Annals of Library and Information Studies
Vol. 66, June 2019, pp. 57-64

Co-authorship and co-occurrences analysis using Bibliometrix R-package: a case study of India and Bangladesh

Samir Kumar Jalal
Deputy Librarian, Central Library, Indian Institute of Technology (IIT) Kharagpur, Kharagpur-721302. Email: jalalsk1971@gmail.com

Received: 17 July 2018; revised: 20 June 2019; accepted: 25 June 2019

Research collaboration between India and Bangladesh based on research output of 1156 papers jointly produced by both countries were retrieved from Web of Science for the period of 1991 to 2017.The collaboration network on co-authorship and co-occurrences was build using BibliometrixR package. Top ten keywords have been plotted to show the subject trends in general. VOSviewer was used to build network maps to know the collaborative zones with respect to authors, subjects and keywords. The study shows that the major collaborations were in medical science followed by agriculture and biological sciences. The verification of Lotkas's Law was made using lotka() function in Bibliometrix R package. The finding of the study shows that Lotka's law is still valid for the co-authorship data at the international level.

Keywords: Research Collaboration; Bibliometrix R-Package; Co-authorship; Co-occurrence analysis; Collaboration Network; India; Bangladesh

Introduction

Research collaboration aims at finding out the similarity between two or more researchers in a discipline or topics of their interests. Various associated aspects of collaboration like types of collaboration[1], factors affecting international collaboration and research relationship[2] at different levels advocated that co-authorship is not the only factor in determining the collaboration[3]. International collaborative behavior among scientists reflects that scientists working in the core areas of research have higher international collaboration[4]. The collaboration between China and G7 countries, based on Science Citation Index (SCI), was found to be strong due the proliferation of the growth of science in China[5].International collaboration pattern by Indian scientists through the analysis of multi-authored publications was conducted through correspondence analysis by Anuradha et al and showed that data set in physics, chemistry, clinical medicine are the first, the second and the third largest subjects respectively having international collaboration[6]. Another study on the international collaboration pattern during 1998-2007 showed that Iranian scientists collaborated with 115 countries and that geosciences had the biggest number of co-authored publications internationally[7].

Research collaboration especially at the international level has now-a-days, gaining immense importance due to faster technological change, global competition and innovation. Chen, Zhang and Fu[8] reviewed the intellectual base and main research trajectories of the IRC research domain over the period 1957–2015 through co-citation network analysis, main path analysis and bibliographic coupling analysis and found that co-authorship analysis is the main research methods to study research collaboration(RC). But, the geographic, linguistic, political, cultural distances in the context of IRC are more significant than those in other kinds of RC[9].Research cooperation at the global level among industrialized countries especially in Europe and neighbouring countries was made by Georghiou[10].

International Research Collaboration (IRC) influenced the author to ponder over the issues on how scientists and researchers from India and Bangladesh (being a neighbouring country) work together and collaborate on various subject areas. It would be good to know more about the collaboration pattern between them through publication analysis of author's country affiliation studies. Collaboration between India and Bangladesh may encourage global culture which, in turn, may attract students and researchers and also helps for recruitment. The study explores the possibility of university-institute collaboration and fosters more and more collaborative research.

Co-authorship is not the sole measure in research collaboration because some authors, without any

reason, incorporate the names of other authors and affiliation as a token of love and respect though they may not have contributed anything new to the paper[11].Collaboration may be of various types like inter-institutional, inter-country, inter-state and intra-institutional collaboration. A study on initial-based, heuristic and machine learning approaches was considered to solve the name dis-ambiguity problem while building the co-authorship network[12].Vasantha Kumar, Sendhil kumar and Mahalakshmi[13] made a survey to showcase various additions to the existing co-authorship networks. Interpretations of the co-authorship networks for the detection of research communities are also studied. Direct citation, co-citation and bibliographic coupling help to analyze and understand the structure of science and thereby to find out the research fronts. Garfield's (1955)[14] article on *"Citation Indexes for Science*" made it clear that subject indexes could not able to identify the research fronts but could be traced out the historical development of the subject by direct citation analysis, proposed by Garfield (1964)[15].

