathcal{S}$
(17) **for** $i_x \in \mathbb{I}$ **do**
(18) $\mathcal{S}_x^{(d-1)} \longleftarrow \left\{a_i^{(d-1)} | a_{i,x} \neq 0, \forall a_i \in \mathbb{A}\right\}$
(19) $i_x^{(d)} \longleftarrow \langle i_x^{(d-1)}, \mathcal{S}_x^{(d-1)} \rangle$
(20) Put $\left\{\text{HASH}(i_x^{(d)})\text{: } i_x^{(d)}\right\}$ into $\mathcal{S}$
(21) **for** $f_\alpha \in \mathbb{F}$ **do**
(22) $\mathcal{S}_\alpha^{(d-1)} \longleftarrow \left\{a_i^{(d-1)} | m_{i,\alpha} \neq 0, \forall a_i \in \mathbb{A}\right\} \cup \left\{p_a^{(d-1)} | r_{a,\alpha} \neq 0, \forall p_a \in \mathbb{P}\right\}$
(23) $f_\alpha^{(d)} \longleftarrow \langle f_\alpha^{(d-1)}, \mathcal{S}_\alpha^{(d-1)} \rangle$
(24) Put $\left\{\text{HASH}(f_\alpha^{(d)})\text{: } f_\alpha^{(d)}\right\}$ into $\mathcal{S}$

Algorithm 1: WL relabeling process on the interinstitutional collaboration network.

To embed the subgraphs, we use the SkipGram and negative sampling [33]. This can be formulated as

$$\begin{aligned}\mathcal{L}\left(a_i^{(d)}\right) &= \sum_{\forall \mathcal{S}_a \in \mathcal{N}\left(a_i^{(d)}\right)} \log P\left(\mathcal{S}_a \mid \Phi\left(a_i^{(d)}\right)\right) \\ &\quad - \sum_{\forall \mathcal{S}_b \notin \mathcal{N}\left(a_i^{(d)}\right)} \log P\left(\mathcal{S}_b \mid \Phi\left(a_i^{(d)}\right)\right) \\ &\simeq \sum_{\forall \mathcal{S}_a \in \mathcal{N}\left(a_i^{(d)}\right)} \log \sigma\left(\Phi\left(\mathcal{S}_a\right)^\intercal \Phi\left(a_i^{(d)}\right)\right) \\ &\quad + \sum_{j=1}^{k} \mathbb{E}_{\mathcal{S}_b \sim P_n(\mathcal{S})}\left[\log \sigma\left(-\Phi\left(\mathcal{S}_b\right)^\intercal \Phi\left(a_i^{(d)}\right)\right)\right],\end{aligned} \tag{3}$$

where $P_n(\mathcal{S})$ denotes a noise distribution of subgraphs, k indicates the number of negative samples, and $\Phi(\cdot)$ denotes the projection function. In this study, $P_n(\mathcal{S})$ is a uniform distribution. $\Phi(a_i)$ is obtained by concatenating $\Phi(a_i^{(0)})$ to $\Phi(a_i^{(D)})$.

3.3. Research Collaboration Prediction. We use the conventional MLP (Multilayer Perceptron) model to predict interinstitutional collaborations. The MLP model consists of three fully connected layers and one drop-out layer. Inputs of the model are $2 \times \delta$-dimensional vectors composed by concatenating vector representations of two scholars. An activation function of this model's output layer is the sigmoid function, and the other layers use the ReLu (Rectified Linear Unit) function as their activation functions. This model predicts collaboration probabilities between two scholars, and scholar pairs are classified into two groups that are appropriate for collaboration and not. As a loss function, the binary cross entropy is applied.

In this study, we focus on the interinstitutional collaborations that should consider not only relationships between individual scholars but also relationships between research institutes and between scholars and institutes. Research institutes have their own purposes, and members of the institutes also should concentrate on occupational research. Therefore, we cannot ensure that training the model to predict every collaboration (scholar-publication-scholar relations) in the bibliographic network is the best approach for learning the individual characteristics of research institutes. Therefore, we propose two more approaches based on our research questions (in Section 1) to make the model reflect agendas of the target institutes and compare them with the conventional approach (i.e., learning all the previous collaborations). The three approaches for training the MLP model are as follows:

(i) Case 1: Learning all the collaboration relations in the bibliographic network.

