Scientometrics
DOI 10.1007/s11192-016-1999-x

Toward a better scientific collaboration success prediction model through the feature space expansion

Fahimeh Ghasemian[1] · Kamran Zamanifar[1] · Nasser Ghasem-Aqaee[1] · Noshir Contractor[2]

Received: 13 December 2015

Abstract The problem with the prediction of scientific collaboration success based on the previous collaboration of scholars using machine learning techniques is addressed in this study. As the exploitation of collaboration network is essential in collaborator discovery systems, in this article an attempt is made to understand how to exploit the information embedded in collaboration networks. We benefit the link structure among the scholars and also among the scholars and the concepts to extract set of features that are correlated with the collaboration success and increase the prediction performance. The effect of considering other aggregate methods in addition to average and maximum, for computing the collaboration features based on the feature of the members is examined as well. A dataset extracted from Northwestern University's SciVal Expert is used for evaluating the proposed approach. The results demonstrate the capability of the proposed collaboration features in order to increase the prediction performance in combination with the widely-used features like h-index and average citation counts. Consequently, the introduced features are appropriate to incorporate in collaborator discovery systems.

Keywords Scientific collaboration · Collaborator discovery system · Collaboration network · Collaboration success · Hypergraph · Machine learning

✉ Kamran Zamanifar
zamanifar@eng.ui.ac.ir

Fahimeh Ghasemian
ghasemianfahime@gmail.com

Nasser Ghasem-Aqaee
aghaee@eng.ui.ac.ir

Noshir Contractor
nosh@northwestern.edu

[1] Department of Computer Science, Isfahan University, Hezarjarib Ave., Isfahan, Iran

[2] SONIC Lab, Northwestern University, Evanston, IL, USA

Published online: 03 June 2016

Introduction

Scientific collaboration is a social process in which people share their human capitals to produce knowledge (Bozeman et al. 2013). A study of more than 21 million published articles from 1945 to the present shows that teams act better in the production of high impact, highly-cited articles (Börner et al. 2010). The complexity of scientific problems, requirements to access to new and expensive data and research tools, technology progress which facilitates communication and sharing are some reasons for increasing desire of scientific collaborations (Olson et al. 2008). Therefore, collaboration among the appropriate individuals is becoming more important for the scientific progress (Schleyer et al. 2012).

In response to this, areas of research known as e-science, cyber-infrastructure and science of team science have emerged (Jirotka et al. 2013) for the study of scientific collaboration patterns and developing technologies and infrastructures to support scientific collaborations. Also this topic attracts especial interest as a domain of CSCW[1] studies in the last few years (Schmidt and Bannon 2013).

One of the challenges in these studies is exploring how to build optimal (regardless of one how defines the term) teams (Börner et al. 2010). The prerequisite for this is to extract successful research collaboration patterns or answer this question that what makes a collaboration successful. Recently some efforts are devoted for developing CSCW systems which support individual researchers' effort to form optimal collaborative relationships (Schleyer et al. 2012). Integration of collaboration networks is essential in these systems (Schleyer et al. 2012) for below reasons:

1. Studies have shown that individual characteristics are not the only factors in individual or collaboration success and social capital which is determined based on the social network structure of individuals is another important factor (Abbasi et al. 2014).
2. The link structure among experts and skills is a valuable piece of information for the estimation of skill level of experts. In this manner, the skills that are directly related to the experts are not a sole concern.
3. Previous works point out that trust is an important factor in the collaboration success (Stokols et al. 2008; Bennett and Gadlin 2012). Closeness of the experts in the collaboration network can be used as an estimation of the trust level.

Although some team formation algorithms have been proposed which exploit the collaboration network of experts (e.g. Lappas et al. 2009; Wi et al. 2009; Gajewar and Sarma 2012; Li et al. 2015) and can be used in the collaboration discovery systems but these algorithms have some limitations:

1. These algorithms use the weighted linear combination of two or a limited number of team characteristics (e.g. expertise level and closeness of the members) to evaluate the fitness of the teams (Dorn and Dustdar 2010), while these weights should be determined manually by the users. This approach can be appropriate as they give the flexibility to users to form teams based on the measures that are more important for them, but it has some disadvantages. As long as the number of considered factors (team characteristics) for team formation is limited, this approach can help, but the team success is affected by many factors and determining the weight of each manually is difficult. Even in cases that the weights are determined appropriately, the

[1] Computer supported cooperative work.

assumption that the team success is a linear function of these factors might not be a correct assumption.
2. Most algorithms just exploit the collaboration network to compute the closeness of the experts and minimize the communication cost among team members.

In this article, our goal is to motivate the CSCW and related communities to consider more beneficial features in tools for forming successful research collaborations. Therefore we consider the problem with the prediction of scientific collaboration success to determine these features. The term success means that the collaboration manages to achieve its mission (Bennett and Gadlin 2012). So depending on the missions, the success for a research team in the field of medicine can be the successful development of a vaccine or for another team with the goal of writing a research proposal, obtaining a grant is the success. Generally, in academic environment, success of a research team is measured based on its effect on the scientific community (the number and citation counts of the published articles), while in industry, the success of a research team is usually measured based on the financial gain.

Following most previous works in this area (Bozeman et al. 2013), the scientific collaboration is equated with the co-authorship and the citation number within a time frame of five years is considered as the success factor. The objective here is to extract more discriminating features which if combined with the common features such as h-index or average citation counts (widely-used in the previous works), it would improve the predictive model. Improvement means that the differences between the real success of collaborations and the estimated success values by the model are decreased. Since a predictive model estimates the success based on the characteristics (features) extracted from the collaborations, appropriate selection of features is highly important for improvement of the model.

For this purpose, the link structures among the scholars (experts) and concepts (the subject categories of the articles which are considered as the skills of the scholars) are of concern in order to:

1. Estimate the expertise level of scholars applying a score propagation process that spreads the scores (which are considered as the expertise level) through the relations among the scholars, relations between the scholars and concepts and relations among the concepts. So, having collaborations on a concept is not the only determining factor. The expertise level of the scholars who are in the neighborhood and the scholar's collaborations on the related concepts are other factors that affect on the expertise level of the scholar for a concept.

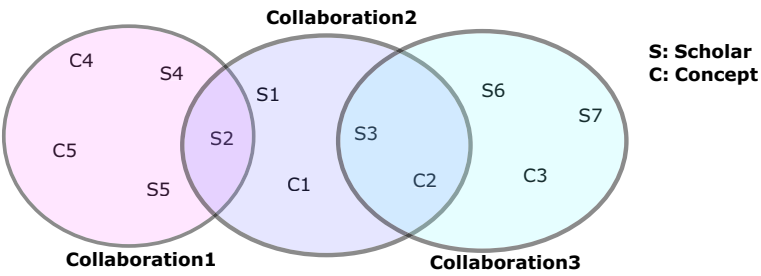

Fig. 1 An example of a hypergraph for modeling three scientific collaborations

2. Rank the scholars based on their relations with the other scholars.

For computing the two above-mentioned features, hypergraph is used for representing the collaboration relations. A hypergraph is a generalization of the ordinary graph which can be used to model high-order relations (Tan et al. 2011). By using this modeling approach, we can accurately capture the high-order relations among the scholars and concepts in each collaboration without any information loss. Also, recent studies have shown the usefulness of this approach for modeling the collaboration relations (Sharma et al. 2014). An example of a hypergraph for modeling three scientific collaborations is illustrated in Fig. 1 which includes two types of vertices: scholar and concept and captures three types of relations: collaboration relation among the scholars, relations among the concepts and relations between the scholars and concepts. Integration of all these three types of relations is important in capturing the competence of the scholars for participating in a collaboration. The relations of a scholar with other scholars show with whom the scholar can collaborate more successfully, also in the light of the count and competence of the neighboring scholars, the competence of a scholar can be estimated better. The relations between the concepts provide semantically rich information about the related concepts and improve estimation of the expertise level of a scholar for a concept. So, a scholar who hasn't participated in any collaboration on a particular concept but has some collaborations on related concepts, will receive some skill level for that concept. Finally the direct relations between scholars and concepts show the collaboration experience of the scholars on concepts. To the best of our knowledge no previous works exploit these relations altogether for analyzing the success of the scientific collaborations.

Also, to preserve more information about the collaborations, in addition to aggregate functions such as the maximum and average, the scholars are clustered based on their values for the selected feature and then the frequency of each cluster in the collaborations is computed. Applying Northwestern University's SciVal Expert as the data infrastructure, our experimental results show that the proposed features significantly improve the prediction performance. So these features are valuable for incorporating in collaborator discovery systems. In addition, we show that the proposed modeling approach which integrates different kind of relations is a better choice compared to a model which just considers the relations among the scholars or the relations between scholars and concepts. The rest of this article is organized as follows: the previous works are reviewed in "Literature review" section; some preliminary knowledge is described in "Background" section; the proposed approach is explained in "Collaboration success prediction" section; experiments are made to validate the approach in "Experiments" section; the article is concluded in "Discussion" section and finally suggestions for further research are discussed in "Future work" section.

