International Journal of Information Science and Management
Vol. 19, No. 2, 2021, 1-18
https://dorl.net/dor/20.1001.1.20088302.2021.19.2.1.9

Original Research

Developing a Prediction Model for Author Collaboration in Bioinformatics Research Using Graph Mining Techniques and Big Data Applications

Fezzeh Ebrahimi
Ph. D. Candidate, Department of Knowledge and Information Science, University of Isfahan, Iran
ebrahimif@yahoo.com
ORCID iD: https://orcid.org/0000-0002-6709-8327

Asefeh Asemi
Ph. D. Doctoral School of Business Informatics, Corvinus University of Budapest, Hungary
Corresponding Author:
asemi.asefeh@uni-corvinus.hu.ir
ORCID iD: https://orcid.org/000-0003-1667-4408

Ahmad Shabani
Professor, Department of Knowledge and Information Science,
University of Isfahan, Isfahan, Iran
shabania@edu.ui.ac.ir
ORCID iD: https://orcid.org/0000-0003-0466-6240

Amin Nezarat
Assistant Prof., Institute of Computer Science, University of Masaryk, Czech Republic
nezarat@ics.muni.cz
ORCID iD: https://orcid.org/0000-0002-2445-7254

Received: 10 October 2020
Accepted: 11 November 2020

Abstract

Nowadays, scientific collaboration has dramatically increased due to web-based technologies, advanced communication systems, and information and scientific databases. The present study aims to provide a predictive model for author collaborations in bioinformatics research output using graph mining techniques and big data applications. The study is applied-developmental research adopting a mixed-method approach, i.e., a mix of quantitative and qualitative measures. The research population consisted of all bioinformatics research documents indexed in PubMed (n=699160). The correlations of bioinformatics articles were examined in terms of weight and strength based on article sections including title, abstract, keywords, journal title, and author affiliation using graph mining techniques and big data applications. Eventually, the prediction model of author collaboration in bioinformatics research was developed using the abovementioned tools and expert-assigned weights. The calculations and data analysis were carried out using Expert Choice, Excel, Spark, and Scala, and Python programming languages in a big data server. Accordingly, the research was conducted in three phases: 1) identifying and weighting the factors contributing to authors' similarity measurement; 2) implementing co-authorship prediction model; and 3) integrating the first and second phases (i.e., integrating the weights obtained in the previous phases). The results showed that journal title, citation, article title, author affiliation, keywords, and abstract scored 0.374, 0.374, 0.091, 0.075, 0.055, and 0.031. Moreover, the journal title achieved the highest score in the model for the co-author recommender system. As the data in bibliometric information networks is static, it was proved remarkably effective to use content-based features for similarity measures. So that the recommender system can offer the most suitable collaboration suggestions. It is

expected that the model works efficiently in other databases and provides suitable recommendations for author collaborations in other subject areas. By integrating expert opinion and systemic weights, the model can help alleviate the current information overload and facilitate collaborator lookup by authors.

Keywords: Recommender System, Co-Author, Graph Theory, Network Analysis, Bibliographic Networks, Research Collaboration

Introduction

The increasing boost of human knowledge has contributed to scientific collaborations. Scientific collaboration can occur in the compilation book, article translation, an article published in journals and presented in conferences, research projects, membership in scientific societies, and cooperation with scholarly journals (Ghanei Rad, 2006). Another example of scientific collaborations is faculty member collaborations in supervising, advising, and refereeing student theses (Tabarzeh, 2018). The collaborations may occur at intra-institutional, inter-institutional, domestic, and international levels. One of the researcher's concerns is to find potential collaborators who can best cooperate in a research project. One of the critical issues for researchers is to identify effective scientific collaborators in co-authorship networks. Identifying the best candidates for scientific collaboration helps save time, increase efficiency, boost research quality, and develop science.

A co-authorship network is a social network constituting a group of researchers. In a co-authorship network, authors function as nodes, while undirected edges represent two authors who have published a joint article (Das, Samanta & Pal, 2018). Static social networks such as bibliometric information networks are a type of social network. PubMed is an example of such a networks. PubMed is an information network constituting bibliometric data on medical sciences provided by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). Bioinformatics is an interdisciplinary field that includes methods and software for understanding biological information and that involves the interaction of computers, mathematics, statistics, physics, and biology sciences (Benton, 1996). Bioinformatics is a branch of biology that develops methods and software tools (i.e., algorithms and databases) to understand and effectively use Biological Data. It involves the storage, analysis, distribution, and retrieving of biological information. Paulien Hogeweg and Ben Hesper coined the term bioinformatics to describe “the study of informatic processes in biotic systems.” (Rose, 2020). Considering the importance of bioinformatics as an interdisciplinary field, researchers have produced and developed science in this field (Figure 1). With the increase in online biomedical articles, including bioinformatics articles in full-text format, it has become vital for most text mining software to understand and cite documents.

