Improving DBLP Efficiency through Social Media Mining

Faryal Shamsi, Irum Sindhu
Department of Computer Science
Sukkur IBA University
Sukkur, Sindh, Pakistan
{faryal.shamsi@iba-suk.edu.pk} {irum.sindhu@iba-suk.edu.pk}

ABSTRACT: *Social media mining provides a system to extract meaningful patterns from rapidly growing social networks. DBLP is a well-known dataset of the computer science bibliography, which gathers the computer science research community under single umbrella. The dataset provides millions of records, which include research publications journals, conferences and author information. This study intends to understand the social network of authors, by utilizing the records available in DBLP. This will help in recognizing the same author appearing with variation in names. For example, John F. Kennedy may appear in different variations, like – John Fitzgerald Kennedy, John Kennedy, J.F Kennedy or even JFK. Because of these variations DBLP is unable to produce correct information about the research index and publications of an author. Understanding the social network of John F. Kennedy will help us in recognizing that, these different names are referring to the same author thus improving overall efficiency of DBLP.*

Keywords: DBLP, Author Names, Database Study

Received: 18 July 2021, Revised 30 August 2021, Accepted 8 September 2021

DOI: 10.6025/ijwa/2021/13/4/97-104

1. Introduction

DBLP is an online database hosting service, established in 1993 at University of Tier, Germany. This service provides a platform to the computer science research community worldwide and allows to explore the computer science bibliography, with more than 3.6 million records. The records include a comprehensive information about publications in the field of computer sconce from 1995 till present [6]. The publication information includes the name of publisher (i.e. Journal or Conference), volume, date and year of publication, number of pages, the URL address, list of authors and the link to profile of the authors [6]. The research work is being published, especially in the field of computer science, with such a huge volume that it becomes almost impossible task to calculate the research index of an author, manually. For that purpose, several indexing services and applications are providing a platform to the researchers to explore the publication information of each other. Unfortunately, the DBLP dataset is unable to produce correct information with respect to the author names, as it relies on how the name of author is spelled. Therefore, a minor variation in the spelling of author name makes a huge impact on his or her research index. There are several author level metrics used to calculate the overall productivity of an author as a researcher. Two major indicators of a researcher's productivity are number of research publications and citations. Practically, the huge number of publications does not ensure the quality of the work while the huge number of citations can be achieved by a single article [7]. Therefore, to calculate impact of an author in research community, h-index is widely acceptable measure [1] which involves both aspects – The number of publications and its citations[4]. The h-index does not only play a major role in accessing an author's profile, but is used as an important measure by the Academic Universities worldwide to access the quality aspects of an author. Academic institutions have to maintain a significant number of research publications to ensure their worth and ranking to compete with other institutes.