(ii) Case 2: Learning previous collaborations between the target institutes (RQ 1 and RQ 2).

(iii) Case 3: Learning collaborations that produced similar publications to previous collaborations between the target institutes (RQ 1, RQ 2, and RQ 3).

The first case supposes that the bibliographic network embedding method can represent characteristics of research

institutes and their collaborations despite their diversity. Thus, this case assumes that scholars' vector representations include information for purposes and preferences of the scholars' affiliations. In this case, the MLP learns all the collaborations in the bibliographic network, as shown in Figure 3(b), and we use the trained model to predict probabilities of further collaborations between scholars from target institutes. Therefore, this approach makes the prediction model reflect the general characteristics of research collaborations. Although the general characteristics cover interinstitutional collaborations, this will not be as clear as focusing on only interinstitutional collaborations. Thus, we use this approach as a baseline for validating whether interinstitutional collaborations have distinctive characteristics compared to the others (RQ 1).

The second case, which is based on RQ 1 and RQ 2, focuses on searching for scholars that are appropriate for collaborations between the target institutes. There will be scholars who prefer collaborations but only intrainstitutional collaborations or only particular partner institutes. If a scholar has preferences according to reputations or types of institutes, our embedding model can extract the information from publications and venues connected with the institutes. The institutes will also concern whether the scholar can conduct research that they expect. For example, POSTECH and RIST are significant research partners of each other. However, not all the scholars in the two institutes participated in collaborative studies between the institutes. Thus, we can assume that there will be a certain type of scholars that are appropriate for mutual interests of the institutes. Therefore, this case uses bibliographic networks that consist of scholars in the target institutes as a dataset. Then, we train the MLP to predict whether a group of scholars from the respective institutes has previous collaborations, as shown in Figure 3(c). By comparing this case with the first one, we can reveal whether research institutes' characteristics affect their employees (RQ 2).

We have designed the third approach based on all the research questions (RQ 1, RQ 2, and RQ 3). This case especially concentrates on the fact that research institutes have individual agendas and preferable kinds of publications (RQ 3). Thus, we first find academic publications that are similar to outcomes of previous collaborations between the target institutes by clustering publications in our bibliographic network according to their vector representations. Then, we search for scholars who have written publications that are in the same clusters with the previous collaboration outcomes. We assume that scholars are capable of conducting research that the target institutes expect from their collaborations. When publications that come from collaborations between POSTECH and RIST are in cluster A, research groups that wrote publications in cluster A will let us know compositions of research groups that are appropriate for collaborations between the two institutes. Thus, in this approach, the MLP model learns only the research groups which produced research outcomes that are similar to the previous collaboration outcomes of the target institutes, as shown in Figure 3(d). By comparing this approach with the others, we can validate whether research institutes have preferences for types or topics of publications (RQ 3).

4. Evaluation

To evaluate the proposed methods, we predicted interinstitutional collaborations by analyzing previous collaboration history. Also, our research questions were validated by comparing the performances of the proposed methods with each other. We supposed that research institutes have preferences for topics and types of their members' research outcomes (RQ 2 and RQ 3). Thus, we should collect multiple types of academic publications, although the existing studies mostly dealt with one type. The multiple types caused a limitation in our experiments. Unlike papers with numerous well-organized academic databases (e.g., DBLP and Scopus), it is not easy to expect accurate publication records for patents or technical reports published by each research institute. Thus, we collected the paper dataset from the open academic databases and acquired a patent dataset by directly requesting it to research institutes. Due to this point, we could not conduct the experiments on a large-scale dataset for multiple research institutes. Nevertheless, publication records of research institutes include their collaborating institutes. Thus, the proposed methods made answers by analyzing hundreds of research institutes' characteristics, although they predict collaborations between a few institutes.