Literature review

Scientific collaboration as a research field is discussed in different disciplines including information science, psychology, management science, computer science, sociology, research policy, social studies of science and philosophy (Sonnenwald 2007). Each discipline focuses on a specific aspect of collaboration. In this section, research collaboration literature that examines what constitutes the factors of a successful collaboration is reviewed. Moreover, due to the consideration of citation count as the success measure of the collaborations in this paper, previous works about the citation count prediction are

considered as well. Also, team formation algorithms are examined in this section as they optimize utility functions that measure the success of the collaborations.

Factors of a successful collaboration

Research collaboration literature which is mostly in the area of social science can be divided into two categories: the works that examine how scholars are motivated to form collaborations and the ones that study how scholars should form collaborations to be successful or what constitute the factors of a successful collaboration. Some of these factors are related to collaborators, some are related to the collaboration process and others are related to the environment in that collaboration occurs (Bozeman et al. 2013). Stokols et al. (2008) point to the regular and effective communication, mutual respect, trust and familiarity among team members as the factors of the collaboration success. They also discuss the importance of personal characteristics for the collaborative work. Bennett and Gadlin (2012) conclude those teams that their members understand the overall goals of the project are more effective. Moreover, they point to the regular communication, and the way that members resolve their conflicts as success factors. Also, they consider trust as a foundation for the team's success. Eslami et al. (2013) assess the effect of social structure of collaborations on the performance of research collaboration (i.e. number of published articles). They conclude that there is a significant association between the number of publications and the manner through which scholars are interconnected to one another. Skilton (2008) analyzes a sample of articles published in high-ranking journals and finds that the articles introduced by teams including frequently cited scholars and teams the members of which have diverse disciplinary backgrounds receive more citation counts. Whitfield (2008) reports that the previous collaboration experience of the members have a positive effect on the collaboration success. Cummings and Kiesler (2008) discuss that prior experience with a collaborator reduces the negative impact of distance and disciplinary difference in collaborations.

Although research collaboration is well-studied in social sciences and there are some guidelines to help scholars form successful collaboration (e.g. Bennett and Gadlin 2012), there is still a need for algorithms to form successful collaborations automatically regarding the factors of successful collaborations (Börner et al. 2010). Also a lot of unanswered questions remain about how to best use information technology for developing these algorithms (Schleyer et al. 2012).

Citation count prediction

When each article is considered as a scientific collaboration and its citation counts as its success, there would be a close similarity between this and citation count prediction problem, of course with some differences. Although the problem of citation count prediction of articles is difficult and the performance of the proposed algorithms are not still satisfactory (Fu and Aliferis 2010), in collaboration success prediction problem, it is not necessary to predict the exact citation count. It would suffice to predict which collaboration will be more successful in relation to other possible collaborations and it makes this problem easier. But in predicting the research collaboration success, no information like venue, order of authors and reference list of the article is available. What is known are the scholars and concepts that they will collaborate on.

Callaham et al. (2002) use decision trees and 204 publications and could explain 0.14 in the variation of citation counts 3.5 years after the publication. They conclude that the journal impact factor is the most predictive feature. Castillo et al. (2007) use linear

regression and decision trees to predict the citation counts 4.5 years ahead. They find that the citation counts accumulated within the first year after the publication are highly correlated with the citation counts. Fu and Aliferis (2010) predict the citation count with machine learning methods and a combination of content-based and bibliometric features. Wang et al. (2012) examine the factors influencing the citation counts and conclude that the paper quality and the reputation of the first author contribute to the creation of future citation impact. Didegah and Thelwall (2013) study different features and find venue prestige and the number of citations attracted by the references of a paper to be the strongest features. Yan et al. (2012) consider features including venue prestige, content novelty, diversity and authors' influence. Yu et al. (2014) propose a model which predicts the citation counts of articles with a mixture of features including the number of references, the number of authors, the h-index of the first author, the number of papers published by the first author, maximum h-index of the authors, maximum average citation to the papers published by the authors and the impact factor of the journal.

To sum up, citation count of an article may be influenced by four factors: author characteristics, venue characteristics, field characteristics and article characteristics (Yu et al. 2014). Algorithms that have been proposed for citation count prediction of the articles mostly focus on the characteristics of the articles and venues and less effort has been made to extract more informative features based on the previous collaborations of scholars with one another. Moreover, few previous works benefit from the link structure among the scholars and concepts in collaboration networks that can contribute to extract more informative features. Another point to mention is that features based on article and venue characteristics could not be used for the collaboration success prediction as we just know the authors and concepts that they will collaborate on.

Team formation algorithms

The team formation problem is mainly studied in the field of operations research. However, the underlying network structure that reflects the relationships among the scholars has been ignored in these traditional algorithms (Li et al. 2015). Lappas et al. (2009) are the first to solve the team formation in the presence of the scholars' collaboration networks. They prove that this problem is NP-hard and propose two approximate algorithms: RarestFirst and Enhanced-Steiner. Both algorithms form teams that their members cover the required skills while the communication cost among them is minimized. The communication cost is measured in terms of the diameter and the minimum spanning tree cost. Dorn and Dustdar (2010) measure the effectiveness of a team in terms of the weighted linear combination of skill fulfillment and team connectivity based on the distance measures. They use Genetic Algorithm and Simulated Annealing to solve the problem. Gajewar and Sarma (2012) consider the communication cost in terms of density-based measures. Awal and Bharadwaj (2014) use the genetic algorithm to form teams with optimized Collective Intelligence Index (CII). CII is defined as the linear combination of knowledge competence and collaboration competence of a team. Knowledge competence computes the expertise score of the participating scholars and collaboration competence captures trust among the scholars.

Previous studies show that collaboration success is affected by many factors but almost all algorithms proposed for collaboration (team) formation try to optimize the linear combination of skill coverage and communication cost which is a nave function for evaluating the success of collaborations. In this paper our goal is to exploit the collaboration network to extract more characteristics of collaborations which help to increase the prediction performance of the

collaboration success. So these characteristics are good candidates for considering in scientific collaboration formation algorithms.

Background

The link structure among the scholars and the concepts is a valuable piece of information for the estimation of the scholars' expertise level. In this manner, the concepts that are directly related to the scholars are not a sole concern. Moreover, some studies have shown that individual characteristics are not the only factors in individual or collaboration success and social capital which is determined based on the social network structure of individuals is another important factor (Abbasi et al. 2014). To include all the relations explained in "Introduction", we model the previous collaborations of the scholars on different concepts with a hypergraph and a ranking algorithm is used to compute the similarity of the scholars to concepts and rank the scholars based on their positions in the collaboration network. The resulted ranking scores are used as a measure of the scholars' social capital. In this section, modeling with hypergraph and the algorithm for ranking on hypergraph are discussed.

Hypergraph

Let $G(V, E, W)$ denote a weighted hypergraph where V is the set of vertices, E is the set of hyperedges and W is the weight of hyperedges. Each hyperedge $e \in E$ is a subset of V which is used to model the high-order relations. Notations used for representing a hypergraph are listed in Table 1 (Tan et al. 2011). As observed, each hypergraph is formally described by four matrices including H, W, D_v and D_e. The matrix H is an incidence matrix for capturing the membership of vertices in hyperedges. W includes hyperedge weights. D_v and D_e are diagonal matrices containing the vertex and hyperedge degrees respectively. The problem of ranking vertices is addressed as assigning a score to each vertex according to its relevance to a query vector through the minimization of the following function (Tan et al. 2011):

$$Q(f) = \frac{1}{2}\sum_{i,j=1}^{|V|}\sum_{e\in E_h}\frac{w(e)h(v_i,e)h(v_j,e)}{\delta(e)}\left|\left|\frac{f_i}{\sqrt{d(v_i)}}-\frac{f_j}{\sqrt{d(v_j)}}\right|\right|^2+\mu\sum_{i=1}^{|V|}||f_i-y_i||^2 \qquad (1)$$

where $y = [y_1, y_2, ..., y_{|V|}]^T$ is the initial score of the vertices, $f = [f_1, f_2, ..., f_{|V|}]^T$ is the score vector that the algorithm assigns to the vertices and $\mu > 0$ is the regularization parameter