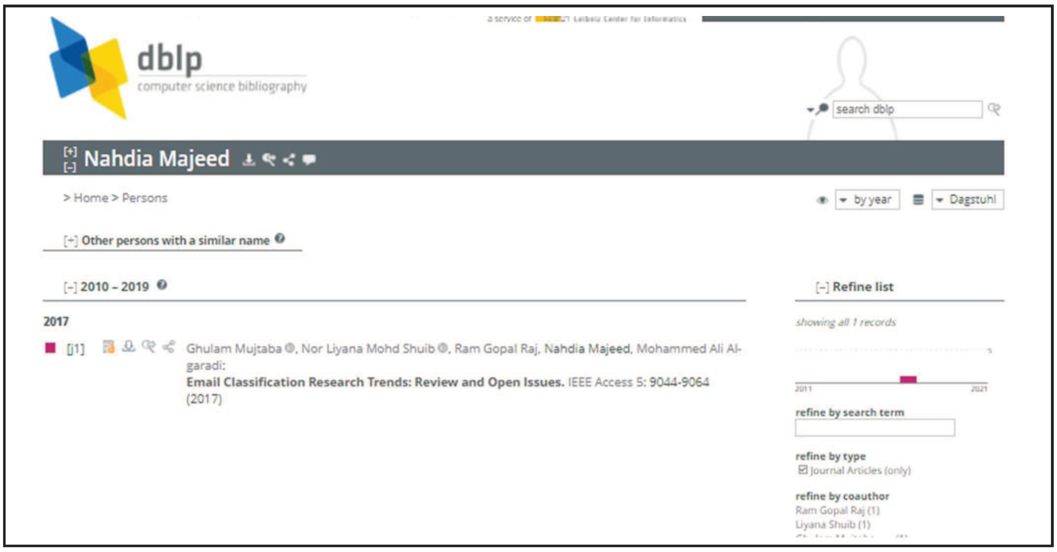
2. Problem Statement

To understand the problem clearly, let's consider a realworld case, where DBLP will display following records for an author. The author have contributed 13 research publications. As, DBLP application recognizes the author by the spelling of name, therefore it fails to display all publications under his name. The problem is that, the same author have used his name with a spelling variation as shown in Figure 3 and Figure 4 respectively. Another example can be seen in Figure 1 and Figure 2 where name inconsistency happened due to surname change of female author named ”Nahdia Majeed” who became ”Nahdia Majeed Lodhro” after marriage. Figure 5 illustrates 3 different spelling used by author ”Muhammad Ajmal Sawand” and DBLP is not able to recognize it as same person. These are some examples, but the DBLP application demonstrates imperfect results for many other authors as well. Therefore DBLP needs an alternative mechanism to identify the author regardless of the variations in spellings of name but considering the social network information already available, in the dataset. The objective of the work as implied by the title, is to improve the overall efficiency of DBLP by removing some of its imperfections. The DBLP system typically recognizes an author by its name. This fact leads to various problems –

1) Many authors may have same name, it causes ambiguity

2) Same authors may use different spellings, thus considered as different persons

3) There is no mechanism to track changes in author details where, author may change name due to religious conversion, addition of prefixes (e.g. Dr. after completion of PhD) or female surname changes after marriage


a service of Leibniz Center for Informatics
dblp
computer science bibliography
search dblp

Nahdia Majeed

> Home > Persons

by year

Dagstuhl

[+] Other persons with a similar name

[–] 2010 – 2019

2017

[j1] Ghulam Mujtaba, Nor Liyana Mohd Shuib, Ram Gopal Raj, Nahdia Majeed, Mohammed Ali Al-garadi:
Email Classification Research Trends: Review and Open Issues. IEEE Access 5: 9044-9064 (2017)