We collected papers and patents published by scholars in POSTECH and RIST from January 2011 to September 2020. The papers were gathered through the affiliation profile pages on Scopus1, and RIST provided bibliographic data for the patents. Our bibliographic network consists of the papers, patents, and every scholar/institute/venue connected with the papers and patents. We composed the network for two time periods: 2011–2015 and 2016–2020. The proposed methods were trained by the bibliographic data from 2011 to 2015 and validated based on the collaborations from 2016 to 2020. In our dataset, papers' author names are in English, and patents' inventor names are in Korean. Thus, we could not build a unified network for both types of publications. We constructed two separate networks and compared the performances of the proposed approaches on the two networks to validate whether research institutes (and their members) have distinct characteristics. Table 1 presents statistics of the bibliographic networks.

The three approaches proposed in Section 3.3 were evaluated based on accuracy for predicting collaboration outcomes between POSTECH and RIST. The accuracy was assessed using three metrics: precision, recall, and F_1 measure. When we measure accuracy of predicting collaborations between i_x and i_y, these metrics are calculated as

$$\begin{aligned}
\widehat{C}(i_x, i_y) &= \left\{ \langle a_i, a_j \rangle \mid a_{i,x} = 1, a_{j,y} = 1, p(a_i, a_j) \geq 0.5 \right\}, \\
p(i_x, i_y) &= \frac{\widehat{C}(i_x, i_y) \cap C(i_x, i_y)}{\widehat{C}(i_x, i_y)}, \\
r(i_x, i_y) &= \frac{\widehat{C}(i_x, i_y) \cap C(i_x, i_y)}{C(i_x, i_y)}, \\
F_1(i_x, i_y) &= 2 \cdot \frac{p(i_x, i_y) \cdot r(i_x, i_y)}{p(i_x, i_y) + r(i_x, i_y)},
\end{aligned} \tag{4}$$