Table 1 Notations used for formal representation of a hypergraph

Incidence matrix	$H(v,e) = \begin{cases} 1, & v \in e \\ 0, & \text{otherwise} \end{cases}$	Matrix that shows the memberships of vertices in the hyperedges
Hyperedge weight matrix	W	Diagonal matrix that elements on the diagonal are the weight of hyperedges
Vertex degree	$d(v) = \sum_{e\in E} h(v,e)W(e)$	The sum of the weight of hyperedges that v is a member of them
Hyperedge degree	$\delta(e) = \sum_{v\in V} h(v,e)$	The number of vertices in e
Vertex degree matrix	D_v	Diagonal matrix that elements on the main diagonal are the degree of vertices
Hyperedge degree matrix	D_e	Diagonal matrix that elements on the diagonal are the degree of hyperedges

which controls the relative importance of these two terms. The optimal score vector (f^*) is obtained when $Q(f)$ is minimized (Eq. 2). This minimization causes the spread of the scores among the vertices regarding their membership in the hyperedges.

$$f^* = argmin_f Q(f) \tag{2}$$

The minimization of $Q(f)$ can be obtained through an iterative approach similar to Random Walk (please refer to Tan et al. 2011 for further details):

$$f^{(k+1)} = \alpha A f^{(k)} + (1 - \alpha)y \tag{3}$$

where α is $\frac{1}{1+\mu}$ and A is a transition matrix:

$$A = D_v^{\frac{-1}{2}} H W D_e^{-1} H^T D_v^{\frac{-1}{2}} \tag{4}$$

In this article, this algorithm is applied for ranking vertices based on their link structure and also, computing the similarity of the vertices to a target vertex. In the first case of ranking, the same initial score values are assigned to all the vertices, and in the second case, the query vector is formed by assigning 1 to the target vertex and 0 to others.

Scientific collaborations as a hypergraph

Let G_{T_1,T_2} denotes the hypergraph which includes all the collaborations from the year T_1 to T_2. n_s and n_c denote the number of scholars and concepts in the hypergraph respectively. A unique number (index) is assigned to each scholar and concept. This index shows the position of the scholar or concept in the query or score vector. For scholars, this index starts from 1 to n_s and for concepts is $n_s + 1$ to $n_s + n_c$. Each collaboration in the year T is denoted by e_T which is composed of a set of scholars V_a and a set of concepts V_c. For each collaboration, the ranking algorithm is run on $G_{T-d,T-1}$ to measure the similarity of the

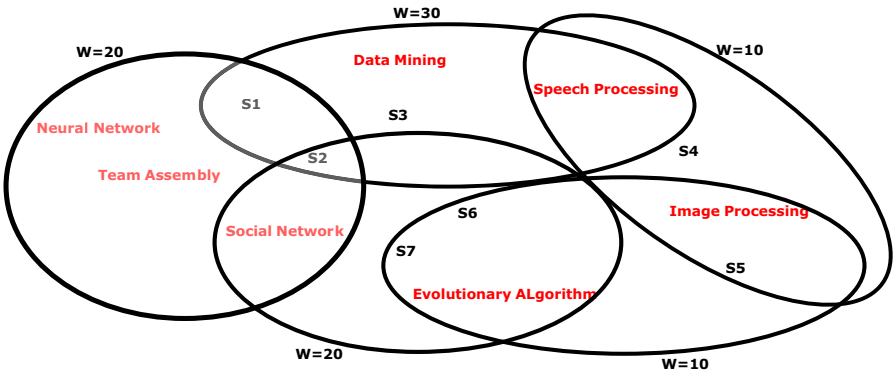

Fig. 2 Example of a small hypergrpah used for showing the result of the ranking algorithm (Fig. 3). Each *oval* (hyperedge) represents a collaboration among a set of scholars on a set of concepts and the weight (w) of each hyperedge shows the citation counts of the collaboration output

Fig. 3 **a** The result of ranking scholars, **b** Similarity of the scholars to Image Processing

scholar members to concepts as an estimation of their skill level based on their past collaborations. Where d is the time interval considered for constructing the hypergraph. Each time a query vector is formed for one of the concepts of V_c by assigning 1 to the selected concept and 0 to the others in the query vector (Eq. 5). Next, the ranking algorithm is run to rank the scholars for each query vector.

$$y_i^{sc} = \begin{cases} 1, & i = index(sc) \\ 0, & \text{otherwise} \end{cases}, \quad 1 \le i \le (n_s + n_c) \tag{5}$$

where sc is the desired concept.

For more clarification, an example of using this algorithm for ranking scholars and computing their similarity to a concept is presented in Fig. 2, where each one of the hyperedges is composed of two types of vertices (scholar and concept) and represents the collaboration among the scholars on some concepts. The final scores of the scholars for different initial scores are shown in Fig. 3. In each case the name of a scholar is depicted based on his/her score. Those scholars that earn higher scores appear larger with darker color. The result of ranking when the same initial values are assigned to all the vertices (scholars and concepts) are shown in Fig. 3a. As expected, $S1$, $S2$ and $S3$ received the best ranking as they had collaboration with the weight of 30, followed by $S6$ and $S7$ who had collaboration with weight equals to 20 and also are neighbor with $S2$ who had a good ranking score. $S4$ and $S5$ received the lowest ranks because the weights of their collaborations are less than the others. But since $S5$ is neighbor with $S6$ and also participated in two collaborations, its score is higher than $S4$'s score.

For computing the similarity of the scholars to the concept of Image Processing, first the query vector illustrated in Fig. 4 is formed. Next, the ranking algorithm is applied to rank the vertices. The names of the scholars with similarity to the concept of Image Processing are presented in Fig. 3b. As expected $S5$ who had two collaborations for Image Processing, received the highest similarity score, followed by $S6$ and $S4$ who had one collaboration on Image Processing. As observed although $S1$, $S2$, $S3$ and $S7$ did not have any collaboration on this concept, due to their relations with the scholars or concepts who are in direct relations with Image Processing, received some non-zero similarity.

S1	S2	S3	S4	S5	S6	S7	Speech Processing	Data Mining	Evolutionary Algorithm	Image Processing	Social Networks	Team Assembly	Neural Networks
0	0	0	0	0	0	0	0	0	0	1	0	0	0

Fig. 4 The query vector for ranking vertices based on their similarity to the concept of Image Processing

Collaboration success prediction

In this section, we introduce our approach for prediction of the collaboration success.

Data gathering

Although in recent years, social networking sites such as LinkedIn[2] have been developed for scholars to facilitate communication among the researchers and also to determine their expertise; recognition of the scholars' expertise just based on the information they enter, has faced many challenges. Individuals often lack motivation to update their profiles. Self-declaration of information usually does not have enough accuracy and correctness. For example, people may exaggerate in expressing their information or maybe they do not enter their information completely in order not to have more responsibilities. Therefore, determining the researcher's expertise based on the manually entered information is not effective and has many problems (Fazel-Zarandi and Fox 2013). Automatic extraction of the information on the research activity of the scholars is on an increase. The term research networking systems has been used to refer to the systems which automatically gather information about the scholars. In these systems, authorized databases and sources are used for the extraction of the scholars' profiles and their collaboration network. Scival Expert is a research networking system which has been developed by Elsevier. Northwestern University applies the Scival Expert to represent the information about its scholars. This information is expressed using VIVO[3] ontology and is accessible from a SparQL Endpoint.[4]

The extracted information such as the first and last names of the scholars, title and citation counts of their articles are saved in a database. Since it is not possible to access the related concepts of articles using SparQL Endpoint, this database is restricted to articles in the field of Medicine. Then, the rPubmed library in R is used for retrieving MeSH[5]. terms of the articles from PubMed Database through PMID of the articles. The MeSH is a standard terminology that PubMed uses for expressing the subject categories or related concepts of articles. PubMed employed skilled subject analysts to examine journal articles and assign the most specific MeSH terms to each article (Pubmed 2005). For instance, the MeSH terms of the article "Diagnosis and management of heart failure with preserved ejection fraction: 10 key lessons" are "Animals", "Echocardiography", "Heart Failure/diagnosis", "Heart Failure/physiopathology", "Heart Failure/therapy", "Heart Rate", "Hemodynamics", "Humans", "Natriuretic Peptide", "Brain/blood".

As for training the model, the citation counts broken by year are necessary; Scopus API[6] is used to add them. The number of collaborations, scholars, MeSH terms and average collaboration size (average number of the scholars in each collaboration) for each year from 1990 to 2006 are shown in Table 2. The tendency for more collaborative research can be observed in this dataset as well, since the average of collaboration size is on an increase.

[2] www.linkedin.com.

[3] www.vivo.com.

[4] http://vivo.scholars.northwestern.edu/.

[5] Medical Subject Heading.

[6] http://dev.elsevier.com/.