[–] Refine list

showing all 1 records

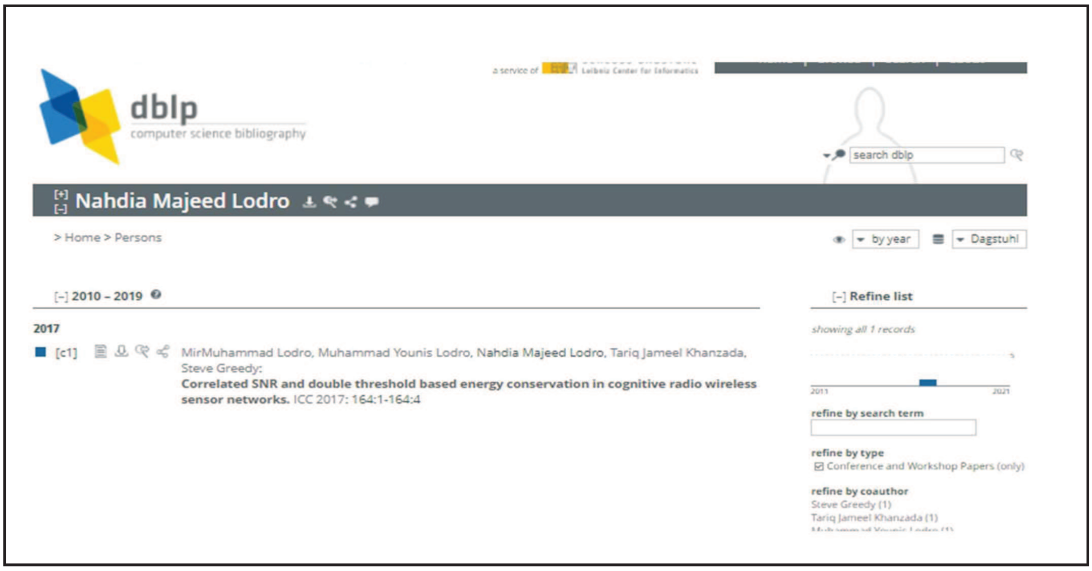
refine by search term

refine by type
☑ Journal Articles (only)

refine by coauthor
Ram Gopal Raj (1)
Liyana Shuib (1)

Figure 1. DBLP Display Records for ”Nahdia Majeed”

To address the above stated problems, this study aims to identify an author within DBLP by utilizing his or her social network, rather than the spelling of name. If observed closely, the DBLP dataset provides some significant information about authors of the research papers. Features like: co-authors information, date and year of publications, the title and keywords related to the research to track the author's field of specialization can mine the similarities and differences of an author regardless of his or her name spellings.



a service of Schloss Dagstuhl Leibniz Center for Informatics
dblp
computer science bibliography
search dblp

Nahdia Majeed Lodro

> Home > Persons

by year | Dagstuhl

[–] 2010 – 2019

2017

[c1] MirMuhammad Lodro, Muhammad Younis Lodro, Nahdia Majeed Lodro, Tariq Jameel Khanzada, Steve Greedy:
Correlated SNR and double threshold based energy conservation in cognitive radio wireless sensor networks. ICC 2017: 164:1-164:4

[–] Refine list

showing all 1 records

2011 2021

refine by search term

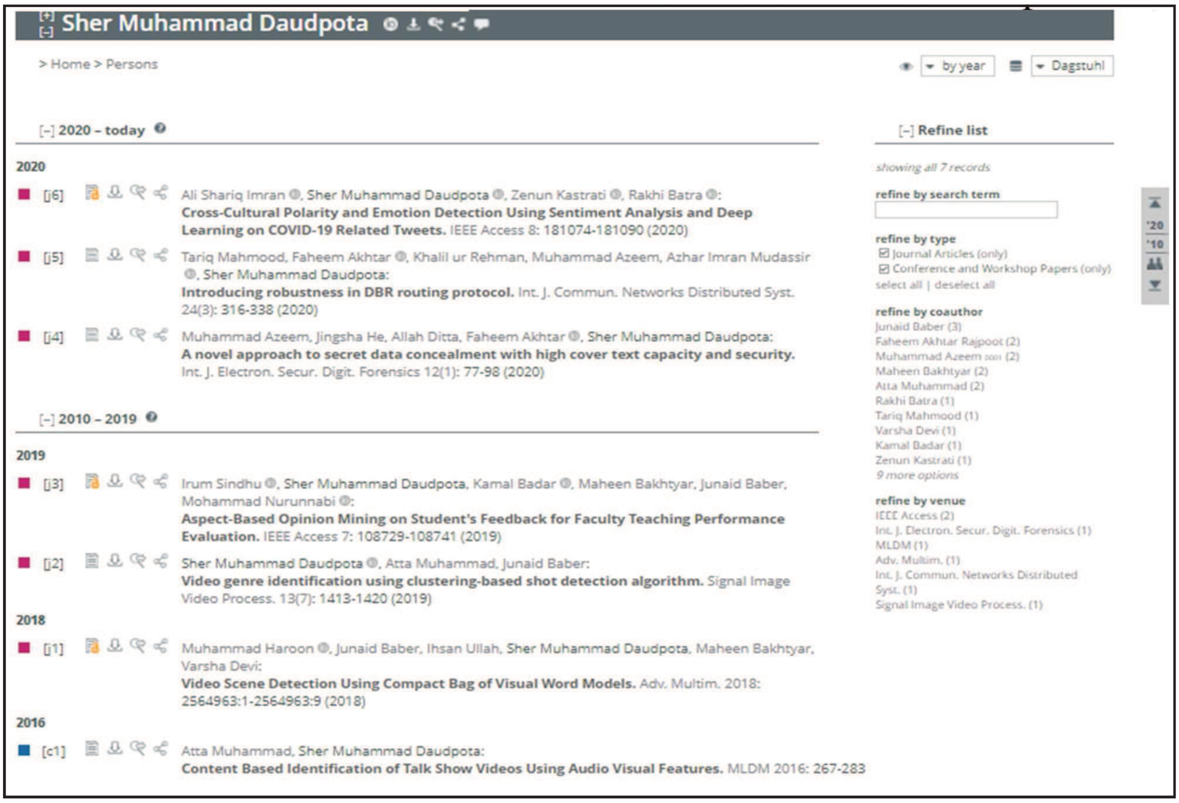
refine by type
☑ Conference and Workshop Papers (only)

