Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status

Paweena Chaowanarom, Chidchanok Lursinsap

Abstract
Currently, the research in computer science has been exponentially expanded beyond its own fields into the other research fields such as medical science, business, and social science in forms of collaborative researches. This collaborative researches stimulate a new recommending algorithm for determining a potential research collaborator under the interdisciplinary environment. Unlike other research fields, the research problems in computer science can be transformed to other known and solvable problems. In this paper, a new hybrid algorithm based on dynamic collaboration over time was proposed for recommending an appropriate collaborator. Besides considering only three basic factors concerning social proximity, friendship, and complementarity skill as employed by others’, three new additional factors related to research interest, up-to-date publication data, and seniority of researcher are involved in our analysis. A set of new measures for all six recommending factors were proposed. The experiments were conducted with real bibliographic data within six continuous years of publication and over six topics in computer science. Our results were significantly higher than the results of the other methods at 90% confidence level.

Keywords
Social network; Research collaboration; Collaborator recommendation; Knowledge discovery; Co-authorship network

1. Introduction
Currently, the research in computer science has been exponentially expanded beyond its own fields into the other research fields such as medical science, business, and social science in forms of collaborative researches. The algorithms or techniques in computer science are rather versatile and capable of finding solutions to various difficult problems in scientific, engineering, medical science, or even social science. Furthermore, the concerned research problems in computer science can be transformed to another known and solvable problem. The solution to this known problem can be adapted to solve the concerned research problems within the same time and space complexities. The collaboration level in computer science papers is rather moderate with respect to other scientific fields [1]. One question usually encountered by a team leader is how to select the appropriate collaborators in a new research. Finding a potential research collaborator in computer science is a challenging problem for the following reasons. Computer science research in the aspects of research problems and algorithms is rapidly and temporarily changed during the past decade [2]. The publications in the field are very heterogeneous and mostly interdisciplinary to many other topics [2] such as software engineering and data mining reported by Bird [3]. This interdisciplinary trend has stimulated a higher degree of research collaboration among the researchers in various fields of computer science. Computer science comprises various sub-fields requiring specialization and characteristic features [2] and [4].
However, it is noticeable that there are some of researchers published their papers in several related fields outside their main research fields. This is due to the nature of interdisciplinary nature of computer science which can be transformed to another problem. Intuitively, a set of famous researchers in a specific research topic seems to be the best choice to collaborate in research. Unfortunately, most of these potential collaborators are typically overloaded with their own research activities. Thus, the problem of finding the best collaborator should be transformed to the problem of finding the appropriate collaborator instead.

The following factors [5] and [6] have been suggested and used by many researchers for analyzing and recommending potential collaborators to an inquiring researcher.

1. **Social proximity** covering cohesive publication force and possibility of publishing new paper. This factor was considered in many researches by analyzing the structure of a co-authorship network [7], [8], [9], [10], [11], [12] and [13].

2. **Friendship** covering co-authorship force based on the list of their common co-authors. The degree of friendship is measured in terms of distance or number of hops between researchers as in Lee's work [11] for more than one hop distance, Cohen's work [12], CollabSee system [13], and our previous work [10] for three and four hops, respectively.

3. **Complementarity skills** covering the research background similarity [10], [11], [12], [13], [14], [15], [16] and [17]. Researchers with more background research experiences in a required topic are more likely to be selected.

4. **Research interest** governed by the probability to work in inquired research topic in the near future. The factor concerning the change or trend research interest in the future has not been studied before. This factor actually occurs among computer scientists due the nature of computer science research.

5. **Up-to-date information** regarding the publication [8], [10] and [18]. This factor refers to the year of publication in social proximity measure.

6. **Seniority status** of researcher [5], [6] and [19]. Senior researchers have more potential to be recommended than junior researchers.

The existing techniques in several research studies did not cover the essence of the six factors. The main disadvantage is the fresh information for supporting dynamic collaboration over time has been lost for measuring. Lack of time stamp in social proximity and research trend lead to mistake in selecting potential research collaborators. An example of unsatisfied output of researcher recommendation is the result obtained from CollabSee system. This system used only the name of seeker for querying, but an inquired research topic was omitted. It was based on the assumption that the research topics in the past and the future are the same. This assumption is too rigid and it may create a pitfall if the inquired researcher wants to change his topic or find his collaborators across different topics based on interdisciplinary environment.

In this paper, we proposed a new algorithm to increase the relevance and accuracy of recommending a collaborator in computer science field. Any researcher who wants to find some research collaborators is called **inquiring researcher**. Not only a given inquiring researcher’s name, but also an inquired topic are used for querying. The output is a ranked list of potential collaborators relating to both researcher’s name and the inquired topic. The overall contributions mainly focusing on the above six factors are the following:

- Our proposed algorithm is a hybrid method consisting of a structural approach based on co-authorship network with social proximity and friendship factors. Moreover, it also uses semantic approach based on the content of papers covering the factor of complementarity skill and the temporal research interest.
- Unlike the previous other studies, time evolution is attended in our algorithm for taking into account with the up-to-date information factor. The year of publication is gradually integrated in both structural approach and semantic approach.
Our algorithm pays attention to seniority factor by studying the seniority relationship among researchers in computer science sub-fields. The possible seniority relationship between a co-author pair are: (1) senior researcher collaborating with senior researcher; (2) senior researcher collaborating with junior researcher; and (3) junior researcher collaborating with junior researcher. The output is the proportions of each seniority relationship which can be used as the parameter in algorithm.

However, there are some sub-fields in computer science having a few researchers when compared with other sub-fields, i.e., pure theoretical computer science. This makes the analysis rather difficult and inaccurate. Therefore, only the topics with a large number of researchers involved are concentrated, i.e., Bio-informatics, Data Mining, Hardware, Neural Network, Software, and Algorithm and Theory. Our study focused on the problem of defining a measure for each determination factor. The following questions are the main concerns.

1. How to measure the cohesive publication force between two researchers who may be friends or who may only know each other via the publications?
2. How to measure the possibility of publishing new papers of the potential collaborator?
3. How to measure the cohesive co-authorship force between two researchers?
4. How to measure the similarity of research backgrounds between two researchers?
5. How to measure the probability of working in the same domain of the inquiring researcher and the potential collaborator?
6. How to measure the research trend of the potential collaborator?
7. How to measure the potential collaboration between two researchers?
8. How to determine the best collaborator based on these seven measures?

The rest of paper is organized into the following sections. Section 2 describes the proposed methodology to compute the relevance between researcher pairs. Section 3 describes the experiments and performance evaluation. Section 4 concludes the paper.

