Ebrahimi *et al. J Big Data* (2021) 8:44
https://doi.org/10.1186/s40537-021-00432-y

Journal of Big Data

RESEARCH Open Access

Developing a mathematical model of the co-author recommender system using graph mining techniques and big data applications

Fezzeh Ebrahimi[1], Asefeh Asemi[1,2*], Amin Nezarat[3] and Andrea Ko[4]

*Correspondence:
asemi.asefeh@uni-corvinus.hu
[1] Department of Knowledge and Information Science, University of Isfahan, Isfahan, Iran
Full list of author information is available at the end of the article

Abstract
Finding the most suitable co-author is one of the most important ways to conduct effective research in scientific fields. Data science has contributed to achieving this possibility significantly. The present study aims at designing a mathematical model of co-author recommender system in bioinformatics using graph mining techniques and big data applications. The present study employed an applied-developmental research method and a mixed-methods research design. The research population consisted of all scientific products in bioinformatics in the PubMed database. To achieve the research objectives, the most appropriate effective features in choosing a co-author were selected, prioritized, and weighted by experts. Then, they were weighted using graph mining techniques and big data applications. Finally, the mathematical co-author recommender system model in bioinformatics was presented. Data analysis instruments included Expert Choice, Excel, Spark, Scala and Python programming languages in a big data server. The research was conducted in four steps: (1) identifying and prioritizing the criteria effective in choosing a co-author using AHP; (2) determining the correlation degree of articles based on the criteria obtained from step 1 using algorithms and big data applications; (3) developing a mathematical co-author recommender system model; and (4) evaluating the developed mathematical model. Findings showed that the journal titles and citations criteria have the highest weight while the abstract has the lowest weight in the mathematical co-author recommender system model. The accuracy of the proposed model was 72.26. It was concluded that using content-based features and expert opinions have high potentials in recommending the most appropriate co-authors. It is expected that the proposed co-author recommender system model can provide appropriate recommendations for choosing co-authors on various fields in different contexts of scientific information. The most important innovation of this model is the use of a combination of expert opinions and systemic weights, which can accelerate the finding of co-authors and consequently saving time and achieving a greater quality of scientific products.

Keywords: Big Data, Bioinformatics, Co-Author, Content-Based Recommender System, Data Science, Graph Theory, Mathematical Model

Introduction

Scientific collaboration in various fields has increased because of the growth in knowledge production and the increase in interdisciplinary knowledge. Some current scientific research requires the collaboration of hundreds of scientists with different specialties [1]. An increase in scientific collaboration has been a prominent feature of the evolution of science, at least since the beginning of the twentieth century [2–5]. These collaborations can be done at the intra-institutional, inter-institutional, domestic, and international levels. One of the researchers' concerns in choosing a co-author is to find individuals who can help them achieve the best and most appropriate scientific results. Identifying such individuals is researchers' one of the most critical issues that can lead to saving time, achieving more efficiency, and synergizing results. Achieving such a network requires a social network of authors whose members are as nodes and directionless edges represent two authors with a joint article [6]. One of these network types is the static social networks type of which the bibliographic information networks subtype is significant. An example of such a network is the PubMed[1] database. This information network consists of bibliographic data on medical science and information provided by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). Bioinformatics is an interdisciplinary field that combines elements of computer science, information technology, mathematics, statistics, and biotechnology, providing methods for extracting information and biological processes for knowledge discovery [7].

Due to the interdisciplinary nature and large volume of articles in bioinformatics, the need to use the methods and algorithms of recommender systems and link prediction methods is quite necessary for predicting and recommending co-authors. Further, graph theory is an important topic for information network analysis. Typically, in this method, the network dataset is represented as a graph in which the nodes within the network, the vertices of the graph, and the connections among the nodes form the graph links. One of the most important challenges in the field of graphs is the growing size of the data graph that sometimes involves millions of vertices and edges, making it so difficult to understand the graph that even many computer programs cannot analyze them. Therefore, big data tools are recommended for analyzing such large network graphs [8]. Some researchers have also used these tools to establish a recommender system. Dahdouh et al. used the parallel FP growth algorithm provided by the Spark Framework to implement the Hadoop ecosystem [47]. Nassar et al. have worked on establishing a proposed system based on deep learning. By combining deep neural network and matrix factorization, they have introduced a multi-criteria participatory filtering recommender [48]. Given the above, the need for a co-author recommender system to help researchers find their best potential research colleagues seems necessary. Given that no research has been presented to provide such a system in bioinformatics and the expert opinions and weights in this field have not been used in designing the recommender system, the present study attempts to present a co-author recommender system using link prediction algorithms, network analysis, big data tools, content-based systems, co-author

[1] https://pubmed.ncbi.nlm.nih.gov/.

recommender system, and expert opinions. Via graph theory, this system predicts and proposes a potential co-author for a researcher in bioinformatics. This model can be useful for other disciplines and databases. Also, it can help predict and recommend co-authors according to the content-based criteria.

In this study, first, the criteria effective in choosing co-authors were identified. Then, they were prioritized and weighted using AHP. Next, the criteria were weighted using algorithms and big data applications. After that, the weights obtained from both steps were integrated to obtain a mathematical co-author recommender system model. Finally, the proposed model was evaluated.

Research objectives

This study aimed to develop a mathematical co-author recommender system model in bioinformatics using graph mining techniques and big data applications. To achieve this main objective, the following secondary objectives were considered:

1. Identifying and prioritizing the criteria effective in choosing co-authors using AHP;
2. Determining the correlation degree of bioinformatics articles based on the "article titles", "abstracts", "keywords", "journal titles", and "institutional affiliation" criteria using graph data mining techniques and big data applications;
3. Presenting a mathematical co-author recommender system model in bioinformatics using graph mining techniques and big data applications; and.
4. Evaluating the proposed model using graph mining techniques and big data applications.

Methodology

The present study employed an applied-developmental research method and a mixed-methods (quantitative and qualitative) research design. The research was conducted in four stages. It is notable that we developed a mathematical model of the co-author recommender system using graph mining techniques or graph theory and big data applications. The matrix tables represented by graph theory for each question. Also, the tools and modules represented by using big data. Such as normalizer, pyspark.mllib.linalg.distributed.