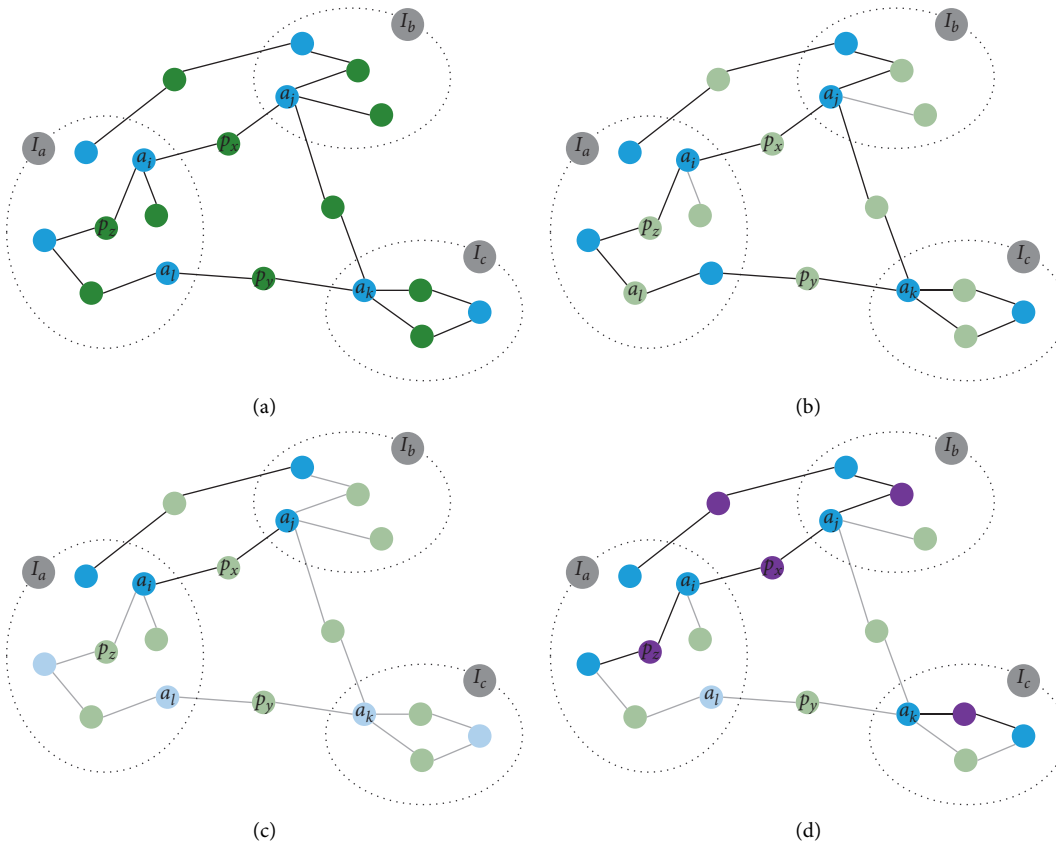

Figure 3: Assumptions for predicting interinstitutional research collaborations. (a) presents an example of the bibliographic network. (b) Case 1, (c) Case 2, and (d) Case 3 describe nodes and relations that are considered by the three proposed models to predict collaborations between I_a and I_b, respectively. Gray ellipses indicate research institutes, blue nodes refer to scholars, and green nodes indicate publications. Purple nodes in (d) denote publications that are in the same cluster.

Table 1: Statistics of datasets. This table compares our two datasets that consist of papers and patents published by scholars in POSTECH and RIST from January 2011 to September 2020.

	Papers	Value		Patents	Value
# Nodes	# scholars	35,692	# Nodes	# scholars	3,942
	# papers	19,524		# patents	5,001
	# venues	3,381		# venues	240
# Relations	# scholar-paper	113,462	# Relations	# scholar-patent	17,406
	# paper-venue	19,678		# patent-venue	7,612

"#" means "the number of."

where $\widehat{C}(\cdot,\cdot)$ and $C(\cdot,\cdot)$ are sets of predicted and actual collaborations between two institutes, respectively, and $p(\cdot,\cdot)$, $r(\cdot,\cdot)$, and $F_1(\cdot,\cdot)$ indicate precision, recall, and F_1 measure for predicting collaborations between two institutes, respectively. We compared the performances of the proposed approaches with a performance of a baseline method and also with each other. As the baseline, we use Case 1, one of the proposed approaches, to predict all the collaborations. A comparison of this case with the proposed approaches exhibits the necessity of methods specialized in predicting interinstitutional collaborations. Table 2 presents experimental results.

Additionally, we heuristically tuned hyperparameters of the proposed methods. The number of dimensions for subgraph vectors was 100, and the maximum degree was 4. The MLP model for predicting collaborations includes three fully connected layers that have 200, 150, and 80 nodes. The threshold of its drop-out layer was 0.2. Also, the number of epochs and learning rates were set as 50 and 0.0008, respectively.

4.1. RQ 1: Distinct Characteristics of Interinstitutional Research Collaboration. The motivation of this study is that we need a collaboration prediction method specialized in interinstitutional collaborations. The necessity can be validated by comparing the performance of Case 1 with the performance of Case All. These two cases present accuracies

Table 2: Experimental result.

Metrics		Case 1	Case 2	Case 3	Case All
Paper	Precision	**0.83**	0.77	0.66	0.86
	Recall	**0.73**	0.58	0.65	0.86
	F_1 Measure	**0.78**	0.66	0.65	0.86
Patent	Precision	0.72	**0.78**	0.74	0.80
	Recall	0.72	**0.77**	0.74	0.80
	F_1 Measure	0.72	**0.78**	0.74	0.80

The last column presents accuracy for predicting all the collaborations, and the remaining columns present accuracy of the proposed approaches for predicting collaborations between POSTECH and RIST.

of the same model for different targets. Case 1 shows accuracy for predicting interinstitutional collaborations, while Case All is for every collaboration. Therefore, the result that Case All had higher accuracy than Case 1 also underpins RQ 1.