Co-occurrences are used to understand the underlying patterns of the document set under study. Co-citation, co-word, and co-link studies are the main aspects of co-occurrences in the information sciences. In a large data set, co-occurrences are managed through co-occurrences matrices, which may be of two types: symmetrical and asymmetrical co-citation matrices. In one of the study[16] it is mentioned similarity measures e.g. cosine technique cannot be used in symmetrical co-citation matrix but can be applied to the asymmetrical citation matrix to derive the proximity matrix. Proximity measure indicates how similar of how dissimilar the two objects. White and Griffith(1981)[17] advocated to use first author as unit of analysis in co-citation analysis and named as author co-citation analysis (ACA).

In connection with co-occurrences and co-authorship, several software tools like R, Publish or Perish, Bibexcel, Ucinet, VOSviewer, Pajek, CiteSpace, HistCite, Scholarometer etc., are available for data analysis and visualization. Among these, Bibliometrix R-package is very popular open source software package for bibliometric and scientometric analysis and hence, the present study exploited the package.

Objectives of the study

- To study the year-wise growth patterns of international collaboration between India and Bangladesh in the last 27 years;
- To identify the most preferred journals;
- To build authorship network visualization using R Programming, VOSviewer and Pajek.
- To build keyword co-occurrences map and keyword growth analysis using Bibliometrix R-Package; and
- To find out the main clusters in research collaboration.

Methodology

The study is based on 1156 papers published jointly by authors from affiliated countries of India and Bangladesh during 1991-2017. Data was downloaded from Web of Science on May 25, 2018. Only articles, review, proceedings papers and book chapters have been included. Document types such as editorials, notes and corrections were excluded from the study.

RStudio is free and open source development environment of R and runs on Windows, Linux and other operating system. There are various open source packages available to execute the specific user-driven functions in RStudio. One such package is Bibliometrix R, which is basically developed for bibliometric and scientometric studies. In this study, Bibliometrix R-package was used for data analysis and interpretation.

Results and discussions

Frequency distribution of publications

Figure 1 shows the growth of joint publications between India and Bangladesh during 1991 to 2017. For convenience, 27 years have been grouped into 5-years span and the rest 2-years span and subsequently average yearly publications has been calculated. Figure1 clearly shows that there is a quantum jump observed in the growth of publications after the year 2012.

Verification of Lotka's Law

Lotka's law shows the relationship between numbers of authors and publications. The number of

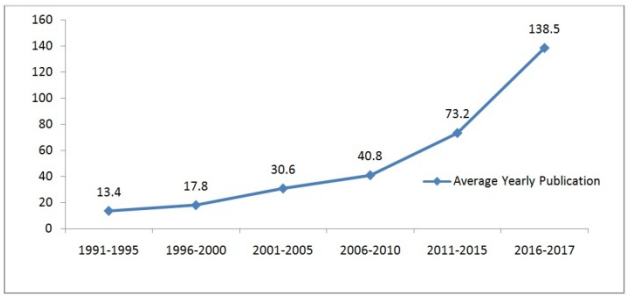

Fig.1 — Trend of international collaboration in publications between India and Bangladesh

authors against their publications or contributions was plotted on a logarithmic scale. The general form of Lotka's law is $Y = \frac{C}{X^n}$ where X is the number of publications; and Y is the relative frequency of authors with X publications, and n and C are constant. Using the syntax in R Programming, L=lotka(results), following data have been generated.

For the given data set, the β = 2.277118, C=1.050457. $R^2 = 0.9467326$ and p value is 0.3364049. The significance of p value is that a small p-value indicates the strong relationship between two variables. Normally, there is an inverse relationship between R^2 and P-value. The higher R-value and lower p-value indicates that data lies on a straight line. R-square value reflects on how much variation is explained by the model. The Lotka's law was also found to be re-established for the case of research collaborative works for the authors at the international level.