Table 2 The characteristics of the extracted dataset from Northwestern University's Scival Expert

Year	Number of collaborations	Number of MeSH terms	Number of scholars	Average collaboration size
1990	1102	7488	3502	4.47
1991	1229	7851	4138	4.61
1992	1290	8344	4369	4.76
1993	1349	8649	4606	4.83
1994	1458	9600	5087	4.96
1995	1517	9971	5603	5.1
1996	1629	10,536	6161	5.49
1997	1731	10,742	6649	5.55
1998	1676	10,765	6829	5.68
1999	1911	11,545	8088	6
2000	1926	11,961	7842	5.75
2001	2023	12,230	9380	6.29
2002	2291	12,802	9103	5.92
2003	2429	13,035	9792	6.07
2004	2627	13,588	10,960	6.11
2005	2720	13,952	11,406	6.24
2006	2805	13,923	12,479	6.63

Success prediction using machine learning techniques

Our approach for predicting the collaboration success consists of two phases: training and evaluation. Collaborations from the year 2000 to 2005 (4238 instances) are used for training and the year 2006 (744 instances) for evaluation of the model. The reason that 2000 is the first year used for constructing the train dataset is that a time window of 10 years length is used to extract the features of collaborations of one year. Also for the construction of the training and evaluation sets, those collaborations of any year whose members do not exist in the previous 10-year collaboration network, are removed. For example, if there is a collaboration in the year 2000 that at least one of its members does not exist in the collaboration network of years 1990 to 1999, that collaboration is removed since its features cannot be extracted.

We construct and save a hypergraph for each time interval listed in Table 3 and use it to extract the features of the collaborations in the target year. Each hypergraph is saved as a set of matrices including D_v, D_e, W, H and A. Also, for each hypergraph, two hash functions are also saved which map the concept name and uid[7] of the scholars to an integer number. This integer is used as an index for determining the corresponding row or column of a scholar or concept in D_v, H and A. Also these indices determine the position of the scholars or concepts in the query and score vectors. As explained before, for scholars, these indices start from 1 to n_s and for concepts start from $n_s + 1$ to $n_s + n_c$. For more clarification, the matrix of H of the hypergraph depicted in Fig. 2, is illustrated in Fig. 5.

[7] Scival Expert assigns a unique identifier (uid) to each scholar.

Table 3 List of the time intervals used for constructing the hypergraphs

Time interval	Target year
1990–1999	2000
1991–2000	2001
1992–2001	2002
1993–2002	2003
1994–2003	2004
1995–2004	2005
1996–2005	2006

These hypergraphs are used for the extraction of the features of the collaborations in the target years

	C1	C2	C3	C4	C5
S1	1	1	0	0	0
S2	1	1	0	0	1
S3	0	1	0	0	0
S4	0	0	1	0	0
S5	0	0	1	1	0
S6	0	0	0	1	1
S7	0	0	0	0	1
Speech Processing	0	1	1	0	0
Data Mining	0	1	0	0	0
Image Processing	0	0	1	1	0
Evolutionary Algorithms	0	0	0	1	1
Social Networks	1	0	0	0	1
Team Assembly	1	0	0	0	0
Neural Network	1	0	0	0	0

Fig. 5 The matrix of H of the hypergraph illustrated in Fig. 2

For computing the features of the collaborations of each target year (e.g. 2000), first the scholars and concept members of each collaboration are mapped to their indices using the hash functions, next, the desired features are extracted using the hypergraph of the related interval (e.g. 1990–1999). Feature extraction for constructing the train dataset is shown in Fig. 6. Beginning from 2000, collaborations during 1990–1999 are used to extract the features of collaborations of the year 2000. Then the time window is moved forward to 1991 and from 1991 to 2000, the features of the collaborations of the year 2001 are extracted and in this way the features of all collaborations during 2000 to 2005 are extracted. The same approach is used to extract the features from the evaluation dataset.

Collaboration relations modeling

For modeling the collaborations, a hypergraph with two types of vertices (scholars and concepts) are used. The weight of the hyperedges is considered as the citation number of

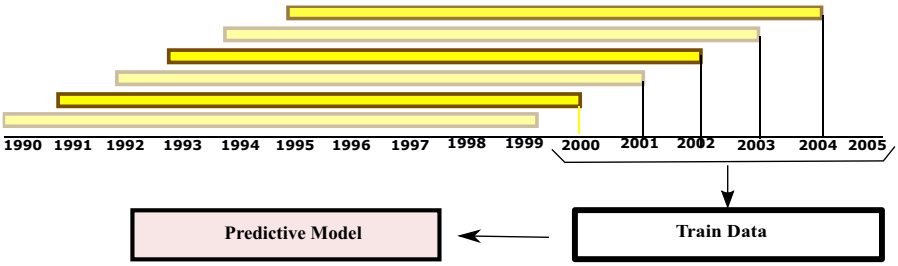

Fig. 6 The train phase of the predictive model

the collaboration output. To eliminate the bias, the weights of each year are normalized through:

$$w'(e_i) = \frac{w(e_i)}{max(w_T)} \tag{6}$$

where $max(w_t)$ is the maximum weight of all hyperedges in the given year and $w(e_i)$ is the weight of the hyperedge.

In the case of scientific collaboration, recent collaborations in the time window are more important as they can better show the current scientific situation of the scholars. Temporal dynamics are accounted by adjusting the weights of the hyperedges. An exponential kernel is used for this adjustment as below:

$$w'_{t_i} = (1-\theta)^{T-t_i}\theta W_{t_i} \tag{7}$$

where w_{t_i} is the weight of the hyperedge, t_i is its time label and T is the target year. The parameter $0 \leq \theta \leq 1$ determines the rate of increase which is set to 0.6. This value is chosen in a way that the extracted features have the maximum correlation with the collaboration success.

For computing the transition matrix in hypergraph, Eq. 4 is applied where the degree of each vertex is considered as the number of hyperedges that it is a member of. Because our experiment shows that for this application, it works better than the sum of the weight of the hyperedges.

Another point about modeling the collaboration relations with the proposed modeling approach is the difference between the degree ($d(v)$) distribution of scholar and concept vertices. Generally concept vertices have higher degree than scholars. For instance the average degree of concepts and scholars of the collaborations from the year 1990 to 2006 are 25.18 and 2.14 respectively. Also, MeSH which is used for specifying the concepts of articles has a hierarchical structure (from general to more specific terms). So the degree of a general term like "Animal" or "Human" is very high and the degree of a term like "Hemodynamics" that is more specific is low. Since the transition matrix is normalized by the degree of the vertices, these differences cause that the score of the scholars be more affected by the scores of the scholars and the specific terms in the neighborhood which is desirable.

Feature extraction

For training a predictive model, feature extraction is the most crucial part. As the only available information is the previous collaborations of the members and their position in

the collaboration network, the set of features should be extracted in a way that would contribute to distinguish the successful collaborations from unsuccessful ones. These features can be extracted based on the expertise level, familiarity of the members with one another or their structural centrality. For each collaboration e_T, the following features are extracted for each scholar member ($v_a \in V_a$):

1. Features that represent the expertise level of the scholars

   (a) The *h_index* is a scientometric indicator for measuring the productivity of the scholars based on their citation pattern. It is defined as "the highest number of papers that received h or more citations" (Egghe 2006). First articles are sorted in decreasing order based on their citation counts, next the *h_index* of the scholar v_a is computed as below:

   $$h\_index(v_a) = argmax_h(c_k \geq h | 1 \leq k \leq h) \tag{8}$$

   where c_k is the citation count of the kth article of the scholar.

   (b) The *g_index* is a modification of the *h_index* which is sensitive to the level of the highly cited papers. It is defined as "the highest number of g of papers that together received g^2 or more citations" (Egghe 2006). First articles are sorted in decreasing order based on their citation counts, next the *g_index* of scholar v_a is computed as below:

   $$g\_index(v_a) = argmax_g\left(g^2 \leq \sum_{k \leq g} c_k\right) \tag{9}$$

   where c_k is the citation count of the kth article of the scholar.

   (c) *The average of citation counts* This feature is computed for the scholar v_a using below equation.

   $$citation\_avg(v_a) = \frac{\sum_{k=1}^{n_a} citation(p_k)}{n_a} \tag{10}$$

   where n_a is the number of the articles and p_k is the kth article of the scholar. $citation(p_k)$ is the citation count of the article p_k.