The performance decrements between Case 1 and Case All were similar in predicting collaborated papers and patents. However, both Case 1 and Case All performed higher accuracy on papers than on patents. Otherwise, Case 2 and Case 3, which focus on the previous collaborations, performed higher accuracy on patents than on papers. We can assume that patents get more influence from the characteristics of interinstitutional collaborations than papers, although we should also consider that scholars in RIST barely write papers (62 papers from 2011 to 2020). For this point, we should experiment again with a larger dataset containing more institutes and publication types in further research.

Different diversities of publication types can also cause this result; papers are more diverse than patents. Precision and recall of Case 1 were similar to each other on predicting collaborated patents between POSTECH and RIST. However, its precision for predicting papers collaborated by the two institutes was much higher than its recall. This problem was worse in Case 2 that learns only the previous collaborated papers. Otherwise, Case All and Case 3 performed small deviations between their precision and recall on both patents and papers. These two cases might be less affected by the diversities of publications. Since Case All learned general characteristics of the research collaboration, this case gains capability for handling the diversities. On the other hand, Case 3 searched for scholars who can produce the same kinds of research outcomes as the previous collaborations between POSTECH and RIST. Thus, this case learned both the diversities and the two institutes' characteristics. Conclusively, both types of research publications were affected by the distinct characteristics of interinstitutional research. However, due to the diversity of papers, we need more samples and better methods for extracting features of papers produced by interinstitutional teams.

4.2. RQ 2: Correlations of Research Collaborations with Affiliations. Case 2 learns only interinstitutional collaborations between target institutes, while Case 1 is based on all the collaborations. Also, Case 2 emphasizes scholars who participated in the collaborations, compared to Case 3 that focuses on publications. Case 2 could not outperform Case 1 in predicting papers collaborated by POSTECH and RIST. However, this result might come from RIST's lack of interest in writing papers; the next section provides detailed discussions. Otherwise, Case 2 exhibited the best performance in predicting collaborated patents of the two institutes. Its accuracy nearly caught up with the accuracy of Case All (prediction for every collaboration). Additionally, Case 2 exhibited a reasonable precision (0.77) for predicting the collaborated papers, despite its low recall. Case 2 could extract characteristics of previous collaborations, but the characteristics did not have enough generality due to the lack of samples. These results underpin that focusing on previous collaborations between the institutes is more effective than learning the general characteristics of the research collaboration.

RQ 2 is the assumption that characteristics of research institutes influence collaborations of their members. By comparing Case 1 with Case All, we found out that interinstitutional collaborations between particular institutes have unique characteristics compared to the other collaborations conducted by the institutes. Then, Case 2 revealed that we could find and utilize the unique characteristics. Also, differences in accuracy for papers and patents might be caused by the fact that scholars in RIST barely write papers, and we did not restrict our dataset to research conducted on duty. In other words, research institutes have preferences for certain types and topics of research outcomes, and the preferences affect research of scholars in the institutes. Conclusively, we can say that characteristics of our affiliations affect our research and research collaborations.

4.3. RQ 3: Preferences of Research Institutes for Collaboration Outcomes. Case 3 aims at finding kinds of publications preferred by the research institutes and forming research groups that are capable of producing the same kinds of research outcomes. Case 3 could not outperform Case 2 that concentrates on previous participants in collaborations between the institutes. However, performance gaps between them were not significant, and Case 3 had a higher recall for predicting collaborated papers than Case 2. These results underpin that styles of expected publications are as significant as characteristics of scholars to predict interinstitutional collaborations. Also, the effectiveness of expected publications means that research institutes have preferences for certain kinds of research outcomes, although we do not know what exactly determines the "kinds."

