A Case Study on A Miner Dataset: Identifying leading research through various Models

Dr.G.Ayyappan[1] , Dr.C.Nalini[2] , Dr.A.Kumaravel[3]
Associate Professor, Department of Computer Science, Bharath University[1]
Professor, Department of Computer Science, Bharath University[2]
Professor & Head, School of Computing, Bharath Unviersity[3]
ayyappangmca@gmail.com[1], drnalinichidambaram@gmail.com[2], drkumaravel@gmail.com[3]

Abstract: The increasing tendency across scientific disciplines to write multi authored papers [1,2] makes the issue of the sequence of contributors' names a major topic both in terms of reflecting actual contributions and in a posteriori assessments by evaluation committees. The reviewers aware that there are different cultures to authorship order. The usual and informal practice of giving the whole credit (impact factor) to each author of a multi authored paper is not adequate and over emphasizes the minor contributions of many authors. Similarly, evaluation of authors according to citation frequencies means often overrating resulting from high-impact but multi authored publications. Teja Tscharntke et al. [72] proposed that four methods. Like as SDC,EC, FLAE, and PCI. Comparison of the credit for contributions to this study under the four different models has been suggested. The proposed systems, such as Individual Frequency (IF) and Weighted Frequency (WF), have no repeated impact for each position.

Keywords: SDC,EC,FLAE,PCI,IF and WF.

I INTRODUCTION

Social network is a structure of human relations and association. It is made up of a organized social actors in a network form. Information has varied number of forms and various purposes for communication. Journals serve as major source of primary information. The Researchers tend to publish more and more research output in journals. This paper focuses on the novel techniques for identifying the leading research contribution in the clusters of research social networks. M. T. Rahman et al. proposed an approach that can be used to measure the impact of an author of a multi-authored paper in a more accurate way than either giving each author full credit or dividing credit equally. The proposed proposal not only resolved the long-standing concern for the fair distribution of each author's credit depending on his/her contribution, but it will also, hopefully, discourage the addition of non-contributing authors to a paper. Tasleem Arif proposed a method that use a token-based similarity score in this first stage of comparison and based on the results of the first stage it uses a character-based similarity score in the second stage. Experimental results obtained on standard datasets indicated that the proposed technique shows a lot of improvements over the existing methods. J. M. Warrender proposed a simple tool that assisted researchers in assessing contributions to a scientific publication, for ease in evaluating which contributor qualify for authorship, and in what order the authors should be listed. The tool identified four phases of activity leading to a publication.

In this paper organizes section one has related works and brief introduction of these fields, In section two represents materials and methods, In section three describes results and discussions and the section four presents conclusion

II MATERIALS AND METHODS

Social network analysis A "social" network is defined as a group of collaborating (or competing) entities that have some type of relationship and interact within a shared environment often referred to as a community. Author collaboration Research collaboration or Author collaboration can be defined as the working together of researchers to achieve the common goal of producing new research knowledge. The dataset named as topic_paper_author in the academic social network data from AMiner [8,9]. The dataset is collected for the purpose of cross domain recommendation. The attributes contain Data Mining, Medical Informatics, Theory, Visualization and Database areas. In this research work used Weka 3.6.9 [11], open source software for Text Mining process, MeSH [12] for identification of domain and SAS University Edition [13,14] which is getting permission to access SAS Studio from SAS Institute for Mining of the research community. Based on this novel metrics, the "Top most Influential researcher" of research community has been identified.

III EXPERIMENTS AND RESULTS

In this section presents experiment ans results of the various methods like as Sequence Credit Method, Equal Contribution (EC) Method, First Last Author Emphasis (FLAE) Method, Percent Contribution Indicated (PCI) Method , Individual Frequency Method and Weighted Frequency Method.

Sequence Determines Credit (SDC) Method

The sequence of authors should reflect the declining importance of their contribution, as suggested by the previous studies. Authorship order only reflects relative contribution, whereas evaluation committees often need quantitative measures.

Therefore, the SDC method suggested that the first author should get the credit for the whole impact (impact factor), the second author one-half of the impact factor, the third author one-third of the impact factor, and so forth, up to rank 10. When papers have more than 10 authors, the contribution of each author from the 10th position onwards was then valuated at just 5%.

The leading research contributors based on the research work in the Topic_Paper_Author dataset were ranked by using the SDC method. Table 1.1 lists the top 20 leading contributors in the dataset. By using the SDC method, the leading author with the highest frequency (58.5) was identified as Mr. Surajit Chaudhuri.