Step 1: identifying and prioritizing the criteria effective in choosing a co-author using AHP

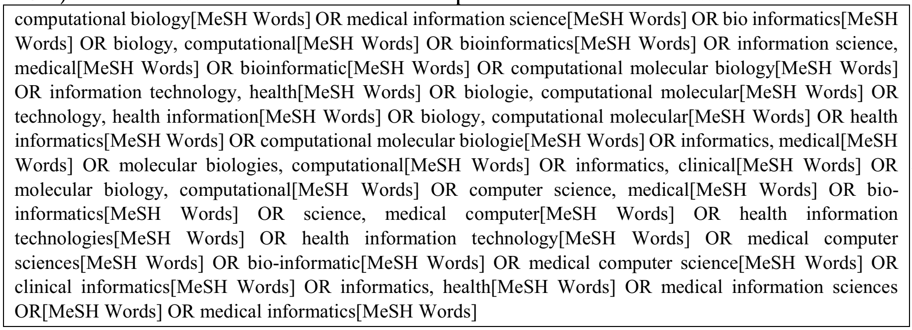
In this step, the criteria for choosing a co-author were identified by reviewing literatures related to the subject. The qualitative focus group method was used to identify and determine the validity of the weighting criteria questionnaire. Focus group research is a way of collecting qualitative data from an informal group discussion on a specific topic [9]. At this stage, the necessary calculations to determine the priority of each element of the decision were done using the data of pairwise comparison matrices. A preferential judgments questionnaire was designed with 30 questions and distributed among eight bioinformatics, biology, and scientometrics researchers and authors. Finally, the final weights of the criteria were obtained with an incompatibility rate of 0.8. The matrices were formed and analyzed via Expert Choice software.

Table 1 Characteristics of the system used for data processing

Specification	System tools
Intel® Xeon (R) CPU X5670 @2.93 GHZ	CPU
32 GB	RAM Size
24	#Core
Centos 6.9	Linux

Step 2: Determining the correlation degree of bioinformatics articles based on the criteria obtained from step 1 using algorithms and big data applications

This step was done via a quantitative approach so that the mathematical co-author recommender system model was implemented using prediction algorithms, text mining, and big data tools based on graph theory and using Python and Scala programming languages. In this section, all scientific productions in bioinformatics, including 699,160 articles published in the PubMed database until the time of reviewing them (January 2010- December 2019), were examined. This dataset was downloaded in XML format with a volume of 18 GB and used in research. The complete graph was plotted separately for the research criteria and the edge weight was calculated separately. At this stage, to retrieve all synonyms and related words in bioinformatics, the Medical Subject Headings (MeSH) database was searched. The search procedure was as follows:



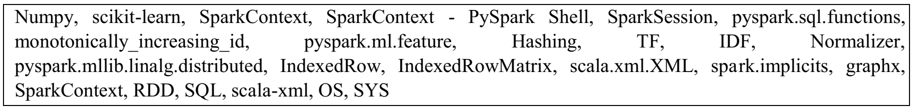
computational biology[MeSH Words] OR medical information science[MeSH Words] OR bio informatics[MeSH Words] OR biology, computational[MeSH Words] OR bioinformatics[MeSH Words] OR information science, medical[MeSH Words] OR bioinformatic[MeSH Words] OR computational molecular biology[MeSH Words] OR information technology, health[MeSH Words] OR biologie, computational molecular[MeSH Words] OR technology, health information[MeSH Words] OR biology, computational molecular[MeSH Words] OR health informatics[MeSH Words] OR computational molecular biologie[MeSH Words] OR informatics, medical[MeSH Words] OR molecular biologies, computational[MeSH Words] OR informatics, clinical[MeSH Words] OR molecular biology, computational[MeSH Words] OR computer science, medical[MeSH Words] OR bio-informatics[MeSH Words] OR science, medical computer[MeSH Words] OR health information technologies[MeSH Words] OR health information technology[MeSH Words] OR medical computer sciences[MeSH Words] OR bio-informatic[MeSH Words] OR medical computer science[MeSH Words] OR clinical informatics[MeSH Words] OR informatics, health[MeSH Words] OR medical information sciences OR[MeSH Words] OR medical informatics[MeSH Words]

.

It is notable that selected "keywords" in this research were extracted from Kiani et al. [10].

Step 3: Developing the mathematical co-author recommender system model

Python and Scala programming languages were used to implement the mathematical co-author recommender system model. Important modules and libraries used in this research are:



Numpy, scikit-learn, SparkContext, SparkContext - PySpark Shell, SparkSession, pyspark.sql.functions, monotonically_increasing_id, pyspark.ml.feature, Hashing, TF, IDF, Normalizer, pyspark.mllib.linalg.distributed, IndexedRow, IndexedRowMatrix, scala.xml.XML, spark.implicits, graphx, SparkContext, RDD, SQL, scala-xml, OS, SYS

Due to the large volume of data, it could not be analyzed by a usual system., so the researchers connected to the Astek big data server and all operations were performed on it. The specifications of the system used are given in Table 1.

Using a mixed-methods approach and the weights obtained from the expert opinions in the first stage as well as the weights obtained by the system in the second step, the final mathematical co-author recommender system model was presented.