Table 1 — Number of articles published by number of corresponding authors

S. No	No. of Articles (X)	No. of Authors (Y)	Frequencies
1.	1	8396	0.6491417968
2.	2	1408	0.1088603680
3.	3	743	0.0574454925
4.	4	814	0.0629349003
5.	5	285	0.0220349467
6.	6	217	0.0167774857
7.	7	189	0.0146126488
8.	8	117	0.0090459255
9.	9	89	0.0068810886
10.	10	93	0.0071903510
11.	11	81	0.0062625638
12.	12	51	0.0039430957
13.	13	67	0.0051801454
14.	14	62	0.0047935673
15.	15	52	0.0040204113
16.	16	48	0.0037111489
17.	17	29	0.0022421525
18.	18	33	0.0025514149
19.	19	25	0.0019328901
20.	20	31	0.0023967837
21.	21	31	0.0023967837
22.	22	15	0.0011597340
23.	23	16	0.0012370496
24.	24	8	0.0006185248
25.	25	7	0.0005412092
26.	26	5	0.0003865780
27.	27	8	0.0006185248
28.	28	4	0.0003092624
29.	29	2	0.0001546312
30.	31	2	0.0001546312
31.	47	1	0.0000773156
32.	50	1	0.0000773156
33.	51	1	0.0000773156
34.	53	1	0.0000773156
35.	65	1	0.0000773156
36.	92	1	0.0000773156

Preferred journals

Scientists preferred to communicate their research papers in journals like Lancet, PLOS Neglected Tropical Disease, PLOS One, BMC Public Health, Field Crops Research etc.

Country collaboration

The study found that countries like USA, UK, Japan, Australia, Canada, Scotland, Switzerland and Germany are the top collaborating countries between India and Bangladesh with respect to this study.

Top 10 authors

There are 12933 unique authors, who have contributed for collaborating the 1156 articles. The result was revealed using Bibexcel at the time of creation of 'author.out' file. With the help of Bibliometrix R-Package and based on full count method, the study revealed that the top five authors having highest publications with joint-institutional affiliations between India and Bangladesh are G.B. Nair (92) from International Centre for Diarrheal Disease Research, Bangladesh, Centre for Health and

Table 2 — Top ten preferred journals

S.N	Sources/ Journals Name	No. of Articles
1.	Lancet	57
2.	PLOS Neglected Tropical Diseases	19
3.	PLOS One	19
4.	Journal of Clinical Microbiology	17
5.	Clinical Infectious Diseases	15
6.	BMC Public Health	14
7.	Field Crops Research	14
8.	Journal of Medical Microbiology	12
9.	Vaccine	12
10.	Hematology International	11

Table 3 — Most productive countries (based on corresponding authors)

S.N	Countries	Articles	Frequency	SCP	MCP
1.	India	318	0.2765	22	296
2.	Bangladesh	288	0.2504	38	250
3.	USA	144	0.1252	0	144
4.	England	57	0.0496	0	57
5.	Japan	53	0.0461	0	53
6.	Australia	33	0.0287	0	33
7.	Canada	30	0.0261	0	30
8.	Scotland	24	0.0209	0	24
9.	Switzerland	19	0.0165	0	19
10.	Germany	17	0.0148	0	17

Note: SCP: Single Country Publications; MCP: Multiple Country Publications

Population Research, Dhaka, Bangladesh and Translational Health Science and Technology Institute, Haryana 122016, India and WHO, South East Asia Registered Off, New Delhi, India; ZA Bhutta (53)being affiliated to more organizations i.e. a) The Hospital for Sick Children, The Centre for Global Child Health, Toronto, ON M5G 0A4, Canada; b) University of Toronto, Toronto, ON, Canada; and c) Aga Khan University, Centre of Excellence in Women & Child Health, Karachi, Pakistan; T. Ramamurthy (51) from Translational Health Science and Technology Institute (THSTI), Faridabad, India and Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States; and M. Rahman (50) from International Centre for Diarrheal Disease Research Bangladesh. But if fractionalized method is adopted, the result becomes fundamentally changed i.e. out of top ten (10) authors only four are being placed as top ten authors.

Authorship pattern

The function *dominance* calculates the authors' dominance ranking as proposed by Kumar & Kumar[18].Function arguments are: *results* (object of class *bibliometrix*) obtained by *biblioAnalysis*; and *k* (the number of authors to consider in the analysis). The Dominance Factor (DF) is a ratio indicating the fraction of multi-authored articles in which a scholar appears as the first author.

DF <-**dominance(**results, k =10) and DF (for Print the output)

Figure 2 shows the trend line, which is an inverse relationship between the dominance factor and multi-authored papers. The linear equation: y = -78.37x + 55.28 indicates an inverse relationship between two variables i.e. (x) and multi-authored paper (y) as the value of the co-efficient of x variable is negative.

Although there is an inverse relationship in Figure 2, a few ups and downs with top 10 authors was observed. But, it is quite evident that values of dominances are higher for Rahman MM and Ramamurthy T with respect to their corresponding multi-authored papers.