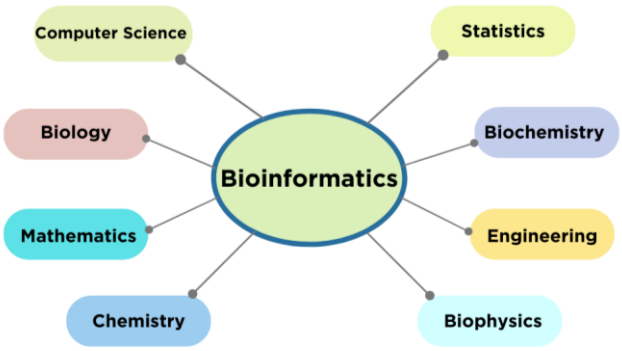

Figure 1. The fields related to bioinformatics (Rose, 2020)

One of the methods of predicting scientific collaborators is to use the procedures and algorithms of recommender systems and link prediction methods. Besides, graph theory is essential in analyzing information networks. In this method, the network data set is usually shown as graphs in which the nodes within the network constitute graph heads, and relations among nodes constitute graph links. One of the graphs' challenges is enlarging the data graph volume, including millions of nodes and edges. Such enormous volume makes it difficult to understand graphs. Even many computer programs may fail to analyze these graphs. Thus, big data tools should be used to analyze such graphs (Chaoji, Al Hasan, Salem, Besson & Zaki, 2008). Due to the rapid growth of scientific production, high volume of information, specialization, and interdisciplinarity of bioinformatics, Big data of PubMed database, the importance of time for researchers, and items like the lack of a recommender system in the field of bioinformatics articles, the lack of a recommending system using expert opinions, Not using essential components such as abstracts, etc. The need for a recommender system that helps researchers find their best potential co-author and scientific partner seems necessary.

According to this study, no article is done to provide a scientific collaborators system in bioinformatics, and research done is mainly based on collaborative filtering or content using a very low number of components and small data. On the other hand, in any of the researches, the opinions of experts are not included. Therefore, a study similar to this study was not found. In this regard, the present study aimed to map the complete graph of co-authorship network in PubMed using link prediction algorithms, network analysis, and big data tools, and content-based system and expert opinions method in order to design the recommender system that uses graph theory to predict the best potential collaborations for a researcher in the field of bioinformatics. Furthermore, the model developed in this study may be helpful in other databases to illustrate author collaborations in a given field by applying a given dataset. Therefore, the study develops a predictive model that can predict scientific collaborators based on content components.

Research objectives

The main objective of the present study is to provide an author collaboration prediction model in bioinformatics research using graph mining techniques and big data applications. To this end, the following specific goals are pursued:

Calculating the weights and correlations of bioinformatics research articles based on article titles using graph mining and big data applications

Calculating the weights and correlations of bioinformatics research articles based on

journal titles using graph mining and big data applications

Calculating the weights and correlations of bioinformatics research articles based on keywords using graph mining and big data applications

Calculating the weights and correlations of bioinformatics research articles based on abstracts using graph mining and big data applications

Calculating the weights and correlations of bioinformatics research articles based on affiliation using graph mining and big data applications

Developing an author collaboration prediction model for bioinformatics research articles using graph mining and big data applications

Methodology

The study is applied-developmental research adopting a mixed-method approach, i.e., a mix of quantitative and qualitative measures. The methodology entails three main phases: 1) identifying and weighting the components contributing to similarity measurement of authors; 2) implementing the co-authorship prediction model; and 3) integrating the weights obtained in the previous phases.

Step 1: the literature was reviewed to identify the criteria affecting the selection of scientific collaborators. The focus group method was used to determine the weighting questionnaire. Focus groups provide a method for collecting qualitative data through an informal group discussion on a specific subject (Wilkinson 2004). The components identified for prioritization using pairwise comparisons were rendered into 30 questionnaire items on a 9-point Saaty (1980) scale and an open-ended item. In this phase, the data were collected from scientometrics and bioinformatics experts, professors, and professionals. The data were analyzed using Expert Choice and Excel software. The face validity of the questionnaire was examined in the focus group by drawing on the opinions of eight experts in the fields of software, artificial intelligence, scientometrics, library and information science, and bioinformatics who were available and answered the questions. The reliability of the instrument showed an inconsistency rate of 0.8.

Subsequently, the research questionnaire was designed using the 9-point Saaty scale (1980). Eventually, the pairwise comparison matrix of expert opinions was calculated based on group AHP. The matrices involved six sub-criteria, which were rendered to 30 questionnaire items based on pairwise comparisons and formula n*(n-1).

Step 2: this step involved a quantitative approach in which the co-authorship prediction model was implemented using prediction algorithms, text mining, and big data tools based on graph theory in Python and Scala. All bioinformatics research output indexed in PubMed including 699160 articles was examined in the modelling phase on December 2019. The dataset, sized 18 GB, was downloaded from PubMed in XML format.Accordingly, the complete matrix of research variables was plotted per variable, and the edge weights were computed by each individual edge. At this stage, The Medical Subject Headings (MESH) database was searched to retrieve all the synonymous keywords and terms relevant to bioinformatics. The search query is as follows:

computational biology[MeSH Terms] OR medical information science[MeSH Terms] OR bio informatics[MeSH Terms] OR biology, computational[MeSH Terms] OR bioinformatics[MeSH Terms] OR information science, medical[MeSH Terms] OR bioinformatic[MeSH Terms] OR computational molecular biology[MeSH Terms] OR

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

The keywords were acquired from Kiani (2020). Python and Scala were used to implement the prediction model. The modules and libraries used in the research included:

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 enormous volume of data, it was impossible to do the processing on a PC. Thus, we connected to the ASTEC big data server to do the operations. The configurations of the system are illustrated in Table 1.

Table 1
System configurations

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

Step 3: for adopting a mixed-method approach, the final co-authorship prediction model was calculated using the expert weightings and system weightings in Step 2.

Results

Identifying and weighting the components contributing to similarity measurement of authors

Following an extensive literature review, some 79 criteria were identified to classify and weight the components. Then some six components, including common journals, citation, titles, affiliations, keywords, and abstract similarity, were selected based on expert opinion and PubMed data. The questionnaire was designed on a 9-point Saaty scale. The pairwise comparison matrix for the expert opinion was calculated based on group AHP (Table 2). During the implementation phase, the citation component was excluded from the calculation due to a

high rate of errors. This is because the citation counts were not available for all PubMed documents but only for open access articles.