2. Proposed scoring measures and algorithms

The proximity and friendship are key factors of structure approach in co-authorship network. The complementarity skill and research interest were taken into account in semantic approach. Some relevant definitions are defined as follows.

**Definition 1.**

Co-authorship network $G = (V, E, L)$ is an undirected multigraph consisting of a set of all researchers' names, $V = \{v_1, \ldots, v_n\}$, appearing in the considered publication database, their relationships in terms of edges, $E = \{(v_a, v_b) | v_a, v_b \in V\}$, and a set of attributes, $L = \{(v_a, v_b) | (v_a, v_b) \in E\}$, attached to each edge. Each $(v_a, v_b)$ consists of (1) paper ID, (2) year of publication, and (3) number of authors in each paper.

**Definition 2.**

The degree of separation between $v_a$ and $v_b$, denoted as $d_{v_a,v_b}$, is the minimum number of edges forming a path from $v_a$ to $v_b$.

**Definition 3.**

A neighbor of any $v_i$ is a vertex $v_a$ such that $(v_i,v_a) \in E$ and $1 < d_{v_i,v_a} < 6$.

**Definition 4.**

Researcher $v_a$ is a **friend** researcher of $v_i$ if $v_a$ is a neighbor of $v_i$ and they co-authored some papers. Let $Q_{v_i}$ denote a set of friends of $v_i$.

**Definition 5.**

Researcher $v_a$ is a **non-friend** researcher of $v_i$ if $v_a$ is a neighbor of $v_i$ but they never co-authored any papers. Let $Q'_{v_i}$ denote a set of non-friend researchers of $v_i$. 

---

**Recommended articles**

Modeling knowledge need awareness using the ...

2015, Knowledge-Based Systems  more
Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status

Paweena Chaiwanarom *, Chidchanok Lursinsap

Advanced Virtual and Intelligent Computing Research Center, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, 254 Phayathai Road, Pathumwan, Bangkok 10330, Thailand

ARTICLE INFO

Article history:
Received 4 June 2014
Received in revised form 17 October 2014
Accepted 27 November 2014
Available online 4 December 2014

Keywords:
Social network
Research collaboration
Collaborator recommendation
Knowledge discovery
Co-authorship network

ABSTRACT

Currently, the research in computer science has been exponentially expanded beyond its own fields into the other research fields such as medical science, business, and social science in forms of collaborative researches. This collaborative researches stimulate a new recommending algorithm for determining a potential research collaborator under the interdisciplinary environment. Unlike other research fields, the research problems in computer science can be transformed to other known and solvable problems. In this paper, a new hybrid algorithm based on dynamic collaboration over time was proposed for recommending an appropriate collaborator. Besides considering only three basic factors concerning social proximity, friendship, and complementarity skill as employed by others*, three new additional factors related to research interest, up-to-date publication data, and seniority of researcher are involved in our analysis. A set of new measures for all six recommending factors were proposed. The experiments were conducted with real bibliographic data within six continuous years of publication and over six topics in computer science. Our results were significantly higher than the results of the other methods at 90% confidence level.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Currently, the research in computer science has been exponentially expanded beyond its own fields into the other research fields such as medical science, business, and social science in forms of collaborative researches. The algorithms or techniques in computer science are rather versatile and capable of finding solutions to various difficult problems in scientific, engineering, medical science, or even social science. Furthermore, the concerned research problems in computer science can be transformed to another known and solvable problem. The solution to this known problem can be adapted to solve the concerned research problems within the same time and space complexities. The collaboration level in computer science papers is rather moderate with respect to other scientific fields [1]. One question usually encountered by a team leader is how to select the appropriate collaborators in a new research. Finding a potential research collaborator in computer science is a challenging problem for the following reasons. Computer science research in the aspects of research problems and algorithms is rapidly and temporally changed during the past decade [2]. The publications in the field are very heterogeneous and mostly interdisciplinary to many other topics [2] such as software engineering and data mining reported by Bird [3]. This interdisciplinary trend has stimulated a higher degree of research collaboration among the researchers in various fields of computer science. Computer science comprises various sub-fields requiring specialization and characteristic features [2,4]. However, it is noticeable that there are some of researchers published their papers in several related fields outside their main research fields. This is due to the nature of studied problems in computer science which can be transformable to another problem. Intuitively, a set of famous researchers in a specific research topic seems to be the best choice to collaborate in research. Unfortunately, most of these potential collaborators are typically overloaded with their own research activities. Thus, the problem of finding the best collaborator should be transformed to the problem of finding the appropriate collaborator instead.

The following factors [5,6] have been suggested and used by many researchers for analyzing and recommending potential collaborators to an inquiring researcher.
1. Social proximity covering cohesive publication force and possibility of publishing new paper. This factor was considered in many researches by analyzing the structure of a co-authorship network [7–13].

2. Friendship covering cohesive co-authorship force based on the list of their common co-authors. The degree of friendship is measured in terms of distance or number of hops between researchers as in Lee’s work [11] for more than one hop distance, Cohen’s work [12], CollabSeer system [13], and our previous work [10] for three and four hops, respectively.

3. Complementarity skills covering the research background similarity [10–17]. Researchers with more background research experiences in a required topic are more likely to be selected.

4. Research interest governed by the probability to work in inquired research topic in the near future. The factor concerning the change or trend research interest in the future has not been studied before. This factor actually occurs among computer scientists due the nature of computer science research.

5. Up-to-date information regarding the publication [6,10,18]. This factor refers to the year of publication in social proximity measure.

6. Seniority status of researcher [5,6,19]. Senior researchers have more potential to be recommended than junior researchers.

The existing techniques in several research studies did not cover the essence of the six factors. The main disadvantage is the fresh information for supporting dynamic collaboration over time has been lost for measuring. Lack of time stamp in social proximity and research trend lead to mistake in selecting potential research collaborators. An example of unsatisfied output of researcher recommendation is the result obtained from CollabSeer system. This system used only the name of seeker for querying, but an inquired research topic was omitted. It was based on the assumption that the research topics in the past and the future are the same. This assumption is too rigid and it may create a pitfall if the inquired researcher wants to change his topic or find his collaborators across different topics based on interdisciplinary environment.

In this paper, we proposed a new algorithm to increase the relevance and accuracy of recommending a collaborator in computer science field. Any researcher who wants to find some research collaborators is called inquiring researcher. Not only a given inquiring researcher’s name, but also an inquired topic are used for querying. The output is a ranked list of potential collaborators relating to both researcher’s name and the inquired topic. The overall contributions mainly focusing on the above six factors are the following:

- Our proposed algorithm is a hybrid method consisting of a structural approach based on co-authorship network with social proximity and friendship factors. Moreover, it also uses semantic approach based on the content of papers covering the factor of complementarity skill and the temporal research interest.