Table 1.1 The top 20 authors ranking by using Sequence-Determines-Credit (SDC) Method

S. No.	Author	Sequence-Determines-Credit (SDC)
1	**Surajit Chaudhuri**	**58.5**
2	Jiawei Han	56.5
3	Rakesh Agrawal	55
4	Philip S. Yu	51.5
5	Richard T. Snodgrass	48.5
6	NogaAlon	46.5
7	H. V. Jagadish	46.5
8	Hector Garcia-Molina	42
9	Michael J. Franklin	41.5
10	Christos Faloutsos	41
11	Michael Stonebraker	38.6
12	Baruch Awerbuch	38
13	Michael J. Carey	37.5
14	Jennifer Widom	35
15	Piotr Indyk	34.5
16	Jon M. Kleinberg	34.5
17	S. Muthukrishnan	34.5
18	Hans-Peter Kriegel	33
19	Divesh Srivastava	32.5
20	Kenneth A. Ross	32.5

Equal Contribution (EC) Method

Authors use alphabetical sequence to acknowledge similar contributions or to avoid disharmony in the collaborating groups. EC method suggested that the contribution of each author is valuated as an equal proportion (impact divided by the number of all authors, but a minimum of 5%).The leading research contributors based on the research work in the Topic_Paper_Author dataset were ranked by using the EC method.

In Table 1.2 lists the top 20 leading contributors in the dataset. By using the EC method, the leading author with the highest frequency (19) was identified as Mr. MiklósAjtai.

Table 1.2 The top 20 authors ranking by using Equal Contribution(EC) method

S.No.	Author	Equal Contribution (EC)
1	**MiklósAjtai**	**19**
2	Rakesh Agrawal	17.5
3	Matthew Andrews	17
4	Dimitris Achlioptas	15
5	Timothy M. Chan	14
6	Shree K. Nayar	13.86
7	Hector Garcia-Molina	13.26
8	Jiawei Han	13.02
9	Graham Cormode	13
10	Susanne Albers	13
11	Michael J. Franklin	12.4
12	Surajit Chaudhuri	12.24
13	Alan M. Frieze	12.21
14	Michael Alekhnovich	12
15	Beng Chin Ooi	11.75
16	Christos Faloutsos	11.56
17	David Eppstein	11.55
18	Yannis E. Ioannidis	11.55
19	Hongjun Lu	11.5
20	Wei Wang	11.25

First Last Author Emphasis (FLAE) Method

In many laboratories, the great importance of last authorship is well established. FLAE method suggested that the first author should get the credit of the impact factor, the last author half, and the credit of the other authors should be the impact divided by the number of all authors.

The leading research contributors based on the research work in the Topic_Paper_Author dataset were ranked by using the FLAE method.

Table 1.3 lists the top 20 leading contributors in the dataset. By using the FLAE method, the leading author with the highest frequency (49.92) was identified as Mr. Surajit Chaudhuri.

Table 1.3 The top 20 authors ranking by using First-Last-Author-Emphasis(FLAE) method

S.No.	Author	First-Last-Author-Emphasis (FLAE)
1	**Surajit Chaudhuri**	**49.92**
2	Rakesh Agrawal	48
3	Richard T. Snodgrass	40.36
4	H. V. Jagadish	32.9
5	Jiawei Han	30.58
6	Pankaj K. Agarwal	30.49
7	Michael J. Franklin	29.5
8	David Eppstein	29.48
9	Michael J. Carey	28.14
10	Michael Stonebraker	28.14
11	C. Mohan	27.32
12	Jon M. Kleinberg	27.24
13	Avrim Blum	27.15
14	SudiptoGuha	26.88
15	Piotr Indyk	26.7
16	Kenneth A. Ross	26.26
17	David R. Karger	25.25
18	Ling Liu	24.52
19	Christos Faloutsos	23.84
20	Daniel A. Keim	23

Percent Contribution Indicated (PCI) Method

There is a trend to detail each author's contribution. It has used to establish the quantified credit.

The leading research contributors based on the research work in the Topic_Paper_Author dataset were ranked by using the PCI method.

Table 1.4 lists the top 20 leading contributors in the dataset. By using the PCI method, the leading author with the highest frequency (80.5) was identified as Mr. Jiawei Han.