Table 2
Matrix of expert opinions and final weight calculation

criteria	Weight achieved
citations	0.374
journal titles	0.374
paper titles	0.091
affiliations	0.075
Keywords	0.055
Abstracts	0.031

In the second phase of the research, the datasets saved in PubMed in XML format were recalled, and the data were distributed, parsed, and crawled using Spark. Subsequently, PMID, author name, affiliation, article title, keyword, abstract, publication year, and journal title tags were extracted. Scala contains a scala-XML library that is used to parse XML documents. For Using the scala-XML library, the raw file was parsed to extract the tags. Figure 2 illustrates the data output.

Department of Andrology, Shanghai General Hospital, Shanghai Jiao Tong U	Zhonghua yi xue	Non-obstructive azoosp	Screening of candidate proteins r	Objective:	2019
Department of Systems Biology, City of Hope Comprehensive Cancer Cen	Advances in exp	Cell subpopulationColor	RPPAs for Cell Subpopulation An	Understand	2019
Division of Biomedical, Research and Development, Institute of Biomedical	Advances in exp	Cell subpopulationColor	RPPAs for Cell Subpopulation An	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 ar	Reverse ph	2019
Department of Bioinformatics and Computational Biology, University of Tex	Advances in exp	Antibody validationRPPA	Generation of Raw RPPA Data ar	Reverse ph	2019
Department of Genomic Medicine, University of Texas MD Anderson Cance	Advances in exp	Antibody validationRPPA	Generation of Raw RPPA Data ar	Reverse ph	2019

Figure 2. Extraction of data frames

The key was defined as a hash to accelerate the searches and unify the authors' first and last names as keys and nodes (Figure 3).

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

Figure 3. Definition of keywords as hash

Calculating the weights and correlations of bioinformatics articles based on article titles using graph mining techniques and big data applications

In this phase, the Spark was recalled using Python. In order to identify the similarities among article titles, a complete graph was produced of all authors in pairs in which authors represented graph nodes, and each edge between a pair of authors represented the similarity

weight of two titles (Figure 4).

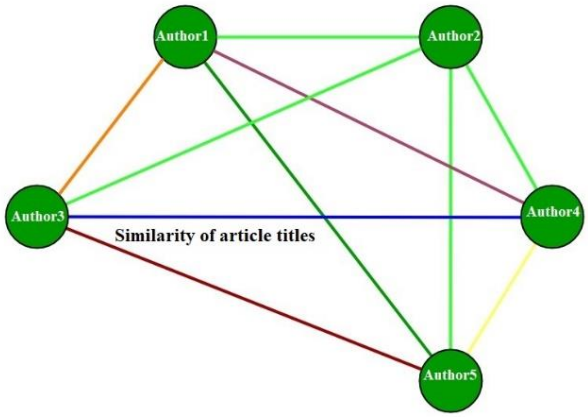

Figure 4. Graph of article titles

In order to measure similarity weights in article titles, the titles were first segmented into words by using the CountVectorizer provided by the scikit-learn library for sentence vectorization. The CountVectorizer parses sentences into a set of tokens. It also deletes the tags and special characters and applies the preprocessing to every individual word. We then rendered the texts into a feature vector to build the incidence matrix for article titles. Term frequency (occurrence) vector was calculated for every article title (Table 3) to measure the distance between every pair of article titles based on cosine similarity. To this end, the words were construed as vectors firstly. For example, Article 1 and Article 2 vectors were formed as (2,1,0,0,0,0,0,0) and (1,1,0,4,0,0,0,1) respectively. Subsequently, their cosine similarity was measured in pairs (Table 4). Article Title 1 and Article Title 2; Article Title 1 and Article Title 3; and Article Title 1 and Article Title n were compared in pairs. Cosine similarity value ranges between 0 and 1. When the two vectors (article titles) are the same, the cosine distance is 1; however, when the two vectors (article titles) are utterly different, the cosine distance is 0.

Table 3
Occurrence of words and Incidence Matrix for each article title

Article title	text	terms
Article title 1	Internet 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

words	Internet	Genom	informatic	Retrieval	Journal	Information	graph	Fuzzy
Article title 1	2	1						
Article title 2	1	1		4				1
Article title 3			1	1	1	1	1	

The cosine distance between Article Title 1 and Article Title 2 is as follows:
Article Title Vector 1: (2, 1, 0, 0, 0, 0, 0, 0)
Article Title Vector 2: (1, 1, 0, 4, 0, 0, 0, 1)

$$\cos\theta = \frac{t1 \,.\, t2}{|t1||t2|} = \frac{2\times1+1\times1+0\times0+0\times4+0\times0+0\times0+0\times0+0\times1}{\sqrt{2^2+1^2}\times\sqrt{1^2+1^2+4^2+1^2}}$$

$$= \frac{3}{\sqrt{5*19}} = 0.31$$

Table 4
Cosine similarity of article titles

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

The following process calculates the Inverse Document Frequency (IDF) that is the normalization of the word frequency. IDF is calculated as follows:

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

When a word appears in all documents, the IDF value for that word is zero. For example, if we have 1000000 article titles with 1000 titles containing the word “internet,” IDF is calculated as follows:

IDF (internet)=log(1000000/1000)= 3

In the next step, TF-IDF is calculated in general. That is, the occurrence of every individual

word in the text is multiplied by the IDF. The calculation is done using the following equation:

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

In the next step, the weights are assigned as the weights of the article title edges for a given pair of authors.