Table 2 **train/test dataset percentage**

Dataset/Year	2010–2018	2019	No. of the articles
Training dataset		≃20%	137,035
Testing dataset	≃80%		562,125
Total			699,160

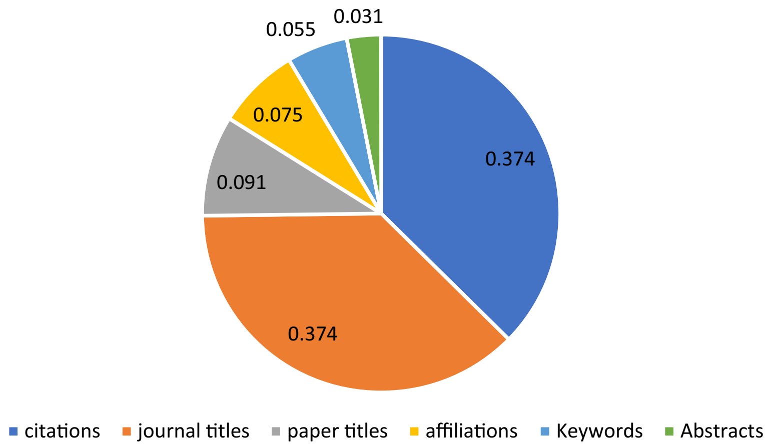


Fig. 1 The weights of the criteria using AHP

Step 4: Evaluating the mathematical co-author recommender system model

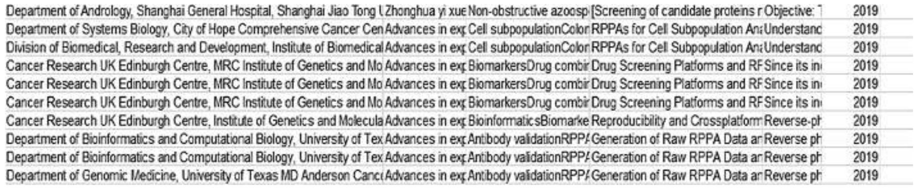
In the evaluation step of the proposed system, the 2019 articles were selected as a test, and the rest of the articles 2010-2018 as training (Table 2). According to the Pareto principle, many outcomes account for approximately 80% of the consequences of 20% of the causes [49]. Here, for predictive modeling, the train:test considered 80:20 percent. They then were evaluated using the accuracy criterion of the proposed system. It is notable that accuracy is only used to test. Also, as the amount of data in the dataset increases, the amount of data for testing decreases.

Findings

In this section, the findings obtained based on the research questions are expressed in order:

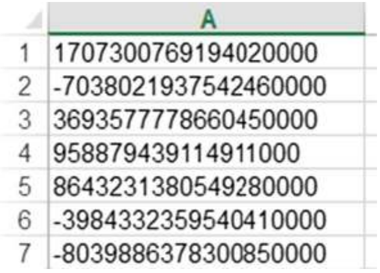
Identifying and prioritizing the effective criteria in choosing a co-author using AHP

At this step, six criteria including "journal titles, citations, article titles, organizational affiliation, keywords, and abstracts" were selected based on the expert opinions, the data available in the PubMed database, and feedback from the members of the focus group. The relevant questionnaire was designed Saaty's [11] 9-point scale and the pairwise comparisons matrix of expert opinions (eight experts) was calculated based on group AHP. The "journal titles, citations, article titles, organizational affiliation, keywords, and abstracts" criteria gained the priority weights of 0.374, 0.374, 0.091, 0.075, 0.055, and



Department of Andrology, Shanghai General Hospital, Shanghai Jiao Tong U	Zhonghua yi xue	Non-obstructive azoosp	[Screening of candidate proteins r	Objective: T	2019
Department of Systems Biology, City of Hope Comprehensive Cancer Cen	Advances in exp	Cell subpopulationColon	RPPAs for Cell Subpopulation Ana	Understand	2019
Division of Biomedical, Research and Development, Institute of Biomedical	Advances in exp	Cell subpopulationColon	RPPAs for Cell Subpopulation Ana	Understand	2019
Cancer Research UK Edinburgh Centre, MRC Institute of Genetics and Mo	Advances in exp	BiomarkersDrug combir	Drug Screening Platforms and RF	Since its in	2019
Cancer Research UK Edinburgh Centre, MRC Institute of Genetics and Mo	Advances in exp	BiomarkersDrug combir	Drug Screening Platforms and RF	Since its in	2019
Cancer Research UK Edinburgh Centre, MRC Institute of Genetics and Mo	Advances in exp	BiomarkersDrug combir	Drug Screening Platforms and RF	Since its in	2019
Cancer Research UK Edinburgh Centre, Institute of Genetics and Molecula	Advances in exp	BioinformaticsBiomarke	Reproducibility and Crossplatform	Reverse-ph	2019
Department of Bioinformatics and Computational Biology, University of Tex	Advances in exp	Antibody validationRPPA	Generation of Raw RPPA Data an	Reverse ph	2019
Department of Bioinformatics and Computational Biology, University of Tex	Advances in exp	Antibody validationRPPA	Generation of Raw RPPA Data an	Reverse ph	2019
Department of Genomic Medicine, University of Texas MD Anderson Canc	Advances in exp	Antibody validationRPPA	Generation of Raw RPPA Data an	Reverse ph	2019

Fig. 2 Extracting and saving the criteria



	A
1	1707300769194020000
2	-7038021937542460000
3	3693577778660450000
4	958879439114911000
5	8643231380549280000
6	-3984332359540410000
7	-8039886378300850000

Fig. 3 Converting keys to array indices via the hash function

0.031, respectively (Fig. 1). The citations criterion was removed during the implementation phase because this criterion was not available for all PubMed database articles and increased the calculation errors.

Calculating the correlation degree of bioinformatics articles with each other based on the "article titles", "abstracts", "keywords", "journal titles" and "organizational affiliation" criteria using graph extraction techniques and big data applications

At this stage, the datasets stored in the PubMed database were retrieved in XML format by using the PubMed tool. Then, they were distributed and parsed using the Spark. The PMID tags, authors name, organizational affiliation, article titles, keywords, abstracts, year of publication, and journal titles were saved in an Excel file. The Scala programming language has a library called scala-xml for analyzing XML documents via the original file was parsed and its tags extracted. The data output is illustrated in Fig. 2.

Then, to accelerate the searches and unify the authors' first and last names as keys and nodes, the keys are converted to array indices via the hash function (Fig. 3).

At this stage, using the Python programming language, the Spark was called. To obtain the weight of similarity between the criteria, first, a complete graph of all the authors was formed in pairs, in which the authors used as graph nodes, and edges between the two authors are an expression from the similarity weights of "article titles", "abstracts", "keywords", "journal titles", "organizational affiliation" (Fig. 4).