We can see this point also in a comparison of accuracy for collaborated papers with that for collaborated patents. Case 2 and Case 3 outperformed Case 1 in predicting collaborations that produced patents, while they showed contrary results in predicting collaborated papers. Case 1 had a strong point in learning more diverse scholars, publications, and venues, while Case 2 and Case 3 restricted the ranges of training data. Thus, we can assume that papers are more various than patents. Case 1 and Case 2 performed lower recall than their precision in predicting collaborated papers. For Case 2, the collaborated papers from 2011 to 2015 might not be enough to represent collaborations

between POSTECH and RIST. However, different results of Case 1 and Case All are difficult to be explained. We carefully conjecture that there were changes in their collaborations for papers between the two time periods (2011 to 2015 and 2016 to 2020). To understand these results more clearly, we should conduct experiments with more institutes and publication types in further research. Differences in purposes of POSTECH and RIST could worsen the issue. We interviewed staff of the technology licensing office of RIST to find the differences. According to the staff, RIST concentrates on applying its research outcomes for patents and restricts its members' academic papers. Otherwise, POSTECH is a research-oriented university, and it barely intervenes in the dissemination of research outcomes. Thus, the paper dataset was not enough to represent scholars in RIST; only 62 papers were written by members of RIST during the recent ten years, while 2,862 and 2,762 patent applications were published by RIST and POSTECH during the same period.

Conclusively, institutes expected particular styles of publications from their collaborations, and the expectation was effective for the interinstitutional research team formation. Also, there were significant differences between types of publications (e.g., journal articles, patents, books, etc.), and research institutes occasionally had preferences for publication types. However, we should construct a unified bibliographic network that includes more research institutes and various academic publications to validate RQ 3 more clearly.

5. Conclusion

This study aims at the interinstitutional research team formation. The existing methods for composing research teams barely considered characteristics of research institutes, although the institutes have individual research interests and aims. We have proposed methods for extracting features of both research institutes and scholars and methods for composing interinstitutional teams based on both sides' characteristics. First, we extended the conventional bibliographic network to represent research institutes' characteristics and embedded the network. Based on vector representations of scholars and publications, we have proposed three methods for predicting collaboration probabilities between scholars in target institutes. The three methods have different ranges of training data: (i) all the previous collaborations, (ii) collaborations between target institutes, and (iii) publications preferred by the target institutes.

We evaluated the three prediction methods and validated our assumptions by predicting collaborations between POSTECH and RIST from 2016 to 2020 by learning their collaborations from 2011 to 2015. From the experimental results, we found that interinstitutional research collaborations have distinct characteristics compared to other types of collaborations. Also, as we expected, publications of scholars were affected by their affiliations, and this influence obviously had correlations with collaborations of the scholars. Lastly, some institutes had preferences for particular types of publications. These correlations and preferences were helpful for predicting future collaborations.

Despite the reasonable accuracy of the proposed methods, they have also shown several limitations as follows:

(i) Scale of dataset: We conducted experiments for only two institutes, and we could not integrate the bibliographic data for papers and patents due to the author name disambiguation problem. We should construct a unified bibliographic network that includes more research institutes and various academic publications to validate RQ 3 more clearly. Also, our experiment for predicting collaborated papers could not be generalized enough due to the lack of Scopus-indexed papers published by RIST. Thus, we should diversify our data sources, for example, collecting domestic journals and conferences. Considering more research institutes can improve this problem.

(ii) Collaboration prediction methods: To predict interinstitutional collaborations, we simply used the conventional MLP model. Our assumptions were applied to only adjusting ranges of training and testing data. Although this approach performed reasonable accuracy and was enough to validate the assumptions, the accuracy can be improved by employing more sophisticated team formation methods. Also, we will attempt to combine the assumptions with prediction models.

(iii) Content of academic publications: We supposed that publications' venues and authors imply the publications' content. However, this approach could not be as accurate as analyzing the content directly. Also, in the case of patents, their venues are patent offices of each nation. Thus, their venues can be correlated to their impact but not to research domains. This point will be the same for technical reports and preprints. In further research, we will attempt to combine content analysis for academic publications with the proposed team formation methods.