- Unlike the previous other studies, time evolution is attended in our algorithm for taking into account with the up-to-date information factor. The year of publication is gradually integrated in both structural approach and semantic approach.

- Our algorithm pays attention to seniority factor by studying the seniority relationship among researchers in computer science sub-fields. The possible seniority relationship between a co-author pair are: (1) senior researcher collaborating with senior researcher; (2) senior researcher collaborating with junior researcher; and (3) junior researcher collaborating with junior researcher. The output is the proportions of each seniority relationship which can be used as the parameter in algorithm.

However, there are some sub-fields in computer science having a few researchers when compared with other sub-fields, i.e. pure theoretical computer science. This makes the analysis rather difficult and inaccurate. Therefore, only the topics with a large number of researchers involved are concentrated, i.e. Bio-informatics, Data Mining, Hardware, Neural Network, Software, and Algorithm and Theory. Our study focused on the problem of defining a measure for each determination factor. The following questions are the main concerns.

1. How to measure the cohesive publication force between two researchers who may be friends or who may only know each other via the publications?
2. How to measure the possibility of publishing new papers of the potential collaborator?
3. How to measure the cohesive co-authorship force between two researchers?
4. How to measure the similarity of research backgrounds between two researchers?
5. How to measure the probability of working in the same domain of the inquiring researcher and the potential collaborator?
6. How to measure the research trend of the potential collaborator?
7. How to measure the potential collaboration between two researchers?
8. How to determine the best collaborator based on these seven measures?

The rest of paper is organized into the following sections. Section 2 describes the proposed methodology to compute the relevance between researcher pairs. Section 3 describes the experiments and performance evaluation. Section 4 concludes the paper.

2. Proposed scoring measures and algorithms

The proximity and friendship are key factors of structure approach in co-authorship network. The complementarity skill and research interest were taken into account in semantic approach. Some relevant definitions are defined as follows.

Definition 1. Co-authorship network \( G = (V, E, L) \) is an undirected multigraph consisting of a set of all researchers’ names, \( V = \{v_1, v_2, \ldots, v_n\} \), appearing in the considered publication database, their relationships in terms of edges, \( E = \{(v_a, v_b) \mid v_a, v_b \in V\} \), and a set of attributes, \( L = \{l_{(v_a, v_b)}(v_a, v_b) \in E\} \), attached to each edge. Each \( l_{(v_a, v_b)} \) consists of (1) paper ID, (2) year of publication, and (3) number of authors in each paper.

Definition 2. The degree of separation between \( v_a \) and \( v_b \), denoted as \( d_{v_a, v_b} \), is the minimum number of edges forming a path from \( v_a \) to \( v_b \).

Definition 3. A neighbor of any \( v_a \) is a vertex \( v_b \) such that \((v_a, v_b) \in E \) and \( 1 \leq d_{v_a, v_b} \leq 6 \).

Definition 4. Researcher \( v_a \) is a friend researcher of \( v_b \) if \( v_a \) is a neighbor of \( v_b \) and they co-authored some papers. Let \( Q_{v_a} \) denote a set of friends of \( v_a \).

Definition 5. Researcher \( v_a \) is a non-friend researcher of \( v_b \) if \( v_a \) is a neighbor of \( v_b \) but they never co-authored any papers. Let \( Q_{v_b}^\prime \) denote a set of non-friend researchers of \( v_b \).
Fig. 1 shows an example of co-authorship network constructed from some data in our experimental data set. Each vertex is denoted by a rectangle with author's name inside the rectangle. The attributes are from the table above the graph. The network consists of eight researchers and eight co-authored papers which are summarized in the table of Fig. 1. The attributes on each edge are (1) paper ID, denoted as p, (2) year of publication of p, and (3) number of authors in p. For example, the attributes at edge connecting authors Versaci and Calcogno are (p2, 2006, 4). Both of them co-authored paper p2 in the year 2006 and there are four authors in paper p2. Let vi be the vertex in G representing the inquiring researcher. The detail of proposed algorithm is the following.

Algorithm 1. Selecting and Ranking Potential Collaborators

| Inputs: | The inquiring researcher vi and a research topic. |
| Output: | Selected collaborators and ranking of relevant scores. |
| 1. | Construct a co-authorship network G from a collection of papers in a given database. |
| 2. | Extract all neighbors Ni = { vj | 1 ≤ d(vj, vi) ≤ 6} |
| 3. | For each vj ∈ Ni do |
| 4. | If vj is a friend researcher of vi then |
| 5. | Compute cohesive publication force S(vj, vi) between vi and vj based on their co-authored papers. |
| 6. | If vj is a non-friend researcher of vi then |
| 7. | Compute cohesive publication force S(vj, vi) between vi and vj based on their transitive relation. |
| 8. | Compute possibility of publishing new paper score T(vj, vi) based on the list of their common co-authors. |
| 9. | Compute existing publication evidence score E(vj, vi) between vi and vj using S(vj, vi), T(vj, vi), P(vj), and F(vj, vi). |
| 10. | EndFor |
| 11. | For each vj ∈ Ni do |
| 12. | Compute research background similarity score B(vj, vi) between vi and vj. |
| 13. | Compute probability ϕ(vj, vi) of vj to work in vi’s inquired research topic. |
| 14. | Compute research trend score T(vj, vi) between vi and vj using B(vj, vi) and ϕ(vj, vi). |
| 15. | EndFor |
| 16. | Compute potential collaboration score C(vj, vi) between vi and vj using E(vj, vi), T(vj, vi), d(vj, vi), and seniority status. |
| 17. | EndFor |
| 18. | Partition ∀ vj ∈ Ni into 6 groups according to d(vj, vi). |
| 19. | For each group in step 18, rank ∀ vj according to C(vj, vi). |

The details of how to compute each force and score in steps 4, 5, 6, 7, 8, 11, 12, 13, and 16 are given in the following sections.

2.1. Computing cohesive publication force S(vj, vi)

The cohesive publication force S(vj, vi) between vi and friend researcher vj is based on the following observations. The first observation concerns the number of co-authors appearing in any paper written by vi and vj. The more number of authors appears in the paper, the less cohesive publication force binds vi and vj. This implies that the contribution of each vj to the paper is not outstanding enough to attract the attention of vi to join any work in the future. The second observation is the freshness of a published paper with respect to the current year. If vj and vi used to publish a paper long time ago, then the chance that they may co-author any paper again is low. Let

1. uy be the number of authors appearing in the jth paper co-authored by vi and vj.
2. yj be the number of years counted between the current year and the year when paper j was published.
3. P be a set of all papers co-authored by vi and vj.