refine by coauthor
Steve Greedy (1)
Tariq Jameel Khanzada (1)
Muhammad Younis Lodro (1)

Figure 2. DBLP Display Records for ”Nahdia Majeed Lodhro”

3. Literature Review

The DBLP is a very imperfect dataset. Michael Ley, the leader of DBLP team himself admits its imperfections in [8].



Sher Muhammad Daudpota

> Home > Persons

by year | Dagstuhl

[–] 2020 – today

2020

[j6] Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Rakhi Batra:
Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets. IEEE Access 8: 181074-181090 (2020)

[j5] Tariq Mahmood, Faheem Akhtar, Khalil ur Rehman, Muhammad Azeem, Azhar Imran Mudassir, Sher Muhammad Daudpota:
Introducing robustness in DBR routing protocol. Int. J. Commun. Networks Distributed Syst. 24(3): 316-338 (2020)

[j4] Muhammad Azeem, Jingsha He, Allah Ditta, Faheem Akhtar, Sher Muhammad Daudpota:
A novel approach to secret data concealment with high cover text capacity and security. Int. J. Electron. Secur. Digit. Forensics 12(1): 77-98 (2020)

[–] 2010 – 2019

2019

[j3] Irum Sindhu, Sher Muhammad Daudpota, Kamal Badar, Maheen Bakhtyar, Junaid Baber, Mohammad Nurunnabi:
Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation. IEEE Access 7: 108729-108741 (2019)

[j2] Sher Muhammad Daudpota, Atta Muhammad, Junaid Baber:
Video genre identification using clustering-based shot detection algorithm. Signal Image Video Process. 13(7): 1413-1420 (2019)

2018

[j1] Muhammad Haroon, Junaid Baber, Ihsan Ullah, Sher Muhammad Daudpota, Maheen Bakhtyar, Varsha Devi:
Video Scene Detection Using Compact Bag of Visual Word Models. Adv. Multim. 2018: 2564963:1-2564963:9 (2018)

2016

[c1] Atta Muhammad, Sher Muhammad Daudpota:
Content Based Identification of Talk Show Videos Using Audio Visual Features. MLDM 2016: 267-283