Table 1.4 The top 20 authors ranking by using Percent-Contribution-Indicated (PCI) method

S.No.	Author	Percent-Contribution-Indicated (PCI)
1	**Jiawei Han**	**80.5**
2	Philip S. Yu	79.8
3	Surajit Chaudhuri	68.5
4	Hector Garcia-Molina	66.8
5	Rakesh Agrawal	65.3
6	H. V. Jagadish	57.3
7	Christos Faloutsos	57.1
8	Michael J. Franklin	55.5
9	Richard T. Snodgrass	54.5
10	Hans-Peter Kriegel	51.2
11	Jeffrey F. Naughton	50.1
12	Rajeev Motwani	49.2
13	Jennifer Widom	48.7
14	S. Muthukrishnan	48.4
15	NogaAlon	46.9
16	Divesh Srivastava	46.5
17	Arie E. Kaufman	46.1
18	Michael Stonebraker	45.9
19	Michael J. Carey	45.8
20	David J. DeWitt	45.1

Individual Frequency

The leading research contributors based on the research work in the Topic_Paper_Author dataset were ranked by using the IF method. Table 1.5 lists the top 20 leading contributors in the dataset. By using the Individual Frequency (IF) method, the leading author with the highest frequency (80.5) was identified as Mr. Jiawei Han. The other authors who formed the top nineteen ranking include Mr. Philip S. Yu (79.8), Mr. Surajit Chaudhuri (68.5), Mr. Hector Garcia-Molina (66.8), Mr. Rakesh Agrawal (65.3), Mr. Christos Faloutsos (68), Mr. H. V. Jagadish (64), Mr. Jeffrey F. Naughton (64), Mr. Divesh Srivastava (62), Mr. Michael J. Franklin (62), Mr. Jennifer Widom (61), Mr. Hans-Peter Kriegel (60), Mr. Rajeev Motwani (60), Mr. Richard T. Snodgrass (60), Mr. Arie E. Kaufman (58), Mr. S. Muthukrishnan (58), Mr. David J. DeWitt (55), Mr. Michael Stonebraker (55), Mr. Raghu Ramakrishnan (52), and Mr. Michael J. Carey (51).

Weighted Frequency

The leading research contributors based on the research work in the Topic_Paper_Author dataset were ranked by using the WF method. Table 1.6 lists the top 20 leading contributors in the dataset. By using the WF method, the leading author with the highest frequency (80.5) was identified as Mr. Jiawei Han. The other authors who formed the top nineteen ranking include Mr. Philip S. Yu (79.8), Mr. Surajit Chaudhuri (68.5), Mr. Hector Garcia-Molina (66.8), Mr. Rakesh Agrawal (65.3), Mr. H. V. Jagadish (42.68), Mr. Michael J. Franklin (38.11), Mr. Baruch Awerbuch (38), Mr. Philip S. Yu (37.61), Mr. Hector Garcia-Molina (36.81), Mr. Michael J. Carey (34.82), Mr. Michael Stonebraker (34.56), Mr. Christos Faloutsos (33.53), Mr. Jon M. Kleinberg (32.83), Mr. Piotr Indyk (31.95), Mr. Charu C. Aggarwal (31), Mr. Marianne Winslett (31), Mr. Pankaj K. Agarwal (30.83), Mr. SudiptoGuha (30.36), and Mr. Kenneth A. Ross (30.29).

Table 1.5 The top 20 authors ranking by using Individual Frequency (IF) method

S.No.	Author	Individual_Frequency
1	Philip S. Yu	100
2	Jiawei Han	93
3	Hector Garcia-Molina	78
4	Surajit Chaudhuri	72
5	Rakesh Agrawal	70
6	Christos Faloutsos	68
7	H. V. Jagadish	64
8	Jeffrey F. Naughton	64
9	Divesh Srivastava	62
10	Michael J. Franklin	62
11	Jennifer Widom	61
12	Hans-Peter Kriegel	60
13	Rajeev Motwani	60
14	Richard T. Snodgrass	60
15	Arie E. Kaufman	58
16	S. Muthukrishnan	58
17	David J. DeWitt	55
18	Michael Stonebraker	55
19	Raghu Ramakrishnan	52
20	Michael J. Carey	51

Table 1.6 The top 20 authors ranking by using Weighted Frequency (WF) method

S.No	Author	Weighted_Frequency
1	Surajit Chaudhuri	57.58
2	Rakesh Agrawal	52.33
3	Jiawei Han	49.45
4	NogaAlon	46.5
5	Richard T. Snodgrass	45.16
6	H. V. Jagadish	42.68
7	Michael J. Franklin	38.11
8	Baruch Awerbuch	38
9	Philip S. Yu	37.61
10	Hector Garcia-Molina	36.81
11	Michael J. Carey	34.82
12	Michael Stonebraker	34.56
13	Christos Faloutsos	33.53
14	Jon M. Kleinberg	32.83
15	Piotr Indyk	31.95
16	Charu C. Aggarwal	31
17	Marianne Winslett	31
18	Pankaj K. Agarwal	30.83
19	SudiptoGuha	30.36
20	Kenneth A. Ross	30.29

Table 1.7 represents all the methods for identifying the frequency of an author contribution up to ten authors in the research articles and considerations of their names commonly occurrence in all the existing and proposed method.