   (d) *The similarity to concepts* The similarity of the scholar v_a to the concept v_c is measured using Eq. 11.

   $$(Sim_c)^{k+1} = \alpha A_{T-10,T-1} Sim_c^k + (1-\alpha) y^c \tag{11}$$

   where Sim_c^{k+1} is the ranking score vector or similarity of all the vertices of $G_{T-10,T-1}$ to the concept $v_c \in V_c$ after $k+1$ iterations, T is the year of the collaboration e_T, $A_{T-10,T-1}$ is the transition matrix of the hypergraph $G_{T-10,T-1}$ and y^c is the query vector for the concept v_c. This query vector is constructed using Eq. 5. Equation 11 is stopped after convergence and the resulted score vector (Sim_c^*) is used as the similarity or skill level of the scholars to the concept v_c. So the similarity of the scholar v_a to the concept v_c is:

   $$Sim_c(v_a) = Sim_c^*(index(v_a)) \tag{12}$$

   where $index(v_a)$ is the index of the scholar v_a.

   (e) *Fuzzy success and unsuccess* First the citation count of all the articles of each year is normalized. Next, all the articles of the scholar v_a are mapped into successful and unsuccessful space based on their normalized citation counts

using the fuzzy membership functions shown in Fig. 7. Finally the scholar is mapped into these two spaces using the average of success and uncusess of his/her articles. These features are formulated in the below equations.

$$Fuz\_success(v_a) = \frac{\sum_{k=1}^{n_a} success(p_k)}{n_a} \tag{13}$$

$$Fuz\_unsuccess(v_a) = \frac{\sum_{k=1}^{n_a} unsuccess(p_k)}{n_a} \tag{14}$$

where n_a is the number of the articles and p_k is the kth article of the scholar v_a.

2. Familiarity of the scholars:

   (a) *The average of Jaccard similarity* This similarity measure is applied to compute the familiarity level of the scholars with one another. This measure is based on the number of common neighbors of the two vertices. The average of the Jaccard similarity of the scholar v_a to other member scholars in the collaboration is considered as another feature of the scholar v_a and is computed as below.

$$Jaccard\_sim(v_a) = \frac{\sum_{k=1, v_{a_k} \neq v_a}^{|V_a|} Jaccard\_sim(v_a, v_{a_k})}{|V_a|} \tag{15}$$

where $Jaccard\_sim(v_a, v_{a_k})$ is the Jaccard similarity of the scholar v_a to the scholar v_{a_k} and is computed as below (Liben-Nowell and Kleinberg 2007).

$$Jaccard\_sim(v_a, v_b) = \frac{\Gamma(v_a) \cap \Gamma(v_b)}{\Gamma(v_a) \cup \Gamma(v_b)} \tag{16}$$

where $\Gamma(v_a)$ and $\Gamma(v_b)$ are the set of the scholars who are the neighbor (the collaborator) of v_a and v_b respectively.

3. Structural centrality of the members

   (a) *Degree centrality* is a simple centrality measure that computes the number of direct neighbors of the individuals in a network (Freeman 1978) regardless of who the neighbors are. The degree centrality of the scholar v_a is computed as below.

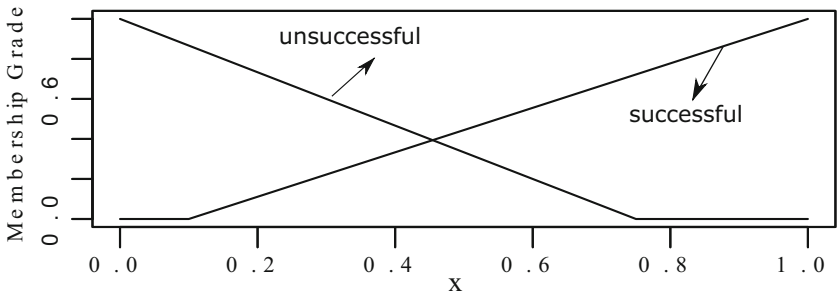

Fig. 7 Fuzzy membership functions for mapping the articles to successful and unsuccessful space based on their normalized citation counts

$$Degree(v_a) = |\Gamma(v_a)| \tag{17}$$

where $\Gamma(v_a)$ is the set of the scholars who are the neighbor (the collaborator) of the scholar v_a.

(b) *Scholar's rank* the rank of the scholar v_a ($rank(v_a)$) in the hypergraph is computed while a query vector with equal initial values is formed for all the scholars and concepts of the hypergraph $G_{T-10,T-1}$ (applying the ranking algorithm explained in "Background" section). In this feature not only the number of neighbors of a scholar but also who these neighbors are affect the final scores of the scholars.

After computation of the above features for the members, aggregate functions including average and maximum are applied to compute the feature value of the collaborations based on the feature values of their members. These features are listed in Table 4.

As explained before, the most common approach for computing the features of the collaborations is to apply the aggregate functions like average and maximum, but when

Table 4 Collaboration features used in the predictive model

$$avg\_hindex(e_T) = \frac{\sum_{v_a \in V_a} h\_index(v_a)}{|V_a|} \tag{18}$$

$$max\_hindex(e_T) = max(h\_index(v_{a_1}), ..., h\_index(v_{a_{|V_a|}})), \quad v_{a_i} \in V_a \tag{19}$$

$$avg\_gindex(e_T) = \frac{\sum_{v_a \in V_a} g\_index(v_a)}{|V_a|} \tag{20}$$

$$max\_gindex(e_T) = max(g\_index(v_{a_1}), ..., g\_index(v_{a_{|V_a|}})), \quad v_{a_i} \in V_a \tag{21}$$

$$avg\_citationAvg(e_T) = \frac{\sum_{v_a \in V_a} citation\_avg(v_a)}{|V_a|} \tag{22}$$

$$max\_citationAvg(e_T) = max(citation\_avg(v_{a_1}), ..., citation\_avg(v_{a_{|V_a|}}), v_{a_i} \in V_a \tag{23}$$

$$avg\_sim(e_T) = \frac{\sum_{v_c \in V_c} max(Sim_{v_c}(v_{a_1}), Sim_{v_c}(v_{a_2}), ..., Sim_{v_c}(v_{a_{|V_a|}}))}{|V_c|}, \quad v_{a_i} \in V_a \tag{24}$$

$$max\_sim(e_T) = max(Sim_{v_{c_1}}(v_{a_1}), Sim_{v_{c_1}}(v_{a_2}), ..., Sim_{v_{c_{|V_c|}}}(v_{a_{|V_a|}})), v_{a_i} \in V_a, \quad v_{c_j} \in V_c \tag{25}$$

$$avg\_FuzSuccess(e_t) = \frac{\sum_{v_a \in V_a} Fuz\_success(v_a)}{|V_a|} \tag{26}$$

$$max\_FuzSuccess(e_T) = max(Fuz\_success(v_{a_1}), ..., Fuz\_success(v_{a_{|V_a|}})), v_{a_i} \in V_a \tag{27}$$

$$avg\_FuzUnSuccess(e_T) = \frac{\sum_{v_a \in V_a} Fuz\_unsuccess(v_a)}{|V_a|} \tag{28}$$

$$max\_FuzUnSuccess(e_T) = max(Fuz\_unsuccess(v_{a_1}), ..., Fuz\_unsuccess(v_{a_{|V_a|}})), v_{a_i} \in V_a \tag{29}$$

$$avg\_JaccardSim(e_T) = \frac{\sum_{v_a \in V_a} Jaccard\_sim(v_a)}{|V_a|} \tag{30}$$

$$max\_JaccardSim(e_T) = max(Jaccard\_sim(v_{a_1}), ..., Jaccard\_sim(v_{a_{|V_a|}})), v_{a_i} \in V_a \tag{31}$$

$$avg\_Degree(e_T) = \frac{\sum_{v_a \in V_a} Degree(v_a)}{|V_a|} \tag{32}$$

$$max\_Degree(e_T) = max(Degree(v_{a_1}), ..., Degree(v_{a_{|V_a|}})), \quad v_{a_i} \in V_a \tag{33}$$

$$avg\_rank(e_T) = \frac{\sum_{v_a \in V_a} rank(v_a)}{|V_a|} \tag{34}$$

$$max\_rank(e_T) = max(rank(v_{a_1}), ..., rank(v_{a_{|V_a|}})), \quad v_{a_i} \in V_a \tag{35}$$

these are applied, some parts of the information on the feature values that exist in the collaborations are omitted. To preserve more information, a similar approach proposed by Torres-Carrasquillo et al. (2002) for converting variable length speech signal to a feature vector with fixed length is used. In this approach, first a clustering algorithm is applied to cluster all the scholars based on the value of the selected feature or features into k groups (clusters). Next, the frequency of different groups of scholars in each collaboration is determined. This approach is illustrated in Fig. 8 where the scholars are clustered into three groups in their corresponding circle. First, based on the feature value of the scholar members $(f_1, f_2, f_3, f_4, f_5, f_6)$, the most similar group to the members is determined. Next, the frequency of each group in the collaboration is computed. In this manner all collaborations will be mapped to feature vectors of size three. These feature vectors determine the number of times each group of the scholars appears in the collaboration.