Data Availability

The bibliographic data used to support the findings of this study were supplied by RIST (Research Institute of Industrial Science and Technology) under license and so cannot be made freely available. Requests for access to these data should be made to RIST (http://www.rist.re.kr).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Consilience Creative program (IITP-2019-2011-1-00783) supervised by the IITP (Institute for Information and communications Technology Planning and Evaluation). Also, this study was supported by the RIST (Research Institute of Industrial Science and Technology), Korea, under Grant no. 2020K011.

References

[1] H.-J. Jeon, O.-J. Lee, and J. J. Jung, "Is performance of scholars correlated to their research collaboration patterns?" *Frontiers in Big Data*, vol. 2, no. 39, 2019.

[2] S. Yu, H. D. Bedru, I. Lee, and F. Xia, "Science of scientific team science: a survey," *Computer Science Review*, vol. 31, pp. 72–83, 2019.

[3] W. Wang, J. Ren, M. Alrashoud, F. Xia, M. Mao, and A. Tolba, "Early-stage reciprocity in sustainable scientific collaboration," *Journal of Informetrics*, vol. 14, no. 3, Article ID 101041, 2020.

[4] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[5] O.-J. Lee, H.-J. Jeon, and J. J. Jung, "Learning multi-resolution representations of research patterns in bibliographic networks," *Journal of Informetrics*, vol. 15, no. 1, Article ID 101126, 2021.

[6] H.-J. Jeon and J. J. Jung, "Discovering the role model of authors by research history embedding," *Journal of Information Science*, Under Review.

[7] X. Kong, H. Jiang, W. Wang, T. M. Bekele, Z. Xu, and M. Wang, "Exploring dynamic research interest and academic influence for scientific collaborator recommendation," *Scientometrics*, vol. 113, no. 1, pp. 369–385, 2017.

[8] J. Zhou and M. A. Rafi, "Recommendation of research collaborator based on semantic link network," in *Proceedings of the 15th International Conference on Semantics, Knowledge and Grids (SKG 2019)*, pp. 16–20, IEEE, Guangzhou, China, September 2019.

[9] X. Kong, H. Jiang, Z. Yang, Z. Xu, F. Xia, and A. Tolba, "Exploiting publication contents and collaboration networks for collaborator recommendation," *PLOS ONE*, vol. 11, no. 2, Article ID e0148492, 2016.

[10] N. Sun, Y. Lu, and Y. Cao, "Career age-aware scientific collaborator recommendation in scholarly big data," *IEEE Access*, vol. 7, pp. 136036–136045, 2019.

[11] C. Xiao, J. Han, W. Fan, S. Wang, R. Huang, and Y. Zhang, "Predicting scientific impact via heterogeneous academic network embedding," in *Trends in Artificial Intelligence - Proceedings of the 16th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2019)*, A. C. Nayak and A. Sharma, Eds., Springer International Publishing, Cuvu, Yanuca Island, Fiji, pp. 555–568, 2019.

[12] Z. Liu, X. Xie, and L. Chen, "Context-aware academic collaborator recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2018)*, Y. Guo and F. Farooq, Eds., ACM, London, UK, pp. 1870–1879, 2018.

[13] A. Narayanan, M. Chandramohan, L. Chen, Y. Liu, and S. Saminathan, "subgraph2vec: Learning distributed representations of rooted sub-graphs from large graphs," http://arxiv.org/abs/1606.08928.

[14] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs, Computing Research Repository (CoRR)," http://arxiv.org/abs/1707.05005.

[15] O.-J. Lee and J. J. Jung, "Story embedding: learning distributed representations of stories based on character networks," *Artificial Intelligence*, vol. 281, Article ID 103235, 2020.

[16] O.-J. Lee, J. J. Jung, and J.-T. Kim, "Learning hierarchical representations of stories by using multi-layered structures in narrative multimedia," *Sensors*, vol. 20, no. 7, p. 1978, 2020.