Table 1.7 Comparison of the high frequency authors by using the existing and proposed methods

S. No.	Author	SDC	EC	FLAE	PCI	IF	WF
1	Surajit Chaudhuri	58.5	12.24	49.92	68.5	72	**57.58**
2	Rakesh Agrawal	55	17.5	48	65.3	70	52.33
3	Jiawei Han	56.5	13.02	30.58	80.5	**93**	49.46
4	Michael J. Franklin	41.5	12.4	29.5	55.5	62	38.12
5	Christos Faloutsos	41	11.56	23.84	57.1	68	33.53

Surajit Chaudhuri was having impact of frequencies respectively 58.5 in Sequence Determines Credit (SDC) method, 12.24 in Equal Contribution (EC) method, 49.92 in First Last Author Emphasis (FLAE) method, 68.5 in Percent Contribution Indicated (PCI) method, 72 in Individual Frequency (IF) method, and 57.58 in Weighted Frequency (WF) method.

Rakesh Agrawal was having impact of frequencies respectively 55 in Sequence Determines Credit (SDC) method, 17.5 in Equal Contribution (EC) method, 48 in First Last Author Emphasis (FLAE) method, 65.3 in Percent Contribution Indicated (PCI) method, 70 in Individual Frequency (IF) method, and 52.33 in Weighted Frequency (WF) method.

Jiawei Han was having impact of frequencies respectively 56.5 in Sequence Determines Credit (SDC) method, 13.02 in Equal Contribution (EC) method, 30.58 in First Last Author Emphasis (FLAE) method, 80.5 in Percent Contribution Indicated (PCI) method, 93 in Individual Frequency (IF) method, and 49.46 in Weighted Frequency (WF) method.

Michael J. Franklin was having impact of frequencies respectively 41.5 in Sequence-Determines Credit (SDC) method, 12.4 in Equal Contribution (EC) method, 29.5 in First Last Author Emphasis (FLAE) method, 55.5 in Percent Contribution Indicated (PCI) method, 62 in Individual Frequency (IF) method, and 38.12 in Weighted Frequency (WF) method.

Christos Faloutsos was having impact of frequencies respectively 41 in Sequence Determines Credit (SDC) method, 11.56 in Equal Contribution (EC) method, 23.84 in First Last Author Emphasis (FLAE) method, 57.1 in Percent Contribution Indicated (PCI) method, 68 in Individual Frequency (IF) method, and 33.53 in Weighted Frequency (WF) method.

The existing one of the method Sequence Determines Credit (SDC) approach and our proposed framework Weighted Frequency (WF) methods provided almost the same result. However, in the case of papers with more than 10 authors, the results had a vast difference for both the methods.

The proposed systems, such as Individual Frequency (IF) and Weighted Frequency (WF), have no repeated impact for each position. However, all the existing systems have almost repeated impact for each of the researcher position.