Based on the above observations, the cohesive publication force S(vj, vi) is computed as follows:

\[
S(vj, vi) = \sum_{j \in P} \left( \frac{1}{uy - 1} \right) \left( \frac{1}{yj} \right)
\]

Table 1 (a) illustrates the cohesive publication forces between vi and friend researcher S(vj, vi) in the co-authorship graph in Fig. 1. The scores and forces were based on the current inquiring year of 2008.

2.2. Computing cohesive publication force \( \bar{S}(vj, vi) \)

Computing the cohesive publication force between vi and a non-friend researcher vj, \( \bar{S}(vj, vi) \), is based on the assumption that there must be a path connecting vi and vj whose \( d(vj, vi) \geq 2 \). Prior to the calculation, all connecting paths between vi and vj must be identified first. All vertices in each path represent non-friend researchers laying in between vi and vj. Let \( L = \{ l1, \ldots, lq \} \) be a set of all possible paths connecting vi and vj. The cohesive publication force \( \bar{S}(vj, vi) \) is viewed as the force propagated and transmitted via all paths in L from vi to vj. We denote \(|l|\) as the length of path \( l \) measured in terms of edges. Therefore, \( \bar{S}(vj, vi) \) is computed as follows:

\[
\bar{S}(vj, vi) = \sum_{l \in L} \frac{|l|}{\sum_{(l, n, m) \in (l, n, m)} \left( \frac{1}{d(n, vj) - 2} \frac{1}{d(m, vi)} \right)}
\]

Suppose vi is Versaci and his non-friend research is Mammmone. From the graph in Fig. 1, there are more than one path between Versaci and Mammmone. To illustrate how to compute this force, we consider only two paths and the force of each path is computed as follows:

1. The first path: Versaci to Morabito to Mammmone. This path is of length 2. The force of this path is

\[
\frac{2}{0.093 + 0.25} = 0.098
\]

2. The second path: Versaci to Costantino to Morabito to Mammmone. This path is of length 3. The force of this path is

\[
\frac{3}{0.0208 + 0.093 + 0.25} = 0.0094
\]

The cohesive publication force between Versaci and Mammmone computed from these two paths is equal to 0.098 + 0.0094 = 0.1074.
2.3. Computing possibility of publishing new paper score $P_{ns}$

We observed that there are two factors stimulating a researcher $v_s$ to publish a new paper. The first factor concerns how many friends $v_s$ has been collaborating with during the past period. If $v_s$ has many collaborating friends, then $v_s$ will have a higher possibility to collaborate again with those friends. The second factor concerns the freshness of each published paper measured in terms of the duration between the current year and the time when the paper was published. If the duration is long, then it implies that $v_s$ may be less diligent to publish a new paper in the future. Let $y_j$ be the number of years counted in between the current year and the year when paper $j$ was published by $v_s$. Let $J$ be a set of all papers published by $v_s$ within a considered period. The period length may be defined by the inquiring researcher $v_t$. The possibility of publishing new paper score is computed as Eq. (5) and Table 1(b) shows the details of computing for eight researchers in Fig. 1.

$$P_{ns} = \sum_{j \in J} \frac{1}{2^{y_j - 1}}$$  (5)

2.4. Computing cohesive co-authorship force $F_{n,n_s}$

The measure of cohesive co-authorship force is based on the observation that any $v_s$ and $v_t$ sharing many common friends should have higher cohesive co-authorship force than those having few common friends. This observation was adapted from the concept introduced in the work by Adamic and Adar [20]. For any researcher $v_t$, let $Q_{nt}$ be a set of friends of $v_t$ and $|Q_{nt}|$ be the cardinality of $Q_{nt}$. The cohesive co-authorship force is measured with respect to the total number of vertices $|V|$ in the co-authorship network $G$ and is defined as follows.

$$F_{n,n_s} = \sum_{v_t \in Q_{nt} \cap Q_{ns}} \log \left( \frac{|V|}{|Q_{nt}|} \right)$$  (6)

In Fig. 1, suppose researchers Versaci and Marra are considered. Both of them have Corsonello, and Morabito are as their common friends. Corsonello has 3 friends and Morabito has 7 friends. The total number of vertices in the graph is equal to 8. Thus, $F_{Versaci,Marra} = \log(8/3) + \log(8/7) = 0.484$.

2.5. Computing existing publication evidence score $E_{n,n_s}$

The existing publication score measures how close the inquiring research $v_t$ to potential collaborator $v_s$ is in terms of relevant $S_{n,n_s}$, $\tilde{S}_{n,n_s}$, $P_{ns}$, and $F_{n,n_s}$. The value of each factor must be normalized to avoid the quantity bias effect. Therefore, the existing publication evidence score is computed as follows.

$$M = \sum_{v_t \in Q_{nt} \cap Q_{ns}} S_{n,n_s} + \tilde{S}_{n,n_s}$$  (7)

$$E_{n,n_s} = \frac{S_{n,n_s} + \tilde{S}_{n,n_s} + P_{ns} + F_{n,n_s}}{M}$$  (8)
Table 1
An example how to compute the cohesive publication forces $S_{n,n}$ among researchers and possibility of publishing new paper score $P_n$ of each researcher based on the current inquiring year of 2008. (a) The cohesive publication forces. (b) The possibility of publishing new paper score.