In this step, to calculate the weights, the features were first converted to words. To do this, CountVectorizer, provided by the scikit-learn library for vectorization of the sentences, was employed. CountVectorizer converted sentences to a set of tokens. Besides, it removes special punctuation and characters and applies prepositions to each word. In the next step, the text is converted to the attribute vector, and the incidence matrix is formed for each attribute. The word frequency vector or the incidence

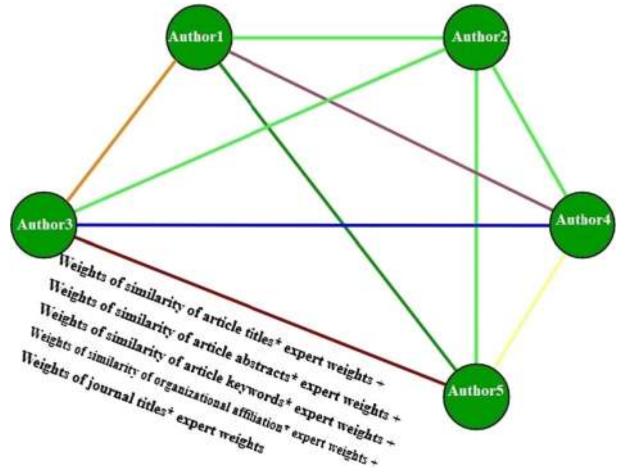
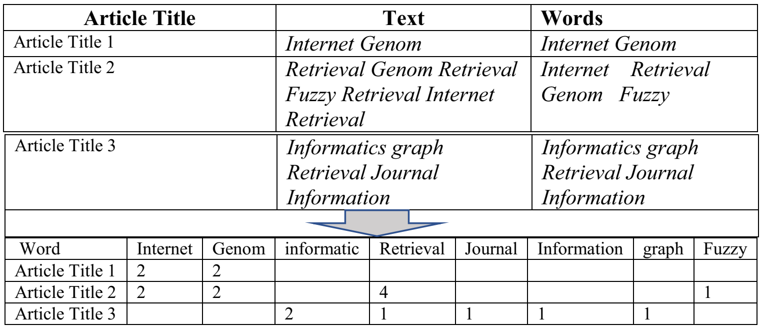

Author1
Author2
Author3
Author4
Author5
Weights of similarity of article titles* expert weights +
Weights of similarity of article abstracts* expert weights +
Weights of similarity of article keywords* expert weights +
Weights of similarity of organizational affiliation* expert weights +
Weights of journal titles* expert weights

Fig. 4 The complete graph among the criteria

Table 3 Incidence of words and matrices for the article titles criterion



Article Title	Text	Words
Article Title 1	*Internet Genom*	*Internet Genom*
Article Title 2	*Retrieval Genom Retrieval Fuzzy Retrieval Internet Retrieval*	*Internet Retrieval Genom Fuzzy*
Article Title 3	*Informatics graph Retrieval Journal Information*	*Informatics graph Retrieval Journal Information*

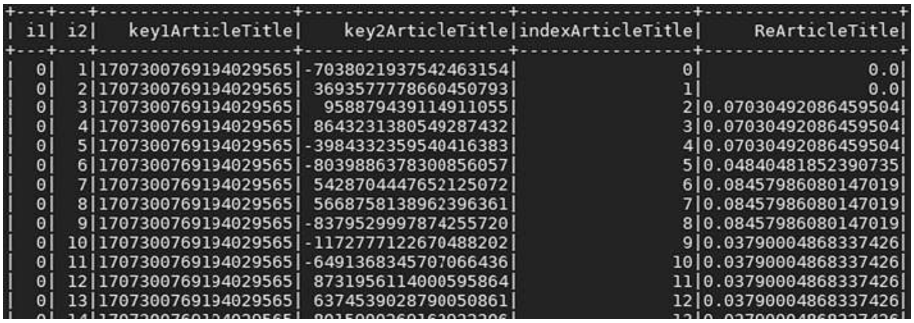
Word	Internet	Genom	informatic	Retrieval	Journal	Information	graph	Fuzzy
Article Title 1	2	2						
Article Title 2	2	2		4				1
Article Title 3			2	1	1	1	1	

Table 4 Cosine similarity of article titles

T3	T2	T1	
0	0.31	1	T1
0.41	1	0.31	T2
1	0.41	0	T3

of words is calculated for each attribute (Table 3). The distance between the two features is obtained based on cosine similarity. To do this, first, the words are used as vectors, for example, the vector of Article title 1 us formed as (2,1,0,0,0,0,0,0), and the vector of Article title 2 is formed as (1,1,0,4,0,0,0,1). Next, their cosine similarities are obtained in pairs (Table 4), i.e. the cosine similarity of Article title 1 with Article title 2; Article title 1 with Article title 2, and Article title 1 with Article title *n*. The numerical cosine similarity ranges from 0 to 1. If the two vectors (article titles) are the same, the cosine distance value is 1, and if the two vectors (article titles) are completely different, the cosine distance value is 0.

The cosine distance between the two Article titles 1 and 2 is as follows:



```
+---+---+-------------------+--------------------+----------------+-------------------+
| i1| i2|   key1ArticleTitle|    key2ArticleTitle|indexArticleTitle|      ReArticleTitle|
+---+---+-------------------+--------------------+----------------+-------------------+
|  0|  1|1707300769194029565|-7038021937542463154|               0|                0.0|
|  0|  2|1707300769194029565| 3693577778660450793|               1|                0.0|
|  0|  3|1707300769194029565|  958879439114911055|               2|0.07030492086459504|
|  0|  4|1707300769194029565| 8643231380549287432|               3|0.07030492086459504|
|  0|  5|1707300769194029565|-3984332359540416383|               4|0.07030492086459504|
|  0|  6|1707300769194029565|-8039886378300856057|               5|0.04840481852390735|
|  0|  7|1707300769194029565| 5428704447652125072|               6|0.08457986080147019|
|  0|  8|1707300769194029565| 5668758138962396361|               7|0.08457986080147019|
|  0|  9|1707300769194029565|-8379529997874255720|               8|0.08457986080147019|
|  0| 10|1707300769194029565|-1172777122670488202|               9|0.03790004868337426|
|  0| 11|1707300769194029565|-6491368345707066436|              10|0.03790004868337426|
|  0| 12|1707300769194029565| 8731956114000595864|              11|0.03790004868337426|
|  0| 13|1707300769194029565| 6374539028790050861|              12|0.03790004868337426|
|  0| 14|1707300769194029565| 8015000260162023206|              13|0.03790004868337426|
```

Fig. 5 Matrix of the criteria of the articles

Article Title Vector 1: (2, 1, 0, 0, 0, 0, 0, 0).