The output of pairwise title weights is illustrated in Figure 5.

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	1173777132670488302	9	0.037900048683237426

Figure 5. Matrix of article titles

Calculating the weights and correlations of bioinformatics articles based on journal titles using graph mining techniques and big data applications

In the next step, all authors produced a complete graph in pairs to determine the similarities among journal titles. The authors represented the nodes, and each edge between a pair of authors represented the similarity weight of common journal titles (Figure 6).

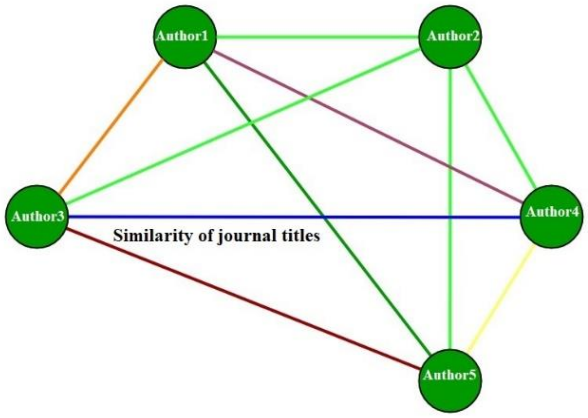

Figure 6. The similarity of common journals

In this step, journal titles were compared in pairs, and their similarities were determined to measure edge weights. The output is illustrated in Figure 6.

i1	i2	key1TitleJournal	key2TitleJournal	indexTitleJournal	ReTitleJournal
0	1	1707300769194029565	-7038021937542463154	0	0.0
0	2	1707300769194029565	3693577778660450793	1	0.0
0	3	1707300769194029565	958879439114911055	2	0.0
0	4	1707300769194029565	8643231380549287432	3	0.0
0	5	1707300769194029565	-3984332359540416383	4	0.0
0	6	1707300769194029565	-8039886378300856057	5	0.0
0	7	1707300769194029565	5428704447652125072	6	0.0
0	8	1707300769194029565	5668758138962396361	7	0.0
0	9	1707300769194029565	-8379529997874255720	8	0.0

Figure 6. Matrix of journal titles

Calculating the weights and correlations of bioinformatics articles based on keywords using graph mining techniques and big data applications

In this step, the weights and correlations of bioinformatics articles were examined based on their keywords. The weights of articles are calculated in pairs based on keywords similarity to measure the similarities of article keywords. The authors represent the nodes, and each edge between a pair of authors represents the similarity weight of article keywords (Figure 6).

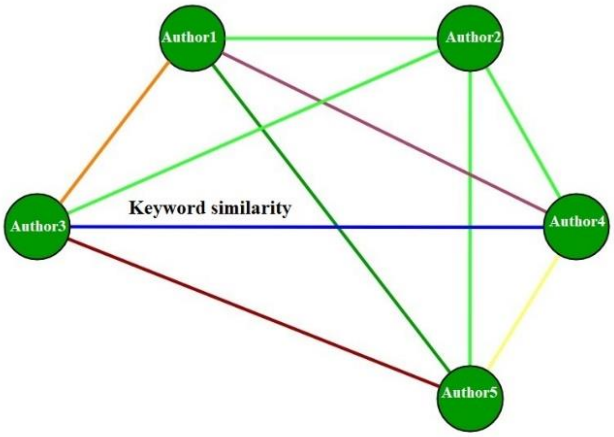

Figure 6. Similarity of article keywords

As with titles, the keywords were compared in article pairs, and the authors were weighted based on keyword similarities.

i1	i2	key1KeywordList	key2KeywordList	indexKeywordList	ReKeywordList
0	1	1707300769194029565	-7038021937542463154	0	0.0
0	2	1707300769194029565	3693577778660450793	1	0.0
0	3	1707300769194029565	958879439114911055	2	0.0
0	4	1707300769194029565	8643231380549287432	3	0.0
0	5	1707300769194029565	-3984332359540416383	4	0.0
0	6	1707300769194029565	-8039886378300856057	5	0.0
0	7	1707300769194029565	5428704447652125072	6	0.0
0	8	1707300769194029565	5668758138962396361	7	0.0
0	9	1707300769194029565	-8379529997874255720	8	0.0
0	10	1707300769194029565	-1172777122670488202	9	0.0
0	11	1707300769194029565	-6491368345707066436	10	0.0
0	12	1707300769194029565	8731956114000595864	11	0.0

Figure 7. Keywords matrix

Calculating the weights and correlations of bioinformatics articles based on abstracts using

graph mining techniques and big data applications

The weights of articles were calculated in pairs based on abstracts similarity to measure the similarities of article abstracts. In this graph, the authors represent the nodes, and each edge between a pair of authors represents the similarity weight of article abstracts (Figure 8). At this stage, to gain the weight of similarities, the words in the abstracts are broken into words. Furthermore, a set of tokens was parsed. Then, the text was converted to the feature vector, and the occurrence matrix was formed for the abstract of articles. Moreover, the similarity of the abstract of the articles was obtained based on cosine similarity and TF_IDF.