H-index and g-index

Bibliometric indicators aim to quantitatively determine the impact of research output as measured through scholarly publications. Impact of any research output cannot be measured directly. Hirsch[19] had introduced the concept of h-index, which is defined as number of articles h that each articles receives at least h citations. The h-index is an author level metrics.

Another popular bibliometric indicator is g-index, developed by Leo Egg he in 2006[20]. It is the unique largest number such that top g papers together receive

Table 4 — Top ten authors based on articles as whole count and fractionalized method

S.N	Authors	Articles	Authors	Articles Fractionalized
1.	Nair GB	92	Nair GB	10.80
2.	Gupta R	65	Rahman MM	6.68
3.	Bhutta ZA	53	Ramamurthy T	5.41
4.	Ramamurthy T	51	Islam MS	5.05
5.	Rahman M	50	Salimullah M	4.71
6.	Rahman MM	47	Takeda Y	3.58
7.	Khang YH	31	Haque R	3.57
8.	Mondal D	31	Yamasaki S	3.30
9.	Sepanlou SG	29	Rathod HT	3.25
10.	Takeda Y	29	Ghosh S	3.22

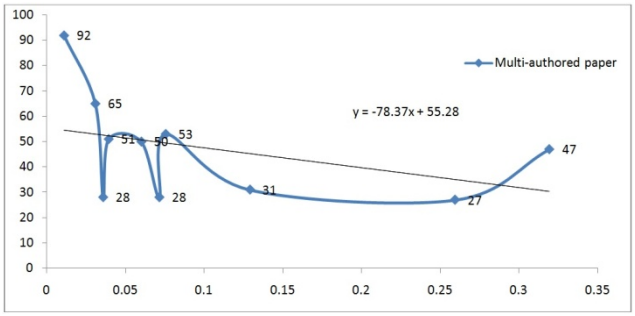

Fig. 2 — Relationship between DF and Multi-authored paper

Table 5 — List of dominance Factor for top 10 authors

S.N	Author	DF (x)	Multi-authored paper (y)	First Authored	Rank by Articles	Rank by DF
1	Nair GB	0.01086957	92	1	1	10
2	Gupta R	0.03076923	65	2	2	9
3	Yamasaki S	0.03571429	28	1	9	8
4	Ramamurthy T	0.03921569	51	2	4	7
5	Rahman M	0.06000000	50	3	5	6
6	Murray CJL	0.07142857	28	2	8	5
7	Bhutta ZA	0.07547170	53	4	3	4
8	Mondal D	0.12903226	31	4	7	3
9	Haque R	0.25925926	27	7	10	2
10	Rahman MM	0.31914894	47	15	6	1

g^2or more citations. It is true that g is greater than or equal to h and i (i.e.$g \geq h > i$). The concept of successive g-index was proposes by Tol[21]. Mathematically, g-index may be written as: $g2 = \sum_{i \leq g}^{n} ci$.From Bibliometrix R-Package, following R Chunk are used to calculate h-index, g-index, m-index etc.

```
>authors=gsub(","," ",names(results$Authors)[1:10])
>indices<- Hindex(M, authors, sep = ";",years=50)
>indices$H
```

Network analysis

Network approach is widely used technique in bibliometric and scientometric studies. The important techniques are co-authorship studies and building the co-citation map.

Co-authorship and co-citation map

VOSviewer is used to create co-authorship, keyword co-occurrences, citations, bibliographic coupling or co-citation map based on bibliographic data. The supported file formats are .txt or .ris or .csv from the database like web of Science, Scopus and PubMed. It is also possible to create term co-occurrences map based on text data. The question is how to handle large amount of data from Web of Science (WoS) because WOS has a limitations to download only 500 records. There may be multiple files containing 500 or less number of records in each file. All such files are to be merged using 'copy' command in DOS-prompt to create a single file, which may be used for network analysis and building network graph. The co-authorship network graph has been built using VOSviewer software as shown in Figure 3.