Article Title Vector 2: (1, 1, 0, 4, 0, 0, 0, 1).

$$\cos\theta = \frac{t1.t2}{|t1||t2|} = \frac{2\times1+1\times1+0\times0+0\times4+0\times0+0\times0+0\times0+0\times1}{\sqrt{2^2+1^2}\times\sqrt{1^2+1^2+4^2+1^2}}$$
$$= \frac{3}{\sqrt{5*19}} = 0.31$$

Formula 1 Cosine similarities.

The next process is the inverse document frequency (IDF) calculation, which is the normalization of the word frequency. The IDF calculation is based on the following formula 2:

$$IDFj = \log(N/|\{\mathrm{i} : tj \in di\}|$$

Formula 2. IDF calculation.

If a word is present in all documents, its Inverse Document Frequency (IDF) value is zero.

For example, if the number of article titles is 1,000,000 and the number of article titles containing the word “internet” is 1000, the IDF is obtained as follows:$IDF(\mathrm{internet}) = \log\,(1000000/1000) = 3$

In the next step, the TF-IDF is calculated. That is, for each word, the number of incidence of that word in a text is multiplied by the frequency of the document (the Formula 3).

$$wm, = freqm, i \times \log(N/nm)$$

Formula 3 TF-IDF calculation.

In the next step, the obtained weights are placed as the weights of the features of two authors’ articles. The output of the weights obtained is shown in Fig. 5.

Table 5 Evaluating the co-author recommender system model of bioinformatics

Number of each author's articles	Accuracy
>4	72.26
==3	No processability
==2	No processability

$Accuarcy = \frac{Correctly\ det\ ectedlinks}{alllinks}$

Presenting a mathematical co-author recommender system model in bioinformatics using graph mining techniques and big data applications

At this stage, the mathematical co-author recommender system model of bioinformatics products using graph mining techniques and big data applications is presented. To this end, a complete graph of the authors of all articles that were formed in the second stage, nodes of authors and edges between two authors, the similarity weights of the article titles, abstracts, keywords, journal titles, and organizational affiliation criteria were integrated with the weights obtained by the experts obtained in the first step. Then, the final weight between the two nodes was calculated. The final weight for predicting co-author was obtained via the formula below:

$$\boxed{\begin{array}{l}\text{Similarity nodes} = \text{weightArticleTitle} * 0.091 + \text{weightabstrac} * 0.031 + \text{weightkeyword} * 0.055 \\ + \text{weightaffiliation} * 0.075 + \text{weightTitleJournal} * 0.374\end{array}}$$

Formula 4 Mathematical Model of the Co-author Recommender System.

Evaluating the mathematical co-author recommender system model of bioinformatics using graph mining techniques and big data applications

For evaluating the model, the authors with more than 4 articles were filtered to obtain a database that is compatible with the hardware used for the present study. Then 2019 articles were selected for testing and the rest of the articles for training. The evaluation criterion is accuracy. The accuracy of the co-author recommender system in bioinformatics was 72.26 (Table 5).

If the authors with more articles are filtered, higher accuracy can be achieved. The more articles authors have in the database, the more knowledge they inject into the model, and the more accurate the model becomes.

By filtering the authors with five articles, an accuracy of 79, more than 10 articles, an accuracy of 98, and more than 15, an accuracy of 100 were achieved (Fig. 6).

Discussion and conclusion

The increasing speed of scientific production has become a serious challenge for researchers. Computer algorithms with access to high knowledge can make a great contribution to the spread of science [12]. Many previous studies [13–16] have examined the authors' collaboration using Social Network Analysis (SNA). Some researchers have used semantic networks to predict research trends [12] while others have predicted scientific productions in databases [17–19]. However, no research has been found to predict and co-author recommender systems in bioinformatics using systems methods and expert opinions.

One of the important issues for researchers is to use the researchers' opinions and ideas to select a scientific partner. According to experts in the focus group, identifying

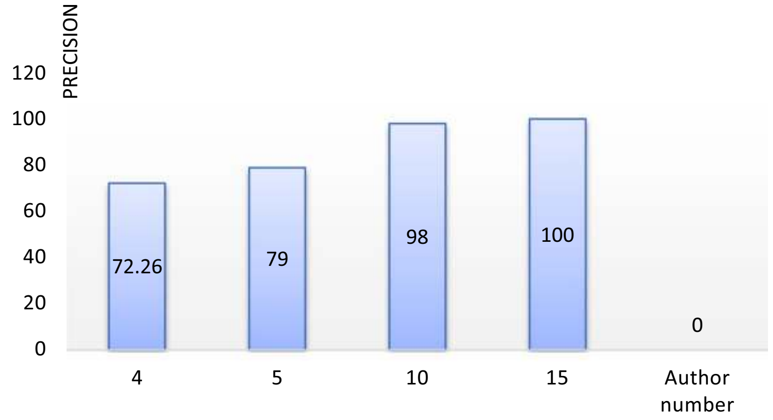


Fig. 6 Investigating the number of authors and accuracy

core authors are not the only factor in choosing a co-author. The authors think that core authors have no desire to co-author with them at all, or that university researchers are reluctant to work with a colleague at the same university [20]. The success of scientific articles depends on other factors such as social impact [21]. As many criteria involved in choosing co-authors, the expert opinions are critical.