[17] D. Purwitasari, C. Fatichah, I. K. E. Purnama, S. Sumpeno, and M. H. Purnomo, "Inter-departmental research collaboration recommender system based on content filtering in a cold start problem," in *Proceedings of the 10th IEEE International Workshop on Computational Intelligence and Applications (IWCIA 2017)*, pp. 177–184, IEEE, Hiroshima, Japan, November 2017.

[18] N. Hernandez-Gress, H. G. Ceballos, and N. Galeano, "Research collaboration recommendation for universities based on data science," in *Proceedings of the 6th International Conference on Computational Science and Computational Intelligence (CSCI 2018)*, pp. 1129–1132, IEEE, Las Vegas, NV, USA, December 2018.

[19] J. D. T. Guerrero-Sosa, V. H. Menéndez-Domínguez, M.-E. Castellanos-Bolaños, and L. F. Curi-Quintal, "Analysis of internal and external academic collaboration in an institution through graph theory," *Vietnam Journal of Computer Science*, vol. 07, no. 4, pp. 391–415, 2020.

[20] W. Wang, J. Liu, Z. Yang, X. Kong, and F. Xia, "Sustainable collaborator recommendation based on conference closure," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 311–322, 2019.

[21] W. Wang, B. Xu, J. Liu et al., "CSTeller: forecasting scientific collaboration sustainability based on extreme gradient boosting," *World Wide Web*, vol. 22, no. 6, pp. 2749–2770, 2019.

[22] Y. Zhang, C. Zhang, and X. Liu, "Dynamic scholarly collaborator recommendation via competitive multi-agent reinforcement learning," in *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017)*, P. Cremonesi, F. Ricci, S. Berkovsky, and A. Tuzhilin, Eds., ACM, Como, Italy, pp. 331–335, 2017.

[23] B. Xu, L. Li, J. Liu, L. Wan, X. Kong, and F. Xia, "Disappearing link prediction in scientific collaboration networks," *IEEE Access*, vol. 6, pp. 69702–69712, 2018.

[24] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)*, pp. 739–744, IEEE Computer Society, Omaha, NE, USA, October 2007.

[25] X. Zhou, L. Ding, Z. Li, and R. Wan, "Collaborator recommendation in heterogeneous bibliographic networks using random walks," *Information Retrieval Journal*, vol. 20, no. 4, pp. 317–337, 2017.

[26] G. J. S. Ganguly, M. Gupta, V. Varma, and V. Pudi, "Author2vec: learning author representations by combining content and link information," in *Proceedings of the 25th International Conference on World Wide Web (WWW 2016)*, J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, and B. Y. Zhao, Eds., ACM, Montreal, Canada, pp. 49-50, 2016.

[27] W. Wang, Z. Cui, T. Gao, S. Yu, X. Kong, and F. Xia, "Is scientific collaboration sustainability predictable?," in *Proceedings of the 26th International Conference on World Wide Web Companion (WWW 2017)*, R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, Eds., ACM, Perth, Australia, pp. 853-854, 2017.

[28] M. V. Dodson, L. L. Guan, M. E. Fernyhough et al., "Perspectives on the formation of an interdisciplinary research team," *Biochemical and Biophysical Research Communications*, vol. 391, no. 2, pp. 1155–1157, 2010.

[29] E. Omodei, M. De Domenico, and A. Arenas, "Evaluating the impact of interdisciplinary research: a multilayer network approach," *Network Science*, vol. 5, no. 2, pp. 235–246, 2016.

[30] W. Wang, S. Yu, T. M. Bekele, X. Kong, and F. Xia, "Scientific collaboration patterns vary with scholars' academic ages," *Scientometrics*, vol. 112, no. 1, pp. 329–343, 2017.

[31] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: large-scale information network embedding," in *Proceedings of the 24th International Conference on World Wide Web (WWW 2015)*, A. Gangemi, S. Leonardi, and A. Panconesi, Eds., ACM, Florence, Italy, pp. 1067–1077, 2015.

[32] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, pp. 2539–2561, 2011.

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26: Proceedings of 27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 3111–3119, Curran Associates, Inc., Lake Tahoe, NV, US, 2013.