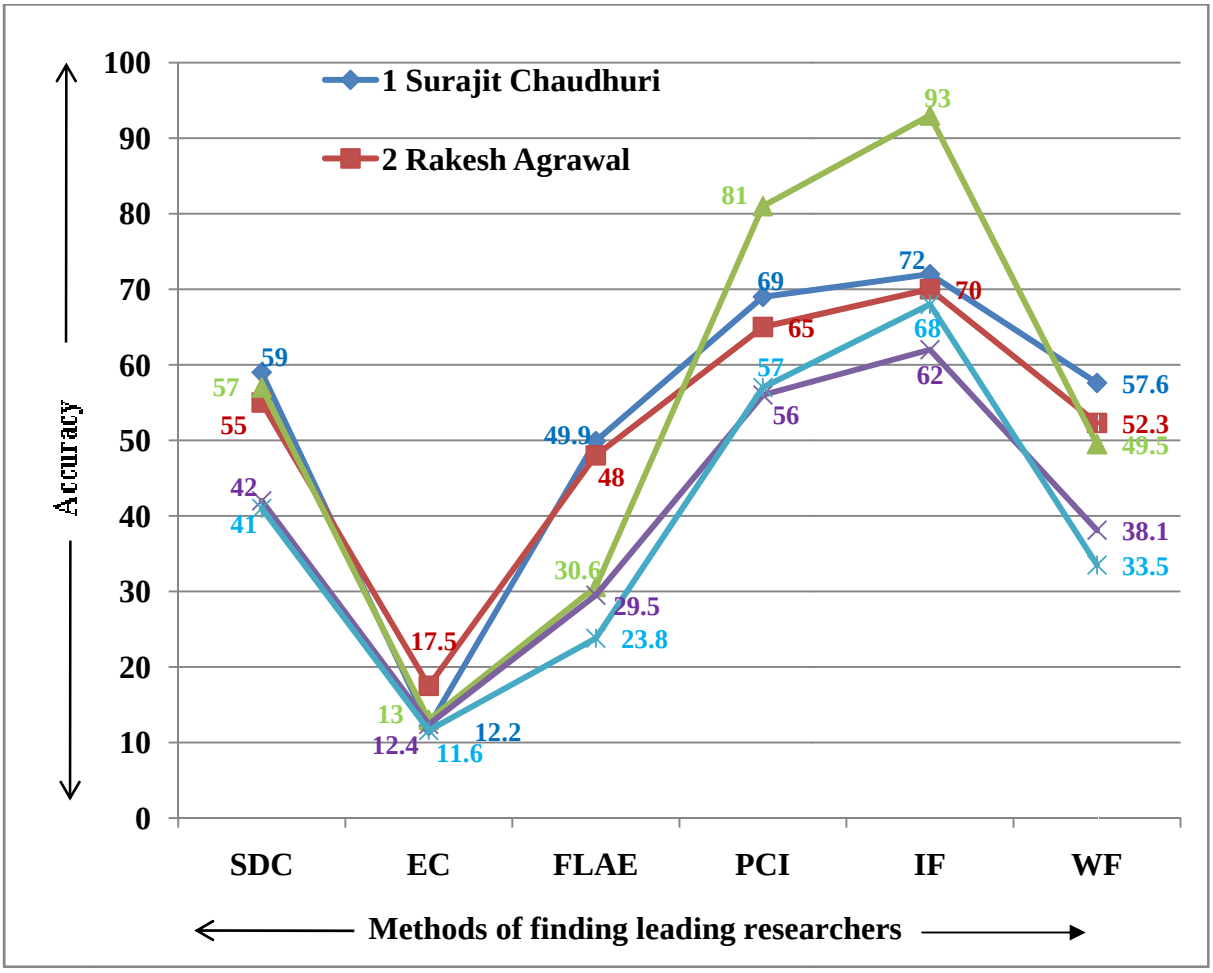

Figure 1.1 The representation of accuracies for the Proposed and Existing Systems.

The proposed framework strongly suggests the approaches for measuring the research contributions by authors in a research community. Although the Weighted Frequency method considers each position of the researcher, the Weighted Frequency (WF) calculation does not assign any repeated impact for more than one position for the researcher. The existing systems assigned the same weights for more than one positions. Therefore, the proposed system is more efficient and useful for the research society when compared with other existing systems.

IV CONCLUSION

The proposed framework finds the leading research contributors based on the individual frequency and the weighted frequency. This system produces the better and reasonable accuracy compare than existing system. The proposed framework strongly suggests the approaches for measuring the research contributions by authors in a research community. Although the Weighted Frequency method considers each position of the researcher, the Weighted Frequency (WF) calculation does not assign any repeated impact for more than one position for the researcher. The existing systems assigned the same impact factor values for more than one positions. Therefore, the proposed system produces the efficient and reasonable accuracy for the research and academic society compare than other existing systems.

References:

[1] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.
[2] T. Tscharntke, M. E. Hochberg, T. A. Rand, V. H. Resh, and J. Krauss, “Author sequence and credit for contributions in multiauthored publications,” PLoS Biol., vol. 5, no. 1, pp. 0013–0014, 2007.
[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] https://meshb.nlm.nih.gov/MeSHonDemand
[5] https://www.sas.com/en_us/software/university-edition/download-software.html
[6] https://support.sas.com/en/support-home.html
[7] https://www.ultraedit.com
[8] M. T. Rahman, J. Mac Regenstein, N. L. A. Kassim, and N. Haque, “The need to quantify authors’ relative intellectual contributions in a multi-author paper,” J. Informetr., vol. 11, no. 1, pp. 275–281, 2017.
[9] Tasleem Arif, “Exploring the Use of Hybrid Similarity Measure for Author Name Disambiguation,” International Journal of Science and Technology Research, vol. 4, no. 12, pp. 171–175, 2015.
[10] J. M. Warrender, “A Simple Framework for Evaluating Authorial Contributions for Scientific Publications,” Sci. Eng. Ethics, vol. 22, no. 5, pp. 1419–1430, 2016.