Further, many criteria are involved in choosing a co-author. In this research, to present the mathematical co-author recommender system model, first, using the AHP technique, effective criteria in choosing the co-author were identified and prioritized. The criteria such as "thematic phrases in article titles", "thematic phrases in abstracts", "thematic phrases in the keywords assigned to articles", "similarity of authors' organizational affiliations", "similarity in citations", and "publication of articles in specialized journals" were considered suitable for weighting according to the expert opinions. Then the desired properties were extracted from the PubMed database in the XML format and saved in an Excel file. Due to the large volume of data, the Spark processing engine, which is scalable and distributable, was employed. Then the complete graph of the authors was drawn and the weight of the features was calculated as the edge. Finally, the mathematical model was designed and evaluated for a co-author recommender system. The results show that according to experts, the similarity in journal titles and joint citations has the highest score and the criterion of abstracts is less important than other criteria. Cabanac [22] considered the content of journals as the most important factor for scientific recommender systems. He also believed that a way for researchers to continuously review scientific texts about their field, as well as scientific texts in related fields, is to read related journal articles and conference papers. It can be inferred that the authors choose joint articles based on their specialty and subject matters. For example, two authors specializing in the genome publish their articles in a genome-related journal. Using the journal titles criterion can fill the gap in documenting the authors' names and journal titles because one of the uses of journal titles is to document the authors' names, as mainly specific authors select related journals based on their specific and thematic expertise.

Cota et al. [23] debugged the authors' names with the hypothesis that authors are more inclined to publish articles on the same subject and place of publication using similar functions. His results showed that it was 12% more accurate than the supervised and unsupervised methods. Han et al. [24] used the similarity measurement of the authors 'names and the article keywords using the probability model to remove the ambiguity of the authors' names.

Regarding the article titles criterion, reviewing articles shows that many authors used TF-IDF [25–29]. Beel et al. [30] stated that about 70% of the weighting methods used the TF-IDF approach. The cosine similarity criterion was also used more than other criteria. Comparing similarity criteria in recommender systems, Magara, Ojo, and Zuva [31] conclude that cosine similarity has the best performance compared to other similarity criteria. Salton and Buckley and Rathipriya, Thiyagarajann, and Thangavel [27, 32] considered the cosine similarity criterion to be superior to the Hamming similarity criterion for the design of the web recommender systems. Hasheminejad, Motieeyan, and Nasiri [33] showed that cosine similarity and Manhattan similarity calculation have better results from Euclidean distance.

One of the reasons for the popularity of the cosine distance is that it is very suitable for evaluation, especially for scattered vectors [34]. Kamyar [35] stated that the cosine distance is one of the best similarity algorithms that has higher accuracy than the Jaccard index and the Levenshtein distance. TF-IDF and cosine similarity are used to find similarities in the article titles. The title of each article is a manifestation of the identity of that article. It is the first manifestation of the text that the reader encounters. The article titles are containers whose objects are the main ideas of the texts.

In humanities research, because some article titles are expressed metaphorically, the degree of adaptation is less, but in bioinformatics, the specificity of the article titles is important. Davarpanah [36] examined the degree of compatibility of Persian article titles with their content in different scientific groups and observed that the article titles in humanities, compared to medical article titles, have less compatibility with the content. In this study, the article titles criterion gained more weight from expert opinions than organizational affiliation, keywords, and abstracts criteria. Nascimento et al. [37] considered the word weights in the article titles to be three times higher than the word weights in the body of texts. Mooney and Roy and Li et al. [17, 28] used the work titles criterion to design a book and articles recommender system. Achary [38] employed the article titles criterion in the content-based part of his recommender system.

Keywords are concepts defining the content of articles. Keyword search not only retrieves relevant documents but also documents that cannot be retrieved by subject search Aanonson & Ghareh Chamani [39, 40] used keywords of articles to recommend articles. Only using this criterion he designed an article recommender system. Mooney and Roy [17] designed a recommender system based on thematic wordinology, which is based on Bayesian algorithms to recommend books to Amazon book customers and buyers. Achary [38] used keyword tags in his recommendation system. Krenn & Zeilinger [12] developed a method for building a semantic network using keywords called SEMNET. They used SEMNET to predict future research trends and to inspire personal and surprising ideas in science. Sun et al. [41] employed the subject matter criterion to predict the co-authorship in heterogeneous bibliographic networks in the DBLP collaboration network. Using the content-based method and

TF-IDF algorithm, Chirita et al. [29] developed a recommender system for keywords on web pages by extracting important keywords from them.

The abstracts criterion is important because they provide comprehensive research information. An abstract contains the most important content and focus of a scientific article that authors tend to write with great effort. Metadata such as article titles, author names, year of publication, and place of publication are common, retrievable, and accessible criteria in most databases for similarity, but retrieving abstracts is not easy in most databases.

Cabanac [22] considered access to text and abstracts of scientific texts to be very costly and difficult to process. According to expert opinions, the similarity criterion was ranked fifth based on the abstracts criterion. When the articles of two authors are very similar in words of their abstracts, it shows that the two authors have a similar scientific trend in a subject area. This criterion was not used in previous research to predict and recommend a co-author. It seems that the massive and heavy processing of this part and also the lack of proper data are the factors that the researchers have omitted.

Organizational affiliation is one of the most important criteria for choosing a co-author. One of the biggest concerns of a personal researcher is to locate potential colleagues whose expertise complements his or her best skills [42]. Some researchers select people who belong to their organization and in their region to create and organize their research team. Some researchers prefer to work with researchers from outside their organizations. Departments, laboratories, schools, and all academic institutions create constraints for researchers due to competition with other institutions. The most important reason for this competition is the financial support of the government and officials [43]. Makarov, Bulanov, and Zhukov [20] found that researchers at the Higher School of Economics (HSE) University often collaborated with researchers from other universities. Organizational affiliation is an important criterion that researchers used in altmetrics and bibliometrics [13, 44–46].

Although evaluating this model is at an acceptable level according to filtering four authors and more than 72.26, no comparison was made because there was no research like. It is assumed that if there were no server restrictions and less than four authors did not filter articles, accuracy would be much higher.