Association strength network approach has been applied while constructing the network graph. It has been seen from the above graph that there are predominantly three clusters with total 500 authors distributed in three clusters namely claster-1 with red colour having 294 authors, cluster-2 having green colour having 202 authors and cluster-3 with blue colour having 4 authors. Co-authorship analysis was performed based on full counting method using VOSviewer where unit of analysis is author. Articles

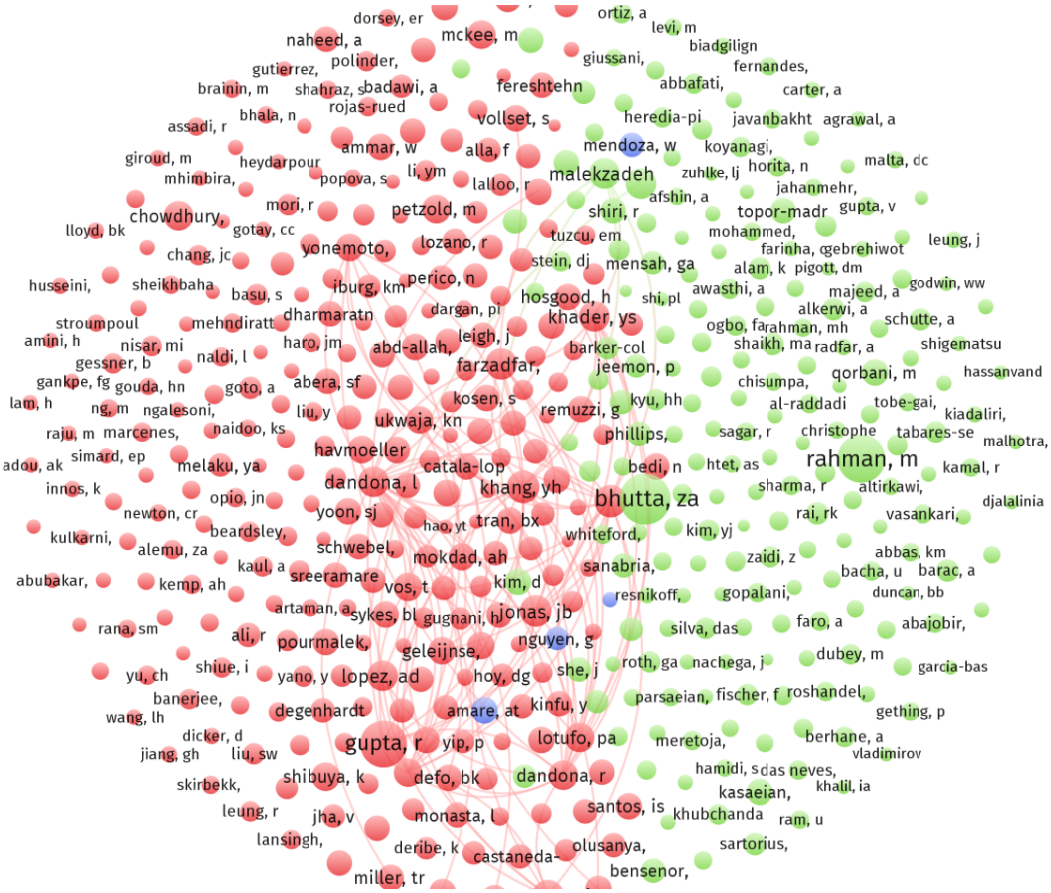

Fig. 3 — Co-authorship network based on available data using VOSviewer

having more than 25 authors have been ignored for analysis. Minimum number of documents for an author should 5 and minimum citation of an author should be 5. Based on this threshold, for each of the 293 authors, the total strength of co-authorship links with other authors was automatically calculated through VOSviewer software and network graph was generated. Thresholds helps to restrict to choose the number of nodes based on citations received. Fazli studied co-authorship patterns and topic networks using the threshold 5:5:20 (citation: co-citation: co-citation cosine coefficient)[22].

The major subjects under collaboration between authors from India and Bangladesh are: a) Public environmental and occupational health (110 papers), medicine general (106), microbiology (101), infectious disease (99), environmental sciences (90), immunology (71), plant sciences (59), Engineering (50), multidisciplinary (48), agronomy (48) and others. Some funding agencies are Bill Melinda Gates Foundation, National Institutes of Health Fogarty International Center, Medical Research Council, National Institute of Health, Population Health Research Institute, Welcome Trust, Glaxo Smithkline etc.

Keyword analysis and building co-occurrences map

Table 7 explains the most relevant top ten author keywords using the summary syntax.