For instance if the similarity of scholars to concepts is considered as the feature and k is set to be 3, cluster 1 can be considered as the scholars with low similarity, cluster 2 as the scholars with medium and cluster 3 as the scholars with high similarity. For computing the features of each collaboration based on the similarity to concepts, in addition to the average and maximum of the scholars' similarity, we count how many scholars with low, medium and high similarity exist in the collaboration and consider these frequencies as the feature value. When these features are combined with the average and maximum, preserve more information about the members of the collaboration.

Experiments

Exploring parameter setting

There are two parameters in the ranking algorithm, the iteration number mentioned in "Hypergraph" section and α in Eq. 3. For the iteration number, the algorithm is stopped on convergence. To explore the influence of α, for each value, the ranking algorithm is applied to compute the similarity of each scholar member to concepts. Next, a linear regression model is used for the prediction of the collaborations' citation counts based on the features

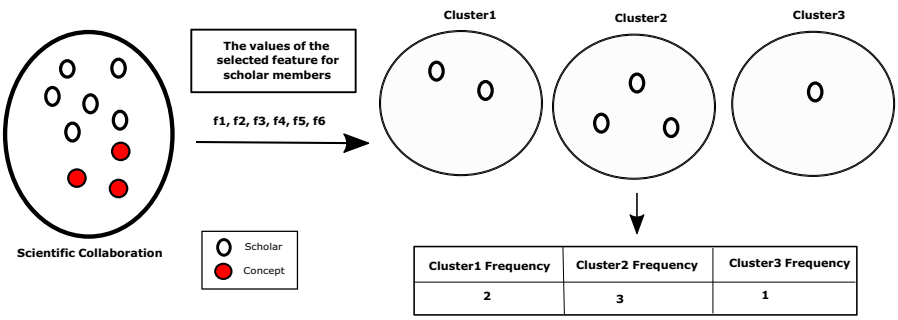

Fig. 8 Computing the feature vector for a collaboration based on the values of the selected feature $(f_1, f_2, f_3, f_4, f_5, f_6)$ using KMeans with 3 clusters

avg_sim and *max_sim* (explained in "Feature extraction" section). Collaborations of the year 2002 are used for setting the parameter and the correlation coefficient of the linear model is used as the evaluation metric. Figure 9 shows the performance measured as a function of α. The ranking algorithm obtains the best result when α is between 0.8 and 0.99.

For $\alpha = 1$, the correlation coefficient is 11.14 % which drops dramatically because the value of 1 means that there isn't any relationship between the ranking result and the query vector. This value is not shown in Fig. 9 to illustrate the differences between the other correlation values more clearly.

Experimental results

In the first set of experiments, we examine the correlation of the features resulted from the ranking algorithm (explained in "Background" section) with the citation counts of the collaborations. For each collaboration (e_T), features including *max_sim*, *avg_sim*, *max_rank* and *avg_rank* are extracted. For computing these features, α is set to be 0.9 and the algorithm is stopped on convergence. Next, the Spearman Correlation test is applied to measure the correlations. This correlation measure is selected because the real values of citation counts are not important and only the collaborations that are more successful in relation to other collaborations should receive higher ranking scores. The correlations are shown in Fig. 10 in chart bars. In all cases, the p-value is less than 2.2e-16, indicating that these correlations are statistically significant. These correlation values show that the extracted features will be helpful in discriminating between successful and unsuccessful collaborations.

In the second experiment, our goal is to show that making a joint hypergraph (scholar-concept-hypergraph) for capturing both scholars and concepts relations is a better modeling

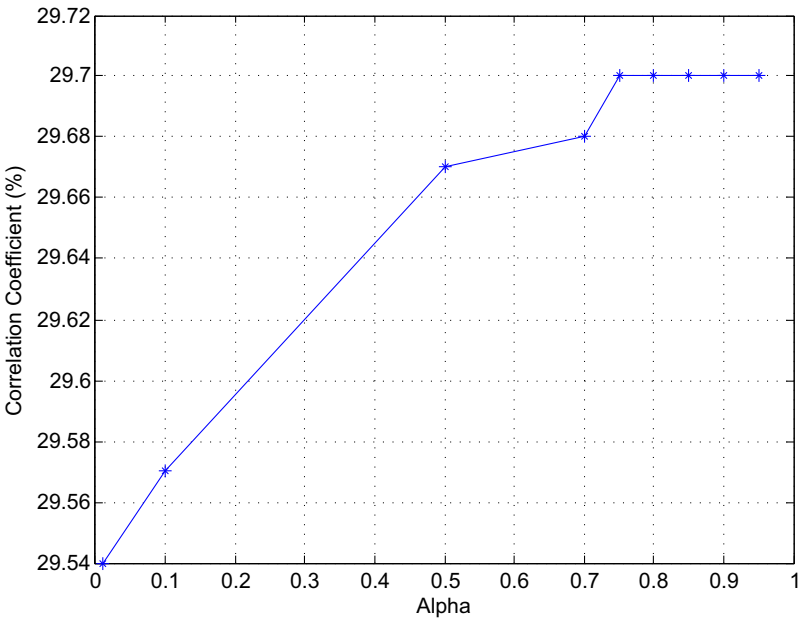

Fig. 9 Exploring the influence of the parameter. The correlation coefficient of the linear model is used as the evaluation metric

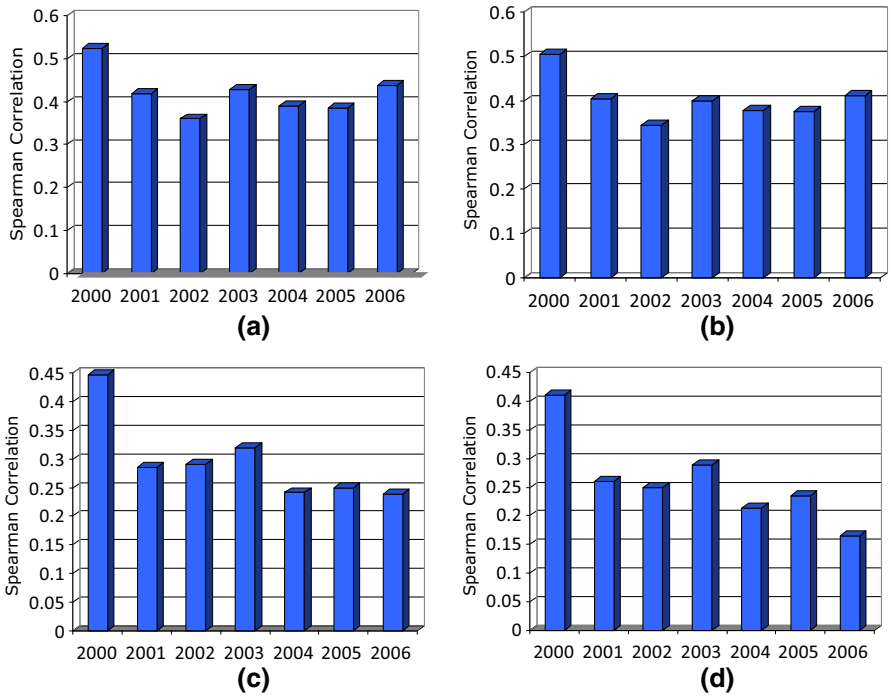

Fig. 10 **a** correlation between the *max_sim* with the citation counts of the collaborations; **b** correlation between the *avg_sim* with the citation counts of the collaborations; **c** correlation between the *max_rank* with the citation counts of the collaborations; **d** correlation between the *avg_rank* with the citation counts of the collaborations

approach (for score propagation) than a hypergraph that only includes the scholars and their social relations (scholar-hypergraph). In the second modeling approach, scores propagate just through the social relations among the scholars while in the first one, the semantic relations among the concepts also affect on the final scores.

For each collaboration from the year 2000–2006, the *max_sim* and *avg_sim* are extracted while a scholar-concept-hypergraph is applied for modeling the collaboration relations. In the case of scholar-hypergraph, the same approach is used for computing the similarity of scholars to concepts (explained in "Feature Extraction") but the query vector is formed using below equation.

$$y_i^c = \begin{cases} A_{T-10,T-1}[i, index(v_c)], & \text{scholar}_i \text{ has atleast one publication on } v_c \\ 0, & \text{otherwise} \end{cases}, \quad 1 \le i \le n_s \tag{36}$$

where i is the ith element of the query vector which corresponds to the ith scholar, index(c) is the index of the desired concept, $A_{T-10,T-1}$ is the transition matrix of the hypergraph $G_{T-10,T-1}$ and T is the target year. So those scholars who has at least one publication on v_c (are directly connected to v_c) are assigned a non-zero initial score equals to the weight of the relation that connect them to v_c. Next, the ranking algorithm is applied to propagate the scores (similarity values of the scholar members to concepts) among the scholars through

the social relations. Finally the features of *max_sim* and *avg_sim* are computed based on the similarity values of the scholar members to concepts.