| $v_1$ | $v_2$ | $j$ | $u_1$ | $U$ | $2^{-|u_1|}$ | $(U - 2^{-|u_1|})^2$ | $v_1$ | $v_2$ | $S_{n,n}$ |
|-------|-------|-----|-------|-----|-------------|----------------|-------|-------|-----------|
| Morabito | Versaci | $p_1$ | 2 | 1.0 | 0.031 | 0.031 | Calcagno | Greco | 0.165 |
| Morabito | Versaci | $p_2$ | 2 | 1.0 | 0.031 | 0.031 | Calcagno | Morabito | 0.165 |
| Greco | Costantino | $p_3$ | 4 | 0.33 | 0.063 | 0.0208 | Greco | Versaci | 0.165 |
| Greco | Morabito | $p_4$ | 4 | 0.33 | 0.063 | 0.0208 | Greco | Morabito | 0.165 |
| Greco | Versaci | $p_5$ | 3 | 0.5 | 0.063 | 0.0315 | Greco | Versaci | 0.165 |
| Marra | Morabito | $p_6$ | 5 | 0.5 | 0.063 | 0.0315 | Greco | Versaci | 0.165 |
| Marra | Versaci | $p_7$ | 5 | 0.5 | 0.063 | 0.0315 | Greco | Versaci | 0.165 |
| Marra | Morabito | $p_8$ | 3 | 0.5 | 0.063 | 0.0315 | Greco | Versaci | 0.165 |
| Marra | Morabito | $p_9$ | 4 | 0.33 | 0.125 | 0.04125 | Marra | Morabito | 0.165 |
| Marra | Morabito | $p_{10}$ | 4 | 0.33 | 0.125 | 0.04125 | Marra | Morabito | 0.165 |
| Marra | Corsonello | $p_{11}$ | 4 | 0.33 | 0.125 | 0.04125 | Marra | Morabito | 0.165 |
| Morabito | Versaci | $p_{12}$ | 4 | 0.33 | 0.125 | 0.04125 | Marra | Morabito | 0.165 |
| Morabito | Corsonello | $p_{13}$ | 4 | 0.33 | 0.125 | 0.04125 | Marra | Morabito | 0.165 |
| Morabito | Versaci | $p_{14}$ | 5 | 0.5 | 0.25 | 0.125 | Morabito | Corsonello | 0.165 |
| Morabito | Versaci | $p_{15}$ | 5 | 0.5 | 0.25 | 0.125 | Morabito | Corsonello | 0.165 |
| Calcagno | Greco | $p_{16}$ | 4 | 0.33 | 0.125 | 0.04125 | Calcagno | Greco | 0.165 |
| Calcagno | Morabito | $p_{17}$ | 4 | 0.33 | 0.125 | 0.04125 | Calcagno | Greco | 0.165 |
| Greco | Morabito | $p_{18}$ | 4 | 0.33 | 0.125 | 0.04125 | Greco | Morabito | 0.165 |
| Greco | Versaci | $p_{19}$ | 4 | 0.33 | 0.125 | 0.04125 | Greco | Morabito | 0.165 |
| Mammmone | Morabito | $p_{20}$ | 2 | 1.0 | 0.25 | 0.25 | Mammmone | Morabito | 0.25 |

(b) The possibility of publishing new paper score $P_n$ of each researcher

| Year | $2^{-|u_1|}$ | $P_n$ |
|------|-------------|-------|
| $p_1$ | 2002 | 0.031 |
| $p_2$ | 2002 | 0.031 |
| $p_3$ | 2003 | 0.063 |
| $p_4$ | 2003 | 0.063 |
| $p_5$ | 2004 | 0.125 |
| $p_6$ | 2005 | 0.25 |
| $p_7$ | 2006 | 0.25 |
| $p_8$ | 2006 | 0.25 |

2.6. Computing research background similarity score $B_{n,n}$

This measure emphasized on the personal information of $v_i$'s and $v_j$'s. The similarity of research backgrounds of the inquiring research $v_i$ and any potential collaborator $v_j$ can influence the correctness of matching result of $v_i$ and $v_j$. If both backgrounds are the same or exactly the same, then both of them may have a higher chance to collaborate. A set of finite research topics must be defined first. Suppose there are $m$ topics. A researcher published different number of papers in different topic within a specified period of years.

The background of $v_j$ is captured in terms of an $m$-tuple of probabilities of different research topics. This background can be viewed as a vector in an $m$-dimensional space. Let $b_{n,n} = [b_{n,1}, ..., b_{n,m}]$ be the background of $v_j$ and $b_{n,m}$ be the probability to publish some papers in topic $j$. The value of $b_{n,m}$ can be estimated by applying the method of Author-Topic Model [21]. The research background similarity score $B_{n,n}$ is computed by adopting the cosine similarity measure as follows.

$$ B_{n,n} = \frac{b_{n,n} \cdot b_{n,m}}{|b_{n,n}| \cdot |b_{n,m}|} \quad (9) $$

2.7. Computing probability $\phi_{n,n}$ of $v_j$ to work in $v_i$'s inquired research topic

A research topic is defined as a set of research topics measured in terms of probabilities similar to the measure previously discussed in $B_{n,n}$. Any neighbor $v_j$ of $v_i$ with the highest probability trend of working in $v_i$'s inquired research topic during a specified period is the most appropriate collaborator. We assume that the considered period is of $T$ years. At the inquiring year $t$, the probability of working in a topic is computed by using the probabilities of the same topic from the preceding years. For researcher $v_j$, let $\rho_{n,k}$ be the probability of working in the inquired research topic domain as that of $v_i$ in year $k$. The value of $\phi_{n,n}$ is computed by the following steps:

1. Let $t$ be the inquiring year.
2. Let $v$ be the number of preceding years in moving average period.
3. For years $t-T+\gamma-1 \leq k \leq t-1$ do
   4. Use $\rho_{n,k}$ with exponential moving average of length $\gamma$ to compute new $\rho_{n,k}$.
   5. EndFor
5. Create a linear function $f(k)$ to interpolate all $(\rho_{n,k})^{t-T+\gamma-1 \leq k \leq t-1}$.
6. Let $\phi_{n,n} = f(t)$
7. EndFor

To demonstrate how to compute the probability $\phi_{n,n}$ of working in $v_i$'s inquired research topic, the data of researcher named Shepperd were used as shown in Table 2. We used his data because there are few records. Note that Shepperd did not appear in the co-authorship network previously mentioned. In this example, there are six
topics. Suppose topic 5 is the inquired research topic of the inquiring researcher \( v_t \) and the current inquiring year is 2008. Each floating point number is \( b_{\text{step,}t} \). To compute \( \phi_{\text{step,}t} \), the values of \( b_{\text{step,}t} \) in each year from years 2002 to 2007 were plotted as shown by the thick line in Fig. 2. Then, the technique of exponential moving average of 3 years was applied to re-compute the probability in each year as shown by the connected dashed line. From this line, a linear function of \( f(t) = 0.0144 t + 0.4152 \) was derived, where \( t \) is the order of years starting in year 2002. The value of \( \phi_{\text{step,}t} \) at year 2008 (\( t = 7 \)) is \( f(7) = 0.0144 \times 7 + 0.4152 = 0.516 \).