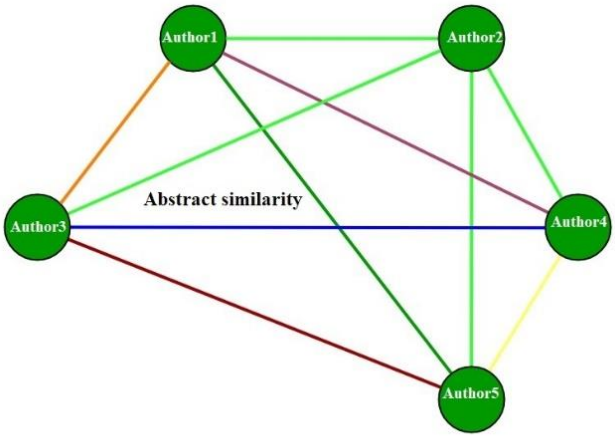

Figure 7. The similarity of article abstracts

In this regard, we measured the similarities of article abstracts in pairs, drew their complete graph, and computed the edge weights:

i1	i2	key1AbstractText	key2AbstractText	indexAbstractText	ReAbstractText
0	1	1707300769194029565	-7038021937542463154	0	0.37559796023114916
0	2	1707300769194029565	3693577778660450793	1	0.37559796023114916
0	3	1707300769194029565	958879439114911055	2	0.29131074204077256
0	4	1707300769194029565	8643231380549287432	3	0.29131074204077256
0	5	1707300769194029565	-3984332359540416383	4	0.29131074204077256
0	6	1707300769194029565	-8039886378300856057	5	0.4297533044894186
0	7	1707300769194029565	5428704447652125072	6	0.34004742044720354
0	8	1707300769194029565	5668758138962396361	7	0.34004742044720354
0	9	1707300769194029565	-8379529997874255720	8	0.34004742044720354
0	10	1707300769194029565	-1172777122670488202	9	0.2632436564159888
0	11	1707300769194029565	-6491368345707066436	10	0.2632436564159888

Figure 8. Abstracts matrix

Calculating the weights and correlations of bioinformatics articles based on author affiliations using graph mining techniques and big data applications

The weights of articles were calculated in pairs based on affiliations similarity to measure the similarities of author affiliations. In this graph, the authors represent the nodes, and each edge between a pair of authors represents the similarity weight of author affiliations (Figure 9).

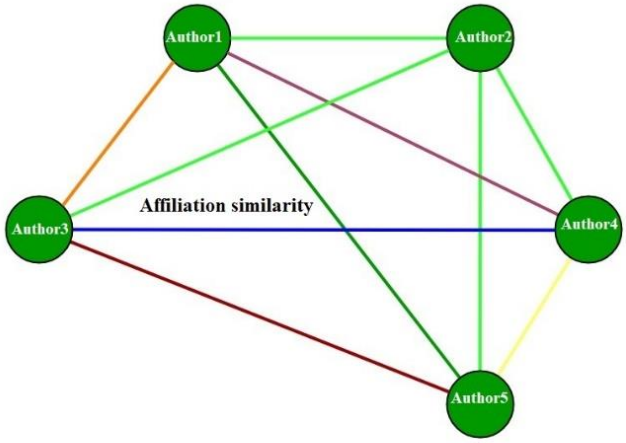

Figure 9. The similarity of affiliations between two nodes

i1	i2	key1Affiliation	key2Affiliation	indexAffiliation	ReAffiliation
0	1	1707300769194029565	-7038021937542463154	0	0.10108753158457662
0	2	1707300769194029565	36935777786604507 93	1	0.07897450086781642
0	3	1707300769194029565	958879439114911055	2	0.07082041768877585
0	4	1707300769194029565	8643231380549287432	3	0.07082041768877585
0	5	1707300769194029565	-3984332359540416383	4	0.06947293668915899
0	6	1707300769194029565	-8039886378300856057	5	0.07082041768877585
0	7	1707300769194029565	5428704447652125072	6	0.12080938583903487
0	8	1707300769194029565	5668758138962396361	7	0.12476168270147527
0	9	1707300769194029565	-8379529997874255720	8	0.13398986921037767
0	10	1707300769194029565	-1172777122670488202	9	0.13398986921037767

Figure 10. Affiliations matrix

Proposing the model of author collaborations in bioinformatics research using graph mining techniques and big data applications

The model for predicting author collaborations was eventually developed using graph mining techniques and big data applications. To this end, the complete graph of all authors of articles was designed such that the authors represented the nodes, and the edges represented the similarity weights of article titles, abstracts, keywords, author affiliations, and journal titles. The weights measured in software were integrated with the weights assigned by experts. So that the final weights were calculated between each pair of nodes

The final weights for predicting co-authorship are as follows:

Similarity nodes = weightArticleTitle * 0.091+ weightabstrac * 0.031+ weightkeyword * 0.055 + weightaffiliation * 0.075 + weightTitleJournal * 0.374

join5 = join5.withColumn('result', 0.091*join5 ['ReArticleTitle'] +0.075*join5 ['ReAffiliation']+0.031*join5['ReAbstractText']+0.374*join5['ReTitleJournal']+0.055*join5['ReKeywordList'])

The final model is illustrated in Figure 12.