The Spearman correlation of these features with the citation counts of the collaborations are illustrated in Fig. 11. As observed, involvement of concept-concept relations in the score propagation process helps to extract better features. These results demonstrate the worth of the inclusion of the semantic relations among the concepts for modeling the collaboration relations.

In the third experiment, we examine if the features of *avg_sim* and *max_sim* which are derived from the scholar-concept-hypergraph are better than the features that are extracted just based on the direct connections between the scholars and concepts. So the question is that "should we consider some skill level for a scholar who doesn't have any publications on a concept but is in the neighborhood of scholars and/or concepts that are related to the concept?". To answer this question, we compare *avg_sim* and *max_sim* with two other features. The first feature computes the scholar's skill level for concept v_c based on the citation counts of the scholar's publications on the concept using below equation.

$$citation\_avg(v_a, v_c) = \frac{\sum_{k=1}^{n_{ca}} citation(p_k)}{n_{ca}} \tag{37}$$

where p_k is the kth article of the scholar v_a on the concept v_c and $citation(p_k)$ is the citation count of p_k and n_{ca} is the number of the total scholar's publications on the concept v_c.

For each scholar member of the collaborations, this feature is computed and the average and maximum across all the members of each collaboration are considered as the features of the collaborations.

The second feature is computed based on the entries of the transition matrix $A_{T-10,T-1}$ (Eq. 4). The score of the scholar v_a for the concept v_c is:

$$weight(v_a, v_c) = A_{T-10,T-1}[index(v_a), index(v_c)] \tag{38}$$

Where $index(v_a)$ is the index of the scholar v_a and $index(v_c)$ is the index of the concept v_c. The average and maximum of this feature across all the members of each collaboration are considered as the features of the collaborations.

The Spearman correlation of these features with the citation counts of the collaborations are shown in Fig. 12. As observed features which are extracted based on the score

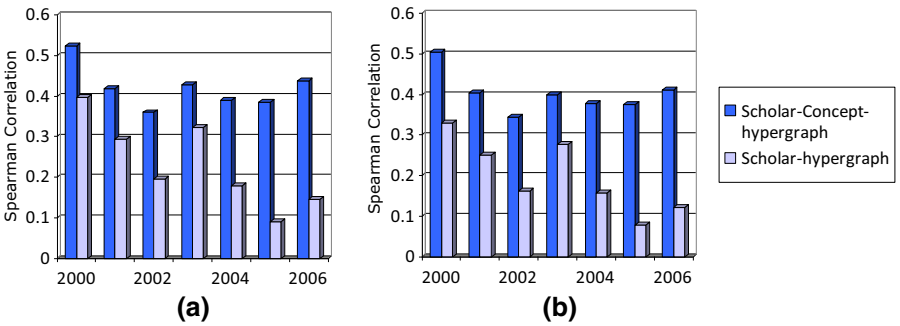

Fig. 11 **a** Comparison of two modeling approaches (concept-scholar-hypergraph and scholar- hypergraph) based on the *max_sim*. Spearman correlation is used as the evaluation metric. **b** Comparison of two modeling approaches (concept-scholar-hypergraph and scholar-hypergraph) based on the *avg_sim*. The Spearman correlation is used as the evaluation metric.

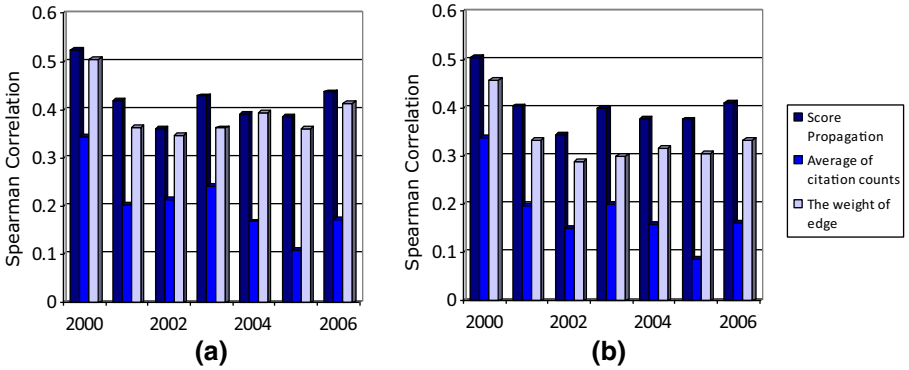

Fig. 12 **a** Comparison of the maximum (across all the members of each collaboration) of different features. **b** Comparison of the average (across all the members of each collaboration) of different features. The Spearman correlation is used as the evaluation metric

propagation are better. This shows that the skill level of a scholar for a concept should be determined in the light of both direct and indirect relations of the scholar to concept.

In the forth experiment, we examine the effect of extending the scholar-concept hypergraph to include venues where the articles are published. In this extended hypergraph (E-hypergraph), each hyperedge is composed of three types of vertices: scholar, concept and venue. Using this E-hypergraph as the modeling approach, we explore if this approach results to a better estimate of the scholars' expertise level (similarity to the concepts). The *max_sim* and *avg_sim* are considered as the features and correlation (Spearman) with the citation count of the collaborations as the evaluation metric. Our experiment shows that the correlation values don't change compare to the situation that scholar-concept hypergraph is used for modeling and adding the venues just increases the time complexity of the ranking algorithm.

Evaluation of the predictive model

In this experiment, classifiers are trained as the predictive models to discriminate between successful and unsuccessful collaborations. A collaboration in year T is considered successful, if the number of citations gained within a time frame of five years after publication is more than the median of citation counts of all collaborations in the same year. We use median because the distributions of citation counts are skewed and median would be a better measure for the center of the distributions.

To assure that our results are not subject to the type of the classifier, classification is done using different classifiers including Naive Bays, Multilayer Perceptron (MLP) and Random Forest. For training and evaluation of these classifiers, RWeka[8] is applied. Also, to avoid overfitting, we use 10 fold cross validation to set the parameters of the classifiers.

Since the cost of considering an unsuccessful collaboration as a successful one is more than that of the successful collaboration as unsuccessful one, we evaluate the classifiers based on $F_{0.5}$ measure that weight precision twice as much as recall and is computed using the below equation:

[8] http://cran.r-project.org/web/packages/RWeka/index.html.

Table 5 Result of evaluation of the classifier with different feature sets

	Precision (%)	Recall (%)	F0.5 (%)
Basic features	65.4	80.9	68.00
Features from hypergraph	64.2	71.5	60.61
Basic features + features from hypergraph	67.4	77.4	**69.19**
Basic features + K means similarity to concepts	66.2	80	68.56

Bold value indicates the best result

Table 6 Ranking the features based on their information gain

Feature name	Information gain
avg_sim & *max_sim*	0.086155
avg_FuzSuccess & *max_FuzSuccess*	0.066635
avg_JaccardSim & *max_JaccardSim*	0.04661
avg_rank & *max_rank*	0.01163
avg_hindex & *max_hindex*	0.010235
avg_gindex & *max_gindex*	0.009445
avg_citationAvg & *max_citationAvg*	0.006270
avg_degree & *max_degree*	0.004295

$$F_{0.5} = \frac{(1.25) * precision * recall}{0.25 * precision + recall} \tag{39}$$

The result of the evaluation of the best classifier (Random Forest) is shown in Table 5. First the classifiers are trained using all the features listed in Table 4 except the features extracted from the hypergraph (*avg_sim*, *max_sim*, *avg_rank* and *max_rank*). Next, each time a new feature set is added to the basic features and examined how it would affect the performance (For each feature set we report the best result among all the trained classifiers). As observed, the features extracted from hypergraph (*avg_sim*, *max_sim*, *avg_rank* and *max_rank*) are good enough to discriminate between collaborations for the defined threshold. Also adding them to the other features increases the classifier performance based on F0.5 measure.

To determine the worth of the features, we measure their information gain. The result is shown in Table 6. As observed, similarity of the scholars to the concepts has the highest information gain, therefore is the best feature. We choose this feature to cluster the scholars into three clusters (the number of clusters was determined experimentally) using KMeans algorithm. The obtained clusters are used to convert the values of this feature into collaboration features as explained in "Feature extraction" section. As observed in Table 5, adding KMeans concept similarity to the basic feature set contributes to the performance improvement.