2.8. Computing research trend score \( T_{v_t,v} \)

The research trend score measures the personal information of such potential collaborators \( v_t \) based on the research background and probability to work in \( v_t \) ’s inquired research topic. To avoid the quantity bias effect, the values of \( b_{v_t,v} \) and \( \phi_{v_t} \) must be normalized. This score is computed as follows:

\[
T_{v_t,v} = \frac{E_{v_t,v} \times 2^{-d_{v_t,v}}}{\sum_i E_{v_t,v_i} \times 2^{-d_{v_t,v_i}}} + \frac{\phi_{v_t}}{\sum_i \phi_{v_t,v_i}}
\]

(10)

2.9. Computing potential collaboration score \( C_{v_t,v} \)

After obtaining the research publication and personal information of \( v_t \) and \( v \) in terms of \( E_{v_t,v} \) and \( T_{v_t,v} \), the potential collaborator score \( C_{v_t,v} \) is computed from \( E_{v_t,v} \) and \( T_{v_t,v} \). But in fact the value of \( C_{v_t,v} \) is not a function of \( E_{v_t,v} \) and \( T_{v_t,v} \). The following two aspects must be involved. The first aspect is the degree of separation \( d_{v_t,v_i} \) and the second aspect is the seniority of \( v \) with respect to \( v_i \). For the first aspect, the higher \( d_{v_t,v_i} \) is, the less \( v_i \) is selected as a potential collaborator. But in the second aspect, any potential collaborator having high seniority compared with the others will have a better chance to be selected. The seniority is measured in terms of the number of published papers in the topic specified by \( v_t \). Let \( s_{v_t,v} \) be the seniority of \( v_t \) with respect to \( v_i \). The value of \( s_{v_t,v} \) depends upon the seniority status of \( v_t \) and \( v_i \). Table 3 shows an example of \( s_{v_t,v} \) for different pairs of seniority status measured in terms of percentage in different topics. Therefore, we proposed to compute \( C_{v_t,v} \) by combining the two aspects as follows:

\[
C_{v_t,v} = \frac{E_{v_t,v_i} \times 2^{-d_{v_t,v_i}} + \frac{T_{v_t,v} \times 2^{-d_{v_t,v_i}}}{\sum_i E_{v_t,v_i} \times 2^{-d_{v_t,v_i}}} + \frac{T_{v_t,v} \times 2^{-d_{v_t,v_i}}}{\sum_i \phi_{v_t,v_i}}}}{ \sum_i \phi_{v_t,v_i}} \times s_{v_t,v} \times s_{v_t,v} \times s_{v_t,v} \times s_{v_t,v} \times s_{v_t,v} \times s_{v_t,v}
\]

(11)

3. Experiments and performance evaluation

This section describes the data preparation, experimental setup, and evaluation method. The results from the experiments were evaluated in various aspects.

3.1. Data collection

In this study, 10 years of research publications from year 2000 and 2009 were considered. The research publication details for our data sets were extracted from SCOPUS,\(^2\) which is one of the largest bibliographic databases containing abstracts and citations for academic articles [22]. Because of active research collaboration as reported in the studies in [23,24,3,25,26], the publications in the area of Computer Science were selected. Prior to obtain the suitable data sets for implementing in the proposed methods, the following steps related to the collection of publications must be processed first:

1. Defining the potential topics. The potential selected topic were explored from Digital Bibliography and Library Project\(^3\) (DBLP) containing one of the largest repositories of CS bibliographic citations. Fortunately, the survey of Laneder [27] in DBLP divided the CS Field into 27 sub-topics.

2. Defining senior researchers. There are two reasons why we define senior researchers. One is that our proposed method takes into account the seniority status of the researchers. The other reason is that the co-authorship network is extremely

\(^2\) http://www.scopus.com

\(^3\) http://dblp.uni-trier.de
sparse since it is a small world network which follows the power law distribution [8,28]. To reduce the number of papers written by authors having published only one paper, we selected only papers written by at least one senior researcher. In this study, any researcher who published more than 10 papers was considered as a senior researcher. Otherwise, the researcher was considered as a junior researcher.

3. Collecting papers. For each possible topic, a query to search papers from SCOPUS was created. In this query, a specific topic name, e.g. "Neural Network", was used as a search phrase to seek out the papers this topic in computer science subject area. The years of publication between 2000-2009 and source types such as conference proceedings and journals are as shown in Fig. 3. The most important process is selecting only the author's name who have published more than ten papers so far. The number of papers published by each author is shown in the parenthesis next to the right of author's name. For example, the number of published papers of Melvin is 76. After submitting all criteria, the potential papers were retrieved. Note that about half of the potential topics must be removed in this step since some queried attributes were missing such as year of publication.

4. Checking downloaded data. There are some topics containing a lot of incomplete information such as abstracts, references. These topics were removed from the data sets. In addition, the number of senior researchers in some topics were too many or too few. To avoid the quantity bias effect, it is appropriate to filter out these topics. Only six topics with about 100 senior researchers per a topic were concentrated in our experiments. The chosen topics and the number of selected papers are shown in Table 4.  

3.2. Experimental set-up  

Since the papers were collected within 10 years, the number of publication years for constructing the co-authorship graph and evaluating the performance of our algorithm must be pre-defined. Fortunately, the study of Yoshikane [29] reported that a new author, who published a paper in one of core SCI listed journals in the present year, would not publish any paper in the preceding 7 years. Thus, it implies that an author may possibly re-publish papers during 7 years. Therefore, we decided to accept papers in preceding six years for building our model called DB-Data set and two years for testing called Test-Data set. These setting was also
used in Sachan’s work [9]. Consequently, the collected papers from years 2000 to 2007 were separated into three snapshots of time slice, each consisting of papers in six years, i.e., 2000–2005, 2001–2006, and 2002–2007. The papers in each set were used to independently created the models underlining the proposed methodology. For Test-Data set, there were two approaches for testing. The first one was tested with the different data sets referred as Approach-A, and the second was tested with the same data set referred Approach-B.

Table 5 shows the statistic of experimental data in particular snapshot based on Approach-A. A particular model was tested with the data set collected in the next two years from the currently considered year. All researchers in the co-authorship graph were subsumed to be possible inquiring researchers. The numbers of inquiring authors appearing in DB-Data set and Test-Data set in the same snapshot were the same. From the experimental database, it was found that about half of researchers in Test-Data set were not in the co-authorship graph constructed from the DB-Data set.