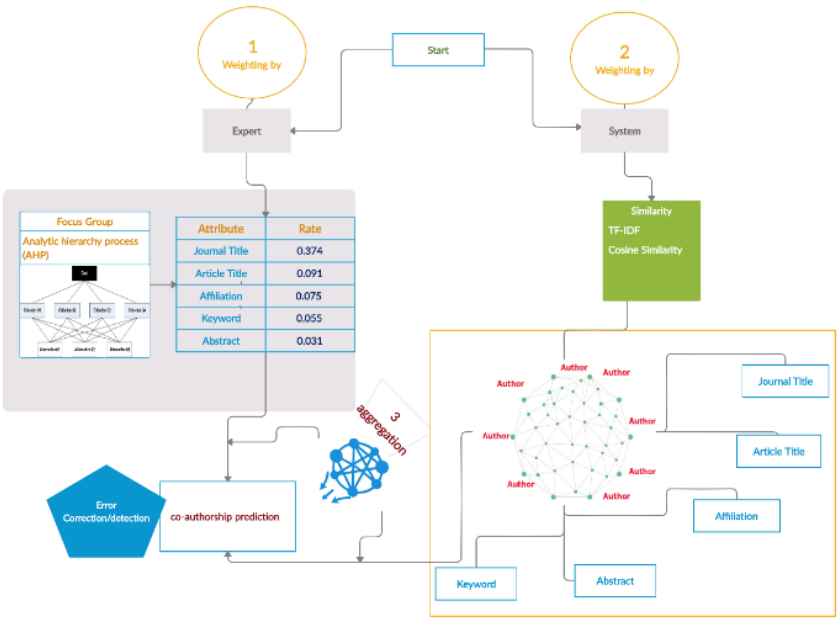

Figure 12. Final model

Discussion

One of the critical issues in proposing scientific collaborators is to use the researchers' opinions because synergy and expert consensus facilitate the selection of scientific partners. The experts of the focus group concurred that identification of core or most-cited authors in a given field was not the critical factor because average authors may assume that core authors show no interest in collaborating with them, or researchers in an institution may be reluctant to collaborate with their colleagues in the same institution (Makarov, Bulanov & Zhukov, 2017). Thus, expert opinions matter in selecting scholarly partners. In this study, "thematic phrases in article titles", "thematic phrases in article abstracts", "thematic phrases in article keywords", "similarity of author affiliations", and "publications in common journals" were selected for weighting. The expert weightings were integrated with system weightings to measure similarities in finding scientific collaborators.

Concerning the weights and correlations of bioinformatics articles based on article titles, a review of the literature revealed that Wu, Mi, Li, Huang & Tong (2018), Wang, Satuluri & Parthasarathy (2007), Li, Chen, Pettit & Rijke (2019), and Chirita, Costache, Nejdl & Handschuh (2007) used TF-IDF feature selection techniques. According to Beel, Gipp, Langer & Breitinger (2016), about 70% of weightings were done using the TF-IDF approach. Salton & Buckley (1988) and Thiyagarajann,Thangavel & Rathipriya (2014) asserted that the cosine similarity method was superior to Hamming similarity criterion in designing a web recommender system. Hasheminezhad, Motieeyan & Nasiri (2018) showed that cosine similarity and Manhattan similarity produced better results than Euclidean distance. One reason for the popularity of cosine distance is that it is highly suitable for assessment, particularly for scattered vectors (Farhadi & Jamzadeh, 2018). Kamyar (2014) contends that the cosine method is one of the best similarity algorithms, with better accuracy than Jaccard and Levenshtein similarities. Magara, Ojo & Zuva (2018) compared similarity criteria in recommender systems and concluded that cosine similarity had the best performance compared with other similarity

criteria.

The title of a work is an echo of its identity; in other words, the title is the first manifestation of the text exposed to the readers. The title is a container whose container is the main idea of the text. In humanities research, some titles assume metaphorical connotations; thus, there is less consistency. However, the title is especially effective in the subject of this thesis addressing the bioinformatics field. DavarPanah (1996) studied the degree of consistency between articles in Persian and their content in different research fields. The results showed that article titles in humanities were less consistent with their contents than medical sciences. The experts attached greater weight to the article titles in the present study than to author affiliations, keywords, and abstracts. Nascimento, Laender, da Silva & Gonçalves (2011) maintained that critical terms in the title weighed three times the critical terms in the article body. Mooney and Roy (2000) and Li et al. (2019) used the title component to design a recommender system for books and articles. Achary (2011) used the title component in his content-based recommender system. Concerning the weights and correlations of bioinformatics articles, the experts attached the most significant weight and priority to journal titles.

Cabanac (2011) argued that journal contents were the main factor for scholarly text recommender systems. He recommended that reading journal articles and conference proceedings were the best way to update the latest developments in a given field. In this section, similarity measurement and weighting were done using TF-IDF and cosine similarity method. In selecting the features, journal ISSNs were also available. Although it was easier to process ISSNs, journal titles were selected as the main component to account for overlaps and proper documentation of author names. This is because one of the functions of journal titles is documentation of author names. Authors typically choose relevant journals based on their expertise. For example, a given author specializing in the genome and who publishes articles in this field tends to publish in genome-related journals. This distinguishes the authors from different fields. Cota, Ferreira, Nascimento, Gonçalves & Laender (2010) disambiguated author names using similarity functions, assuming that authors tend to publish in the same topics and journals. The results showed that this method was 12% more accurate than supervised and unsupervised methods. Han, Giles, Zha, Li & Tsioutsiouliklis (2004) used the probability model for measuring the similarity between author names and article terms to disambiguate author names.

Concerning the weights and correlations of bioinformatics articles, TF-IDF and cosine similarity algorithms were used to calculate keywords similarity. The weights of edges were calculated based on keywords. Keywords are topics and terms that define article contents. Aanonson (1987) showed that keyword search in the titles helps retrieve the relevant documents and the documents that are not retrievable through thematic search. Ghareh- Chamani (2013) used article keywords as the only variable to recommend articles from the CiteSeer website. Mooney and Roy (2000) designed a recommender system based on topical terms. The system was developed to recommend books to Amazon customers based on the Bayesian algorithm. Achary (2011) used keyword tags in Bibsonomy and CiteSeer for his recommender system. Sun, Barber, Gupta, Aggarwal & Han (2011) used the subject component to predict co-authorship in heterogeneous bibliometric networks in the DBLP network. Using the content-based method and TF-IDF algorithm, Chirita et al. (2007) developed a keyword-recommender system in web pages by extracting important keywords from web pages.