Sufficient and efficient information is the basis of any decision making, thinking, and communication. The vast and growing number of publications in all fields of knowledge is incomprehensible to a human researcher. As a result, researchers must specialize in narrow disciplines that challenge the discovery of scientific connections beyond their field of research. Thus, access to structured knowledge from a large collection of journals can help push the boundaries of science [12]. Due to the increasing volume and growth of scientific articles, finding desired co-authors and co-researchers is a very difficult task and one of the main concerns of researchers. In this research, an attempt was made to provide a mathematical model that uses quantitative and systems methods in the macro data environment, as well as the expert opinions to recommend the best and most relevant potential collaborator for an author. The results of this study show that the usability of content-based methods in recommender systems in static bibliographic networks to find co-authors has high efficiency in related retrieval. Content-based methods include using different sections of article contents such as titles, abstracts, and keywords to present relevant articles due to their similarity to input articles. One of the practical and executive achievements of this project is to accelerate the retrieval of relevant authors and consequently saving time, achieving more efficiency, synergizing results, and obtaining a higher quality of research works and scientific development if this system leads to a more appropriate and easier selection of co-authors. Due to the interdisciplinary nature of bioinformatics, researchers in this field cannot specialize in all its narrow subdisciplines. So, it is necessary to find co-authors in this field. In this research, in addition to systems methods, bioinformatics experts' opinions were used to find a suitable co-author recommender system model. This model aligns the needs and concerns of the authors for the most similar co-author with their information demands and ensures the co-author recommendation more reliably. For future research, a comparison of the co-author prediction model based on behavioral characteristics and the co-author prediction and recommendation model based on fuzzy algorithms can be done. Besides, presenting a model in other bibliographic networks such as Web of Science (WoS), comparing the co-author prediction and recommendation model without considering the expert weights, presenting and plotting the co-author recommender system model with other algorithms such as the Jaccard index; Euclidean distance; simple Bayesian inference; and neural networks, and creating and designing a co-author recommender application or website are recommended.

Abbreviations

NLM: National Library of Medicine; NCBI: National Center for Biotechnology Information; MESH: Medical Subject Headings; IDF: Inverse Document Frequency; HSE: Higher School of Economics of National Research University.

Acknowledgements

Project no. TKP2020-NKA-02 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme.

Authors' contributions

EF: Data collection and Data analysis. AA: Supervising and Editing NA: Consulting, KA: Editing. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The relevant data was downloaded from the site https://www.ncbi.nlm.nih.gov/pubmed/in. January 2020.

Ethics approval and consent to participate

The authors Ethics approval and consent to participate.

Consent for publication

The authors consent for publication.

Competing Interests

Not applicable.

Author details

[1] Department of Knowledge and Information Science, University of Isfahan, Isfahan, Iran. [2] Doctoral School of Business Informatics, Corvinus University of Budapest, Budapest, Hungary. [3] Institute of Computer Science, University of Masaryk, Brno, Czech Republic. [4] Corvinus University of Budapest, Budapest, Hungary.

Received: 15 November 2020 Accepted: 22 February 2021
Published online: 06 March 2021

References

1. Boyer-Kassem T, Mayo-Wilson C, Weisberg M. In scientific collaboration and collective knowledge: new essays. Oxford: Oxford University Press; 2017.