Keyword co-occurrences analysis has also been conducted using VOSviewer technique by importing the raw file and generating the map as shown Figure 4.

Table 6 — Bibliometric indicators

S.N	Author	h_index	g_index	m_index	Article	TC	NP
1.	Nair GB	41	67	1.640000	92	5672	164
2.	Ramamurthy T	28	64	1.120000	51	4280	81
3.	Takeda Y	24	44	0.960000	29	2027	51
4.	Yamasaki S	22	34	1.157895	28	1262	51
5.	Bhattacharya SK	24	41	1.142857	27	1762	44
6.	Rahman MM	24	41	0.960000	47	1894	41
7.	Faruque SM	20	38	1.052632	23	1463	39
8.	Alam MM	18	32	1.200000	19	1028	32
9.	Datta S	19	28	1.000000	19	967	49
10.	Chakraborti D	25	31	1.136364	25	2328	31

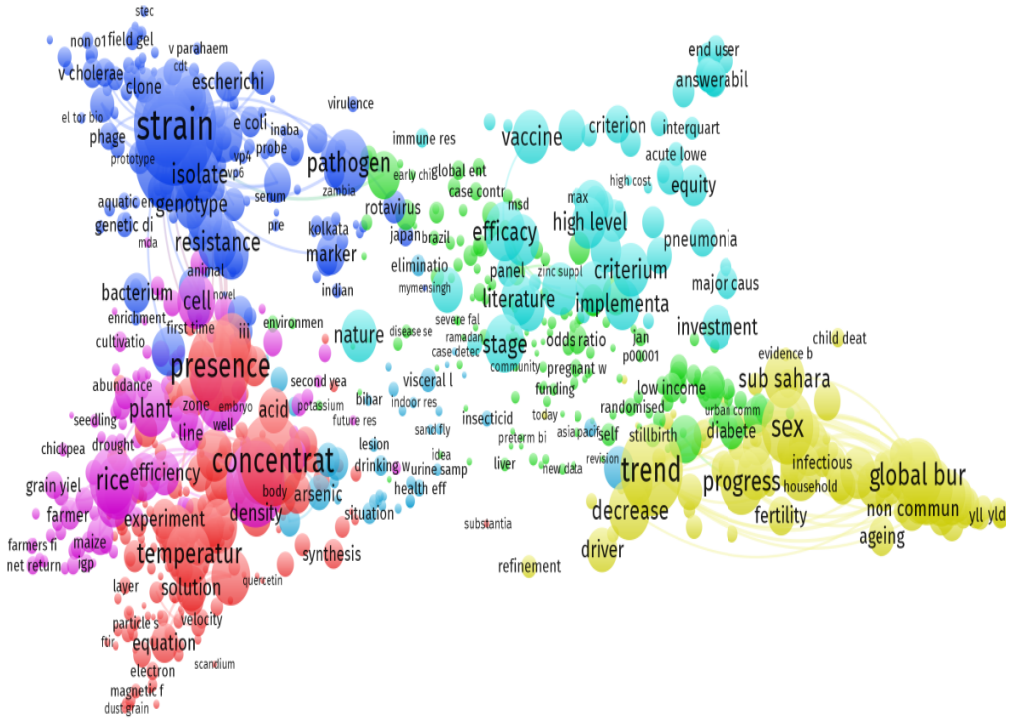

Fig. 4 — Keyword co-occurrences analysis based on WoS data using VOSviewer

Table 7 — Top ten author keywords and Index-keywords

S.N	Author Keywords (DE)	Articles	Keywords-Plus (ID)	Articles
1.	Bangladesh	56	Bangladesh	100
2.	Asia	22	India	84
3.	India	21	Children	62
4.	South Asia	20	Mortality	45
5.	Rice	14	Prevalence	42
6.	Epidemiology	11	Strains	40
7.	*Vibrio cholerae*	11	Disease	39
8.	Cholera	10	Developing-Countries	35
9.	Diarrhea	10	Growth	35
10.	Visceral Leishmaniasis	10	Management	3

There are 857 items distributed over seven clusters: cluster-1(198 items), cluster-2(186), cluster-3(149), cluster-4(113), cluster-5(104), cluster-6(62), and cluster-7(45). Keyword analysis was made by executing the R chunk like installing "reshape2" and "ggplot2 utilities of R-programming.