About the weights and correlations of bioinformatics articles based on abstracts, it should

be noted that article abstracts are important components because they provide a synopsis of the research. Metadata such as title, author names, publication year, and journal title are common and retrievable features used in different databases for similarity measurement; however, it is not easy to retrieve abstracts in most databases. An abstract contains the gist of a research article that is written meticulously by authors. Cabanac (2011) asserts that it is too costly and difficult to access the full texts and abstracts of scholarly texts for processing. According to expert opinions, the abstract component ranked fifth in this study. When two authors have produced similar abstracts, they are likely to have authored similar articles. Thus, their similarity is determined based on their mutual weight. Text mining tools such as cosine similarity and TF-IDF algorithm were used to calculate article similarities in abstracts. Similarity measurement in abstracts has not been carried out in previous studies. Seemingly, researchers have avoided this due to the bulky processing of abstracts and a lack of required datasets.

Concerning the weights and correlations of bioinformatics articles based on author affiliations using graph mining techniques and big data applications, one should note that affiliation is an essential factor for authors to choose collaborators. Some researchers would prefer to collaborate in science with people in their institution or region. Still, some researchers prefer partners from outside their institution. Departments, laboratories, schools, and universities impose limitations on researchers due to competition with their rival counterparts. The main reason for such competitions is government financial support (Roemer & Borchardt, 2015). Makarov et al. (2017) reported that researchers at the Higher School of Economics of National Research University (HSE) often collaborated with researchers from other institutions. Affiliation is a critical component that researchers use in altmetrics and bibliometrics (Yan & Guns, 2014; Brandão & Moro, 2012; Ho, Bui, & Bui, 2019; Andrikopoulos, Samitas & Kostaris, 2016). Finally, a model was designed to predict author collaborations in bioinformatics research articles using graph mining techniques and big data applications by applying a new method to weight the components. Text mining, information retrieval, big data tools, and graph theory were used in this method based on expert opinion and graph theory. The majority of previous studies on co-authorship prediction have already drawn upon topological approaches in an unsupervised manner without expert opinion. However, we addressed content-based methods, expert weighting, and thematic similarity.

Conclusion

Sufficient information is required for decision-making, thinking, and communication. However, due to the dramatic increase in scholarly research and articles, it is exceedingly difficult for researchers to find potential collaborators. The present study drew on quantitative methods in the big data environment and expert opinion to develop a model that could recommend the most relevant potential scholarly collaborators to a given researcher. The results showed that content-based methods in recommender systems in static networks have considerable potential for finding scientific collaborators in relevant retrieval. Content-based methods involve using different sections of the article content, including title, abstract, and keywords, to recommend the relevant articles based on their similarity with a set of input articles.

One of the operational achievements of this study was the acceleration of relevant author retrievals, which in turn led to more efficiency, a higher quality of research, and scientific development. Furthermore, this recommender system leads to a more convenient selection of

authors. Finding a suitable research collaborator is one of the main challenges in interdisciplinary fields such as bioinformatics. In addition to systemic methods, bioinformatics expert opinion was also drawn upon in this study. This model coordinates authors' concerns for finding the most similar research collaborators with their information needs to guarantee good scientific collaborator recommendations. Future studies could address the predictive model for author collaboration based on behavioral characteristics, the predictive model for author collaboration based on fuzzy algorithms, the predictive model for author collaboration in other bibliometric networks such as Scopus, the co-authorship model in scientific social networks, the predictive model for author collaboration without expert weighting, and implementation of the predictive model for author collaboration based on various algorithms such as Jaccard, Euclidean, simple Bayesian, and neural network algorithms.

Acknowledgements

The present publication is the outcome of the project “From Talent to Young Researcher project aimed at activities supporting the research career model in higher education”, identifier EFOP-3.6.3-VEKOP-16-2017-00007 co-supported by the European Union, Hungary, and the European Social Fund.

References

Aanonson, J. (1987). Precision and Recall in Title keyword searchers. *Information technology and libraries*, 14(3), 162-170. https://doi.org/10.6017/ital.v14i3.5290

Achary, R. (2011). *An author recommendation system using both content-based and collaborative filtering methods*. Master thesis. Department of computer engineering and computer science, California state university, Long Beach. ProQuest Dissertations and Theses database.

Andrikopoulos, A., Samitas, A. & Kostaris, K. (2016). Four decades of the Journal of Econometrics: Coauthorship patterns and networks. *Journal of econometrics*, 195(1), 23-32. https://doi.org/10.1016/j.jeconom.2016.04.018

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4), 305-338. https://doi.org/10.1007/s00799-015-0156-0n

Benton D. (1996). Bioinformatics - principles and potential of a new multidisciplinary tool. *Trends in Biotechnology*, 14(8), 261-272. https://doi.org/10.1016/0167-7799(96)10037-8

Brandão, M. A., & Moro, M. M. (2012, June). Affiliation Influence on Recommendation in Academic Social Networks. *In AMW* (pp. 230-234).

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics*, 87(3), 597–620. https://doi.org/10.1007/s11192-011-0358-1

Chaoji, V., Al Hasan, M., Salem, S., Besson, J. & J. Zaki, M. (2008). Origami: A novel and effective approach for mining representative orthogonal graph patterns. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(2), 67-84. https://doi.org/10.1002/sam.10004

Chirita, P. A., Costache, S., Nejdl, W. & Handschuh, S. (2007). P-TAG: Large scale automatic generation of personalized annotation tags for the web. In *Proceedings of the 16th international conference on World Wide Web* (pp. 845-854), New York, NY, USA: ACM Press.