2. Beaver D, Rosen R. Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship. Scientometrics. 1978;1:65–84. https://doi.org/10.1007/bf02016840.
3. Price DJ. Little science, big science. New York: Columbia University Press; 1963.
4. Wagner-Dobler R. Continuity and discontinuity of collaboration behavior since 1800- from a bibliometric point of view. Scientometrics. 2001;52:503–17. https://doi.org/10.1023/A:1014208219788.
5. Heydari M, Safavi Z. The survey of collaborative coefficient of article authors in journal of research in medical sciences. Research Med. 2012;36(2):109–13 **[in Persian]**.
6. Das K, Samanta S, Pal M. Study on centrality measures in social networks: a survey. Soc NetW. 2018. https://doi.org/10.1007/s13278-018-0493-2.
7. Ranganathan S, Gribskov M, Nakai K, Schönbach C. Encyclopedia of bioinformatics and computational biology. Amsterdam: Elsevier; 2019.
8. Chaoji M, AlHasan M. ORIGAMI: a novel and effective approach for mining representative orthogonal graph patterns. Statistical Analysis Data Mining. 2008;1:67–84. https://doi.org/10.1002/sam.10004.
9. Wilkinson S, Silverman D. Focus group research. Qualitative research: Theory, method, and practice.2004; 177-199.
10. Kiani M. Information ecology in field of bioinformatics with emphasis on thematic relationships. [dissertation]. [Isfahan]. Isfahan university; 2020. [in Persian].
11. Saaty TL. The Analytical Hierarchy Process. New York: McGraw-Hill; 1980.
12. Krenn M, Zeilinger A. Predicting research trends with semantic and neural networks with an application in quantum physics. Proc Natl Acad Sci. 2020;117(4):1910–6. https://doi.org/10.1073/pnas.1914370116.
13. Ho T, Bui Q, Bui, M. Co-author Relationship Prediction in Bibliographic Network: A New Approach Using Geographic Factor and Latent Topic Information. SoICT. 2019: 69–77; https://doi-org.ezp.semantak.com/10.1145/3368926.3369668.
14. Cho H, Yu Y. Link prediction for interdisciplinary collaboration via co-authorship network. 2018; 25. https://doi.org/10.1007/s13278-018-0501-6.
15. Sadoughi F, Valinejadi A, Shirazi M S, khademi R. Social Network Analysis of Iranian Researchers on Medical Parasitology: A 41 Year Co- Authorship Survey. Iran J Parasitol. 2016; 11(2): 204-212.
16. Chien S, Chien T, Chang Y, Shih F. Patterns of international coauthor collaboration in bioinformatics. Biomedical Res Netw. 2017;1(6):1783–5.
17. Mooney R J, Roy L. Content-based book recommending using learning for text categorization. in Proceedings of the fifth ACM conference on Digital libraries, 2000: 195–204; https://doi.org/10.1145/336597.336662.
18. Abu-Jbara A, Radev D. Coherent citation-based summarization of scientific papers. in Proceedings of the 49th Annual Meeting of the Association for 60 Computational Linguistics. Human Language Technologies. 2011; 1: 500–509.
19. Teufel S, Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics. 2002;28(4):409–45.
20. Makarov I, Bulanov O, Zhukov L. Co-author Recommender System. Paper presented at the Models: Algorithms, and Technologies for Network Analysis, Cham; 2017.
21. Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F. Predicting scientific success based on coauthorship networks. EPJ Data Science. 2014;3(1):1–16. https://doi.org/10.1140/epjds/s13688-014-0009-x.
22. Cabanac G. Accuracy of inter-researcher similarity measures based on topical and social clues. Scientometrics. 2011;87:597–620. https://doi.org/10.1007/s11192-011-0358-1.
23. Cota RG, Ferreira AA, Nascimento C, Gonçalves MA, Laender AHF. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. J Am Soc Inf Sci. 2010;61:1853–70. https://doi.org/10.1002/asi.21363.
24. Han H, Giles C L, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. In ACM/IEEE-CS.. 2004: 296-305.
25. Wu F, Mi L, Li X, Huang L, Tong Y. Identifying Potential Standard Essential Patents Based on Text Mining and Generative Topographic Mapping. 2018 IEEE International Symposium on Innovation and Entrepreneurship (TEMS-ISIE), Beijing.2018; pp. 1-9
26. Wang C, Satuluri V, Parthasarathy S. Local Probabilistic Models for Link Prediction. ICDM,2007; 322–331; https://doi.org/10.1109/ICDM.2007.108 .
27. Salton G, Buckley C. Term-weighting approaches in automatic text Retrieval. Inf Process Manage. 1988;24(5):513–23.
28. Li X, Chen Y, Pettit B, Rijke M. Personalised Reranking of Paper Recommendations Using Paper Content and User Behavior. ACM Trans. Inf. Syst. 2019; 37, 2019: 23; https://doi.org/10.1145/3312528.
29. Chirita P A, Costache S, Nejdl W, Handschuh S. P-TAG: large scale automatic generation of personalized annotation tags for the web. WWW '07. Proceedings of the 16th international conference on World Wide Web.. 2007: 845-854.
30. Beel J, Gipp B, Langer S, et al. Research-paper recommender systems: a literature survey. Int J Digit Libr. 2016;17:305–38. https://doi.org/10.1007/s00799-015-0156-0.
31. Magara M, Ojo S, Zuva T. A comparative analysis of text similarity measures and algorithms in research paper recommender systems. ICTAS. 2018: 1-5; https://doi.org/10.1109/ictas.8368766.
32. Rathipriya R,.Thiyagarajann,R, Thangavel K. Recommendation of Web Pages using Weighted K-Means Clustering, International Journal of Computer Applications. 2014; 44-48.
33. Hasheminejad M, Motieeyan Z, Nasiri J. Comparison of a recommender text system with three criteria for measuring cosine similarity, Euclidian distance and Manhattan. The 6th International Congress on Development and Promotion of Fundamental Science and Technolpgy in Society. 2019 [in Persian].
34. Farhadi M, JamZad M. Examining similarity criteria in content-based image retrieval. CSJ. 2018;9:13–27 **[in Persian]**.
35. Kamyar M. Automatic extraction of concepts from text based on linguistic methods. [dissertation].[Mashhad]. ferdowsi university of mashhad; 2014 [in Persian].
36. Davarpanah M. Investigating the compatibility of Persian article titles with their content. IRANDOC. 1996;12(2):1–12 **[in Persian]**.

37. Nascimento C, Laender A, Silva A S, Gonçalves M A. A source independent framework for research paper recommendation. ACM/IEEE, 2011: 297–306.
38. Achary R. An author recommendation system using both content-based and collaborative filtering methods [dissertation]. [California]: California state university; 2011.
39. Aanonson J. Precision and Recall in Title keyword searchers. Information technology and libraries. 1987;14(3):162–70.
40. Ghare-Chamani J. Provide a way to suggest referrals in the referral network. [dissertation]. [Tehran]: Sharif University of Technology; 2013 [in Persian].
41. Sun Y, Barber R, Gupta M, Aggarwal CC, Han J. Co-author relationship prediction in heterogeneous bibliographic networks. ASONAM. 2011: 121–128. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5992571 .
42. Yu Q, Long C, Lv Y, Shao H, He P, Duan Z. Predicting Co-Author Relationship in Medical Co-Authorship Networks. PLOS ONE.2014; 9(7); https://doi.org/10.1371/journal.pone.0101214.
43. Roemer R, Borchardt R. Meaningful Metrics: A 21st Century Librarian's Guide to Bibliometrics, Altmetrics and Research Impact. USA: ACRL; 2015.
44. Yan E, Guns R. Predicting and recommending collaborations: an author, institution, and country-level analysis. J Informetrics. 2014;8:295–309. https://doi.org/10.1016/j.joi.2014.01.008.
45. Brandão M, Moro. Affiliation Influence on Recommendation in Academic Social Networks. Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management. 2012; 230-234.
46. Andrikopoulos A, Samitas A, Kostaris K. Four decades of the Journal of Econometrics: coauthorship patterns and networks. J Econometrics. 2016;195(1):23–32. https://doi.org/10.1016/j.jeconom.2016.04.018.
47. Dahdouh K, Dakkak A, Oughdir L, et al. Large-scale e-learning recommender system based on Spark and Hadoop. J Big Data. 2019;6:2. https://doi.org/10.1186/s40537-019-0169-4.
48. Nassar N, Jafar A, Rahhal Y. Multi-criteria collaborative filtering recommender by fusing deep neural network and matrix factorization. J Big Data. 2020;7:34. https://doi.org/10.1186/s40537-020-00309-6.
49. Bunkley, N. (2008, March 3). Joseph Juran, 103, Pioneer in Quality Control, Dies (Published 2008). The New York Times. https://www.nytimes.com/2008/03/03/business/03juran.html.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.