Conclusion

Co-authorship and co-occurrences analysis are two important areas in research collaboration not only at the individual level but also at the country level. The scientific collaboration between India and Bangladesh has been analyzed quantitatively and graphically from different perspectives based on their publication data. The result of the co-authorship study showed that there was a sharp increase in the growth of research publications between India and Bangladesh since 2012. Keyword co-occurrences analysis and its growth pattern based on Web of Science (WoS) have been focused and found a steady growth. Collaborative research studies between these two countries have keywords such as mortality, strain, prevalence, diseases, etc., as the leading keywords. The result of the study also found that there is an inverse relationship between the dominance factor and multi-authored papers between India and Bangladesh.

Acknowledgements

Author is very much grateful to the anonymous reviewers for their valuable inputs that helped me to further enhance the quality of the research work.

References

1 Subramanyam K, Bibliometric studies of research collaboration: a review,*Journal of Information Science*, 6(1) (1983) 33–38.

2 Stead GB and Harrington T F, A process perspective of international research collaboration,*The Career Development Quarterly*, 48(2000) 323–331.

3 Katz J S and Martin B R, What is research collaboration? *Research Policy*, *26* (1997)1–18.

4 Davidson F J and Carpenter M P, International research collaboration, *Social Studies of Science*, 9(4) (1979) 481–497. Available at: https://doi.org/10.1177/030631277900900405(Accessed on 28th June 2019).

5 He T, International scientific collaboration of China with the G7 countries, *Scientometrics*, 80(3) (2009) 571–582. Available at: https://doi.org/10.1007/s11192-007-2043-y (Accessed on 7th April 2019).

6 Anuradha K T and Urs S R, Bibliometric indicators of Indian research collaboration patterns: A correspondence analysis, *Scientometrics*, 71(2) (2007) 179–189.

7 Hayati Z and Didegah F, International scientific collaboration among Iranian researchers during 1998-2007, *Library Hi Tech*, 28(3) (2010) 433–446.

8 Chen K, Zhang, Y and Fu X, International research collaboration: An emerging domain of innovation studies? *Research Policy,*48 (2019)149–168.

9 Fu X and Li J, Collaboration with foreign universities for innovation: evidence from Chinese manufacturing firms,*International Journalof Technology* Management, 70(2-3) (2016)193-217.

10 Georghiou L, Global cooperation in research,*Research Policy*, 27(1998)611-626.

11 Kim KW, Measuring international research collaboration of peripheral countries: taking the context into consideration, *Scientometrics*,66 (2006) 231–240.

12 Savić M, Ivanović M and Jain LC,Analysis of co-authorship networks,*Intelligent Systems Reference Library*,148 (2019).

13 Kumar V V, Sendhilkumar S and Mahalakshmi GS, A power-graph based approach to detection of research communities from co-Authorship networks,*Journalof Computational and Theoretical Nanoscience,*14(12) (2017), 5686–5695.

14 Garfield E, Citation indexes for science:a new dimension in documentation through association of ideas,*Science*, 80(122) (1955) 108–111.

15 Garfield E, Science Citation Index-a new dimension in indexing,*Science New Series*,144 (1964) 649–654.

16 Leydesdorff L and Vaughan L, Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment,*Journalof American Societyof Information Science&Technology,*57 (2006) 1616–1628.

17 White H D and Griffith B C, Author cocitation: A literature measure of intellectual structure, *Journal of the Associationfor Information Science and Technology,*32(3)(1981) 163–171.

18 Kumar S and Kumar S, Collaboration in Research Productivity in Oil Seed Research Institutes of India,In *Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting* Humboldt- Universitat zu Berlin, Institute for Library and Information Science (IBI) (ed. H. Kretschmer & F. Havemann (Eds)) 1–18 (2008).

19 Hirsch JE, An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Science,*102 (2005)16569–16572.

20 Egghe L, Theory and practice of the g-index, *Scientometrics,*69(2006) 131–52.

21 Tol RSJ, A rational, successive g-index applied to economics departments in Ireland,*Journalof Informetrics*,2(2) (2008) 149–155.

22 Fazli F, Karimi M and Hamzehei R, Co-authorship patterns and topic networks in the scientific publication of Hamadan University of Medical Sciences,*Library Philosophyand Practice,* (2018). Available at:https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=5124&context=libphilprac (Accessed on 28th June 2019).