Table 6 shows statistic of data set for Approach-B. There are three DB-Data sets used to construct the co-authorship graph and compare the results tested with the same data. For example in snapshot-3 of both table, there are 3544 possible inquiring researchers in Test-Data set. Among these researchers, 1681 of them also appeared as possible inquiring researchers in DB-Data set. So the rest of 1853 possible researchers were the new coming
authors just appeared in Test-Data set. On the other hand, there are 1681 authors from all 6828 in DB-Data set who continued to produce their papers in Test-Data set. This may imply that the probability to select a researcher who continued to produce his papers from a set of possible inquiring researchers in the preceding six years was approximated as $1681/6828 = 0.25$.

The number of research topics was fixed at 6 as the number of research topics in Table 4. The related setting was defined in the same way as in [21]. To compute the probability $\phi_{n}$ of working in $\nu_i$’s current research topic, a fixed 3-year interval was used for computing the exponential moving average. The reason is that each DB-Data set snapshot consisted of papers within six years. So a half of six was defined for each computing.

3.3. Experimental results

With respect to an inquiring researcher $\nu_i$, a selected collaborator $r_n$ from DB-Data set was the researcher whose potential collaboration score $C_{n_r}$ was highest. If pair $\nu_i$ and $r_n$ existed in the Test-Data set, then the match was counted as a correct match. The number of tested inquiring researchers was taken from Test-Data set. The accuracy of our experiments were measured in terms of precision, mean recall, and mean reciprocal rank (MRR). The precision measure is the ratio between the number of correct matches and the number of recommended researchers from DB-data set. The recall measure is the ratio between the number of correct matches and the number of researchers collaborating with $\nu_i$ in Test-Data set. The reciprocal rank measure is the multiplicative inverse of the rank of the first correct match. Since the number of authors paper in Test-Data set in Tables 5 and 6 is 3.51, the algorithm recommended four researchers for each inquiring $\nu_i$.

Naturally, it is noticed that $C_{n_r}$ is not equal to $C_{n_r}$ due to the number of friends and non-friends of $\nu_i$ and $\nu_i$. The experimental results were evaluated in the following aspects:

1. Comparison of different year-interval snapshots.
2. Comparison of ranking based on existing publication evidence score $\mathcal{E}_{n_r}$, research trend score $\mathcal{T}_{n_r}$, and potential collaboration score $C_{n_r}$.
3. Comparison of research trend score $\mathcal{T}_{n_r}$ and research background similarity score $B_{n_r}$. Note that the score of $\mathcal{T}_{n_r}$ includes $B_{n_r}$ and $\phi_{n_r}$.
4. Comparison of our algorithm with the following algorithms: SSCM4D [10]; Lee’s work [11]; CollabSeer [13]; Liu et al.’s work [8]; Han et al.’s work [7]; and Sachan’s work [9].
5. Comparison of how seniority influences the results of our algorithm. Only senior status and junior status were investigated.
6. Comparison of how different degrees of separation affected the results.

The detail of each comparison is the following.

3.3.1. Comparison of different year-interval snapshots

The following 3-year interval snapshots, i.e. 2000–2005, 2001–2006, and 2002–2007 were focused. The current inquiring year was assumed to occur in year 2008–2009. Table 7(a) shows the accuracy of Approach-A and Table 7(b) shows the accuracy of Approach-B for sets of friend, non-friend, and neighbor researchers of an inquiring research. Notice that, regardless of Approach-A and Approach-B, the accuracy in column friend is highest but that of column non-friend is lowest. This implies that the performance of our algorithm was affected by the data sets and the year

### Table 9
The comparison of research trend score and research background similarity score.

<table>
<thead>
<tr>
<th></th>
<th>Friend</th>
<th></th>
<th>Non-friend</th>
<th></th>
<th>Neighbor</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>MRR</td>
<td>Precision</td>
<td>Recall</td>
<td>MRR</td>
</tr>
<tr>
<td>$T_{n_r}$ (with $\phi_{n}$)</td>
<td>0.405</td>
<td>0.735</td>
<td>0.610</td>
<td>0.099</td>
<td>0.227</td>
<td>0.097</td>
</tr>
<tr>
<td>weighted $B_{n_r}$</td>
<td>0.333</td>
<td>0.603</td>
<td>0.512</td>
<td>0.054</td>
<td>0.124</td>
<td>0.052</td>
</tr>
</tbody>
</table>

### Table 10
The accuracy comparison between our algorithm and the other algorithms.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Mean recalls ($\mu$) and variances ($\sigma^2$)</th>
<th>Difference of two means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Friend</td>
<td>Non-friend</td>
</tr>
<tr>
<td>Our algorithm</td>
<td>0.822</td>
<td>0.102</td>
</tr>
<tr>
<td>SSCM4D [10]</td>
<td>0.692</td>
<td>0.145</td>
</tr>
<tr>
<td>Lee’s work [11]</td>
<td>0.657</td>
<td>0.158</td>
</tr>
<tr>
<td>CollabSeer [13]</td>
<td>0.645</td>
<td>0.162</td>
</tr>
<tr>
<td>Liu et al.’s work [8]</td>
<td>0.625</td>
<td>0.169</td>
</tr>
<tr>
<td>Han et al.’s work [7]</td>
<td>0.591</td>
<td>0.178</td>
</tr>
<tr>
<td>SSCM4D [10]</td>
<td>0.184</td>
<td>0.192</td>
</tr>
<tr>
<td>Lee’s work [11]</td>
<td>0.142</td>
<td>0.216</td>
</tr>
<tr>
<td>CollabSeer [13]</td>
<td>0.109</td>
<td>0.228</td>
</tr>
<tr>
<td>Sachan’s work [9]</td>
<td>0.092</td>
<td>0.234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Friend</th>
<th>Non-friend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>MRR</td>
</tr>
<tr>
<td>Our algorithm</td>
<td>0.453</td>
<td>0.717</td>
</tr>
<tr>
<td>SSCM4D [10]</td>
<td>0.382</td>
<td>0.683</td>
</tr>
<tr>
<td>Lee’s work [11]</td>
<td>0.303</td>
<td>0.524</td>
</tr>
<tr>
<td>CollabSeer [13]</td>
<td>0.356</td>
<td>0.501</td>
</tr>
<tr>
<td>Liu et al.’s work [8]</td>
<td>0.345</td>
<td>0.482</td>
</tr>
<tr>
<td>Han et al.’s work [7]</td>
<td>0.326</td>
<td>0.458</td>
</tr>
<tr>
<td>Sachan’s work [9]</td>
<td>0.326</td>
<td>0.458</td>
</tr>
</tbody>
</table>

(a) The recalls and the statistic t-test was applied to confirm the difference between our accuracy and the others.