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A. & Laender, A. H. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology*, 61(9), 1853-1870. https://doi.org/10.1002/asi.21363

Das, K., Samanta, S. & Pal, M. (2018). Study on centrality measures in social networks: A survey. *Social Network Analysis and Mining*, 8(1), 1-11. https://doi.org/10.1007/s13278-018-0493-2

Davarpanah, M. R. (1996). Investigating the compatibility of Persian article titles with their content. *Iranian Journal of Information Processing & Management*, 12 (2), 1-12. [in Persian]

Farhadi, M. & JamZad, M. (2018). Examining similarity criteria in content-based image retrieval. CSJ. No. 9, 13-27. [in Persian]

Ghanei Rad, M. A. (2006). Status of the scientific community in the field of social science. *Journal of Social Science Letter*, 27, 27-55. [in Persian]

Ghare-Chamani, J. (2013). Provide a way to suggest referrals in the referral network. Master thesis. Sharif University of Technology, Computer Engineering Department. [in Persian]

Han, H., Giles, L., Zha, H., Li, C. & Tsioutsiouliklis, K. (2004, June). *Two supervised learning approaches for name disambiguation in author citations. In Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, 2004. (pp. 296-305). IEEE.

Hasheminejad, M, Motieeyan, Z. & Nasiri, J. (2019). Comparison of a recommender text system with three criteria for measuring cosine similarity, Euclidian distance and Manhattan. In *The 6th International Congress on Development and Promotion of Fundamental Science and Technology in Society*, Tehran. [in Persian]

Ho, T. K. T., Bui, Q. V. & Bui, M. (2019, December). Co-author Relationship Prediction in Bibliographic Network: A New Approach Using Geographic Factor and Latent Topic Information. In *Proceedings of the Tenth International Symposium on Information and Communication Technology* (pp. 69-77). https://doi.org/10.1145/3368926.3369668

Kamyar, M. (2014). *Automatic extraction of concepts from text based on linguistic methods*. (Ph.D. thesis). Ferdowsi University of Mashhad, Department of Computer Engineering. [in Persian]

Kiani, M. (2020). Information ecology in field of bioinformatics with emphasis on thematic relationships. (Ph.D. thesis). Isfahan University, Department of Knowledge and Information Science. [in Persian]

Li, X., Chen, Y., Pettit, B., & Rijke, M. D. (2019). Personalised reranking of paper recommendations using paper content and user behavior. *ACM Transactions on Information Systems (TOIS)*, 37(3). 1-23. https://doi.org/10.1145/3312528

Magara, M. B., Ojo, S. O. & Zuva, T. (2018, March). A comparative analysis of text similarity measures and algorithms in research paper recommender systems. In *2018 conference on information communications technology and society (ICTAS)* (pp. 1-5). IEEE. https://10.1109/ICTAS.8368766

Makarov, I., Bulanov, O. & Zhukov, L. E. (2016, May). Co-author recommender system. In *international conference on network analysis* (pp. 251-257). Springer, Cham.

Mooney, R. J. & Roy, L. (2000, June). Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries* (pp. 195-204). https://doi.org/10.1145/336597.336662

Nascimento, C., Laender, A. H., da Silva, A. S. & Gonçalves, M. A. (2011, June). A source independent framework for research paper recommendation. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (pp. 297-306).

Roemer, R. C. & Borchardt, R. (2015). Meaningful metrics: A 21st-century librarian's guide to bibliometrics, altmetrics, and research impact. American Library Association.

Rose, C. (2020, April, 25). What is Bioinformatics?. Retrieved from https://nuclineers.com/whats-bioinformatics/

Saaty, T. L. (1980). *The Analytical Hierarchy Process*. New York: McGraw-Hill.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text Retrieval. *Information Processing & Management*, 24(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C. & Han, J. (2011, July). Co-author relationship prediction in heterogeneous bibliographic networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining* (pp. 121-128). IEEE. https://doi.org/10.1109/ASONAM.2011.112

Tabarzeh, F. (2018). *Analysis of the scientific cooperation network of university professors in the field of social sciences*. (Master's thesis). Tarbiat Modares University, Faculty of Humanities. [in Persian]

Thiyagarajann, R., Thangavel, K. & Rathipriya, R. (2014). Recommendation of Web Pages using Weighted K-Means Clustering. *International Journal of Computer Applications*, 86(14), 44-48. Retrieved from https://b2n.ir/g96542

Wang, C, Satuluri, V, & Parthasarathy, S. (2007). Local Probabilistic Models for Link Prediction. In Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, Omaha, NE, USA, 322–331. https://doi.org/10.1109/ ICDM.2007.108

Wilkinson, S. (2004). Focus group research. In D. Silverman (Ed.), *Qualitative Research. Theory, Method and Practice*. (2 ed.). (pp 177-199). Thousand Oaks, CA.: Sage Publications.

Wu, F., Mi, L., Li, X., Huang, L., & Tong, Y. (2018, March). Identifying Potential Standard Essential Patents Based on Text Mining and Generative Topographic Mapping. In *2018 IEEE International Symposium on Innovation and Entrepreneurship (TEMS-ISIE)* (pp. 1-9). IEEE.

Yan, E. & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics*, 8(2), 295-309. https://doi.org/10.1016/j.joi.2014.01.008