We also examine the effect of the proposed features on the performance of two regression models for the prediction of the collaborations' citation counts. The correlation coefficients and root mean square errors are shown in Table 7 for different feature sets while linear regression and MLP[9] have been used as the predictive models. As observed, the proposed features, especially the features extracted from hypergraph, significantly

[9] Multilayer Perceptron.

Table 7 Result of evaluation of the regression models with different feature sets

	Linear model		MLP	
	Correlation coefficient (%)	Root mean squared error	Correlation coefficient (%)	Root mean squared error
Basic features	39.51	28.50	50	27.00
Basic features + features from hypergraph	45.54	27.62	61.19	24.62
Basic features + K means similarity to concepts	39.81	28.46	58.56	25.26
All features	45.62	27.60	63.95	24.00

improve the performance of both models. The best results are obtained when all features are used. Another point is that a non-linear model like MLP acts better than a linear model in capturing the relation between the collaboration success and the collaboration characteristics. This shows that for the task of research collaboration formation, optimization of linear combination of different collaboration characteristics is not a good choice.

Finally, the effect of adding a feature based on team diversity is explored. The h-index is considered as the expertise level of the scholars and the entropy-based index (Liang et al. 2007) is used as the measure of diversity between the h-index of the scholar members in each collaboration:

$$Entropy(e_T) = -\sum_{i=1}^{3} freq_i ln(freq_i) \tag{40}$$

where $freq_1, freq_2, freq_3$ represent the fraction of the scholar members with low, medium and high h-index in the collaboration respectively. The thresholds listed in Table 8 are used for quantization of the h-index values to these categories. These thresholds are determined based on the distribution of the scholars' h-index.

The result is illustrated in Table 9 which shows that adding this feature contributes to improve the classifier performance.

Discussion

While there are lots of works especially in the field of social science which examine the factors underlying effective research collaborations, but there is a big gap between these studies and electronic systems designed to help researchers find collaborators. Based on the research agenda and the requirement set for an effective collaborator discovery system proposed by Schleyer et al. (2012), there are many unanswered questions about how to best

Table 8 Thresholds for quantization of the h-index values

Low	$h\_index < 7.42$
Medium	$7.42 \leq h\_index < 14.84$
High	$h\_index \geq 14.84$

Table 9 Result of evaluation of the classifier after adding a feature based on team diversity

Precision	Recall	F0.5
67.6 %	78.7 %	69.56 %

use information technology to facilitate research collaborations. As the exploitation of collaboration network is essential in these systems (Schleyer et al. 2012), in this article an attempt is made to understand how to exploit the information embedded in collaboration networks. We benefit the link structure among the scholars and also among the scholars and the concepts to extract set of features that are correlated with the collaboration success. The experiments on a dataset collected from the Scival Expert, demonstrate that the new feature space improves the performance of the prediction. So these features would be valuable to incorporate in collaborator discovery systems while the citation count is considered as the success measure.

Future work

Research Networking Systems provide rich information about scholars and their research activities. VIVO and Scival Expert are popular systems with growing application by universities and institutions. In this article, we just considered the published articles as the research activity of the scholars and ignored other activities like grants, published books etc. that could contribute to a better estimation of the scholars' competence in a given concept. Also, in this article, we limit our dataset to the collaborations in the field of medicine. The effect of the proposed features can be studied for collaborations in other fields. Our next goal is to use the result of this study in the My Dream Team Assembler' project which is a tool to help form teams of experts.

References

Abbasi, A., Wigand, R. T., & Hossain, L. (2014). Measuring social capital through network analysis and its influence on individual performance. *Library & Information Science Research*, *36*(1), 66–73.

Awal, G. K., & Bharadwaj, K. (2014). Team formation in social networks based on collective intelligence-an evolutionary approach. *Applied Intelligence*, *41*(2), 627–648.

Bennett, L. M., & Gadlin, H. (2012). Collaboration and team science. *Journal of Investigative Medicine*, *60*(5), 768–775.

Börner, K., Contractor, N., Falk-Krzesinski, H.J., Fiore, S.M., Hall, K.L., Keyton, J., Spring, B., Stokols, D., Trochim, W., Uzzi, B. (2010). A multi-level systems perspective for the science of team science. Science Translational Medicine **2**(49), 49cm24–49cm24.

Bozeman, B., Fay, D., & Slade, C. P. (2013). Research collaboration in universities and academic entrepreneurship: The-state-of-the-art. *The Journal of Technology Transfer*, *38*(1), 1–67.

Callaham, M., Wears, R. L., & Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, *287*(21), 2847–2850.

Castillo, C., Donato, D., & Gionis, A. (2007) Estimating number of citations using author reputation. In: String processing and information retrieval (pp. 107–117). Berlin: Springer

Cummings, J. N., & Kiesler, S. (2008). Who collaborates successfully? Prior experience reduces collaboration barriers in distributed interdisciplinary research. In: Proceedings of the 2008 ACM conference on computer supported cooperative work (pp. 437–446). ACM

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064.

Dorn, C., & Dustdar, S. (2010). Composing near-optimal expert teams: A trade-off between skills and connectivity. *On the Move to Meaningful Internet Systems: OTM*, *2010*, 472–489.

Egghe, L. (2006). An improvement of the h-index: The g-index. *ISSI Newsletter*, *2*(1), 8–9.

Eslami, H., Ebadi, A., & Schiffauerova, A. (2013). Effect of collaboration network structure on knowledge creation and technological performance: The case of biotechnology in canada. *Scientometrics*, *97*(1), 99–119.

Fazel-Zarandi, M., & Fox, M. S. (2013). Inferring and validating skills and competencies over time. *Applied Ontology*, *8*(3), 131–177.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257–270.

Gajewar, A., & Sarma, A. D. (2012) Multi-skill collaborative teams based on densest subgraphs. In: SDM (pp. 165–176). SIAM.

Jirotka, M., Lee, C. P., & Olson, G. M. (2013). Supporting scientific collaboration: Methods, tools and concepts. *Computer Supported Cooperative Work (CSCW)*, *22*(4–6), 667–715.

Lappas, T., Liu, K., Terzi, E. (2009). Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 467–476). ACM.

Li, C. T., Shan, M. K., & Lin, S. D. (2015). On team formation with expertise query in collaborative social networks. *Knowledge and Information Systems*, *42*(2), 441–463.

Liang, T. P., Liu, C. C., Lin, T. M., & Lin, B. (2007). Effect of team diversity on software project performance. *Industrial Management & Data Systems*, *107*(5), 636–653.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031.

Olson, G. M., Zimmerman, A., & Bos, N. (2008). Scientific collaboration on the Internet. Cambridge, MA: The MIT Press.

PubMed: MS Windows NT kernel description (2005). http://www.ncbi.nlm.nih.gov/books/NBK3827.

Schleyer, T., Butler, B. S., Song, M., & Spallek, H. (2012). Conceptualizing and advancing research networking systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *19*(1), 2.

Schmidt, K., & Bannon, L. (2013). Constructing cscw: The first quarter century. *Computer Supported Cooperative Work (CSCW)*, *22*(4–6), 345–372.

Sharma, A., Srivastava, J., &Chandra, A. (2014). Predicting multi-actor collaborations using hypergraphs. arXiv preprint arXiv:1401.6404.

Skilton, P. (2008). Does the human capital of teams of natural science authors predict citation frequency? *Scientometrics*, *78*(3), 525–542.

Sonnenwald, D. H. (2007). Scientific collaboration: A synthesis of challenges and strategies. *Annual Review of Information Science and Technology*, *41*, 643–681.

Stokols, D., Misra, S., Moser, R. P., Hall, K. L., & Taylor, B. K. (2008). The ecology of team science: Understanding contextual influences on transdisciplinary collaboration. *American Journal of Preventive Medicine*, *35*(2), S96–S115.

Tan, S., Bu, J., Chen, C., & He, X. (2011). Using rich social media information for music recommendation via hypergraph model. In: Social media modeling and computing (pp. 213–237). New York: Springer.

Torres-Carrasquillo, P. A., Reynolds, D. A., & Deller Jr, J. (2002). Language identification using gaussian mixture model tokenization. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP), Vol. 1, pp. I–757.

Wang, M., Yu, G., An, S., & Yu, D. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics*, *93*(3), 635–644.

Whitfield, J. (2008). Collaboration: Group theory. *Nature*, *455*, 720–723.

Wi, H., Oh, S., Mun, J., & Jung, M. (2009). A team formation model based on knowledge and collaboration. *Expert Systems with Applications*, *36*(5), 9121–9134.

Yan, R., Huang, C., Tang, J., Zhang, Y., & Li, X. (2012). To better stand on the shoulder of giants. In: Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries (pp. 51–60). ACM.

Yu, T., Yu, G., Li, P. Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*(2), 1233–1252.