(b) The precisions and the mean reciprocal ranks (MRR) comparison between our algorithm and the other algorithms’
3.3. Comparison of research trend score and research background similarity score

The objective of this comparison is to test whether the probability $\phi_{n,n}$ of a collaborator $v_n$ to work in an inquired research topic or topic effects the accuracy or not. In Table 9, the research background similarity score $B_{n,n}$ was weighted by using the distance $d_{n,n}$ similar to potential collaboration score in Eq. (11). The weighted $B_{n,n}$ is computed as follows.

weighted $B_{n,n} = \frac{B_{n,n}}{\sum_{i \neq n} B_{i,n} \times 2^{-d_{i,n}}} \tag{12}$

3.4. Comparison of our algorithm and other algorithms

Table 10 shows the comparison between our algorithm and the other’s in case of friend and non-friend researchers. The techniques of SSKCM [10], Lee’s work [11], CollabSee [13] involved friend and non-friend researchers. Liu et al.’s work [8] and Han et al.’s work [7] concentrated only on a potential friend collaborator. The non-friend researchers were not recommended. Finally, Sachan’s work [9] concerned the problem of selecting a potential
non-friend collaborator. Their study focused on how to predict the collaboration links among researchers. Hence, we classified their technique as a non-friend research recommender. For each technique, the mean accuracy and its variance were reported. The statistical t-test was applied to test the following hypotheses between the mean of accuracy by our algorithm and by the other techniques.

\[ H_0: \text{The average accuracy of the given two models are the same, } \mu_1 = \mu_2 \]
\[ H_1: \text{The average accuracy of the given two models are not the same, } \mu_1 \neq \mu_2 \]

From the t-test in Table 10, it can be concluded that the average accuracy of our algorithm is not equal to the average accuracy of the other techniques at the 90% confidence level for friend case and 80% confidence level for non-friend case.

3.3.5. Comparison of seniority influences

This comparison is to find whether the seniority status of an inquiring research effects the accuracy of selecting a potential collaborator or not. We focused only on senior and junior status. Table 11 shows that the accuracy of recommending the collaborators to a senior researcher is lower than the accuracy of recommending to junior researchers. This is because a senior researcher has more collaborating friends than a junior researcher. This makes the pool of friend researchers of the senior researcher larger than that of the junior researcher. Therefore, the probability of selecting the right potential collaborator for a senior researcher is obviously low. Furthermore, the number of non-friends, i.e., friends of friends, proportionally grows in the number of friends and the possibility to select a right recommended researcher is also low. In case of a junior researcher, it is not difficult to recommend a potential collaborator from a small set of potential researchers. Thus, the accuracy in the case of junior status is higher than the case of senior status.

3.3.6. Comparison of effect of degrees of separation on each topic

Table 12 shows the effects of degrees of separation on the accuracy for friend, non-friend, and neighbor sets in each topic. The highest accuracy occurred in neural network topic in all friend, non-friend, and neighbor sets. The lowest accuracy was in the topic of Data Mining. But the average collaborators in Neural Network topic was fewer than those in Data Mining topic. This finding is not surprising since the fewer number of collaborators will increase the possibility to select the right recommended researchers. Therefore, the accuracy of recommendation is inversely proportional to the number of collaborators (friends) in the past period. In the same fashion, we may conclude that the hardest topic for recommending a collaborator is Data Mining. Moreover, Table 12 also shows that the higher the degree of separation is, the less accuracy is obtained. The last line of Table 12 summarizes the average accuracy of each degree of separation.

4. Conclusion

A new algorithm was proposed for recommending a potential collaborator based on five essential personal and academic factors of social proximity between two researchers, acquaintance among researchers, knowledge background of individual researcher as well as knowledge similarity between researcher pair, trend of research interest of each researcher within a studied period, and seniority of other researchers with respect to the inquiring researcher. In the past, some of these factors were not considered in other techniques. Furthermore, the following eight measures were introduced for evaluating the degrees of those five factors: (1) cohesive publication force; (2) possibility of publishing new paper score of potential collaborator; (3) cohesive co-authorship force; (4) existing publication evidence score; (5) research background similarity score; (6) probability of potential collaborators to work in the same research topic as the inquiring researcher; (7) research trend score; and (8) potential collaboration score. From the experiments, the following conclusions were inferred:

1. The recall of our algorithm when selecting the potential collaborator from set of past partners (friends) is 0.822 whereas from set of new partners (non-friends) is 0.294. Since the number of candidates in non-friends data set is larger than friends data set, the recall of non-friend set is rather small when compared with the recall of friend set.
2. The average recall based on the different year intervals of publication are not equal. The recall from snapshot years of 2002–2007 is 0.609 whereas snapshot years of 2000–2005 is 0.424. This shows the accuracy of our algorithm may be affected by the data sets and the year intervals. The accuracy of recommendation strongly depended on up-to-date data.
3. The average recall when used only the proposed structural approach, only the proposed semantic approach and proposed hybrid approach are 0.568, 0.539, and 0.609, respectively. These show that our proposed hybrid method is significantly better than a single approach.
4. The average recall based on only the knowledge background in semantic approach is 0.443. It was increased to 0.539 when the probability of trend toward the interest in inquired topic was involved. Therefore, the trend of research interest can influence the accuracy of recommendation.
5. The comparison results from our algorithm is significantly more accurate than the others’ algorithms. The average accuracy of our algorithm outperformed the other techniques at 90% confidence level for selecting past collaborators and 80% confidence level for selecting new collaborators.
6. Approximately 45% of collaborative work is the combination between senior researchers and junior researchers. Thus, this is the most popular joint working pattern found from the retrieved data sets. Recommending collaborators for a senior researcher publishing a large number of papers is more difficult than recommending collaborators for a junior researcher with a few published papers.
7. The higher the degree of separation is, the less accuracy of recommending a collaborator is obtained.
8. The accuracy of recommendation is inversely proportional to the number of researchers having collaborations in the last period. The hardest topic for recommending a collaborator is Data Mining but the easiest topic is Neural Network.

Although the experimental results of our proposed algorithm and new measures are significant in terms of accuracy, precision, recall, and mean reciprocal rank, there are several other issues worth being further studied. For example, some issues concerning the recommendation in other research fields besides computer science, cross-cultural research areas, and multiple collaborators from different fields may be addressed. Furthermore, the concepts behind each proposed measure can be adapted and modified to other academic fields.

Acknowledgments

The authors thank the supporting foundations of the internship program of the National Institute of Informatics (NII), Japan. We
would also like to express our appreciation to Associate Professor Ryutaro Ichiye for providing the valuable materials, comments and discussion.

References