[–] Refine list

showing all 7 records

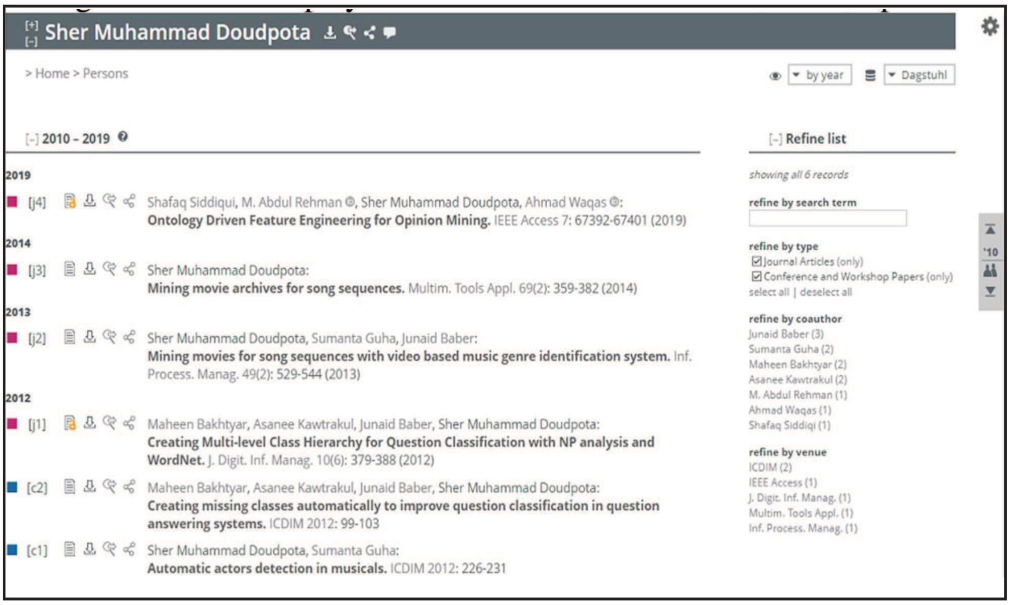
refine by search term

refine by type
☑ Journal Articles (only)
☑ Conference and Workshop Papers (only)
select all | deselect all

refine by coauthor
Junaid Baber (3)
Faheem Akhtar Rajpoot (2)
Muhammad Azeem 0001 (2)
Maheen Bakhtyar (2)
Atta Muhammad (2)
Rakhi Batra (1)
Tariq Mahmood (1)
Varsha Devi (1)
Kamal Badar (1)
Zenun Kastrati (1)
9 more options

refine by venue
IEEE Access (2)
Int. J. Electron. Secur. Digit. Forensics (1)
MLDM (1)
Adv. Multim. (1)
Int. J. Commun. Networks Distributed Syst. (1)
Signal Image Video Process. (1)

Figure 3. DBLP Display Records for “Sher Muhamamd Daudpota”



Sher Muhammad Doudpota

> Home > Persons

[–] 2010 – 2019

2019

■ [j4] Shafaq Siddiqui, M. Abdul Rehman, Sher Muhammad Doudpota, Ahmad Waqas:
Ontology Driven Feature Engineering for Opinion Mining. IEEE Access 7: 67392-67401 (2019)

2014

■ [j3] Sher Muhammad Doudpota:
Mining movie archives for song sequences. Multim. Tools Appl. 69(2): 359-382 (2014)

2013

■ [j2] Sher Muhammad Doudpota, Sumanta Guha, Junaid Baber:
Mining movies for song sequences with video based music genre identification system. Inf. Process. Manag. 49(2): 529-544 (2013)

2012

■ [j1] Maheen Bakhtyar, Asanee Kawtrakul, Junaid Baber, Sher Muhammad Doudpota:
Creating Multi-level Class Hierarchy for Question Classification with NP analysis and WordNet. J. Digit. Inf. Manag. 10(6): 379-388 (2012)

■ [c2] Maheen Bakhtyar, Asanee Kawtrakul, Junaid Baber, Sher Muhammad Doudpota:
Creating missing classes automatically to improve question classification in question answering systems. ICDIM 2012: 99-103

■ [c1] Sher Muhammad Doudpota, Sumanta Guha:
Automatic actors detection in musicals. ICDIM 2012: 226-231

[–] Refine list

showing all 6 records

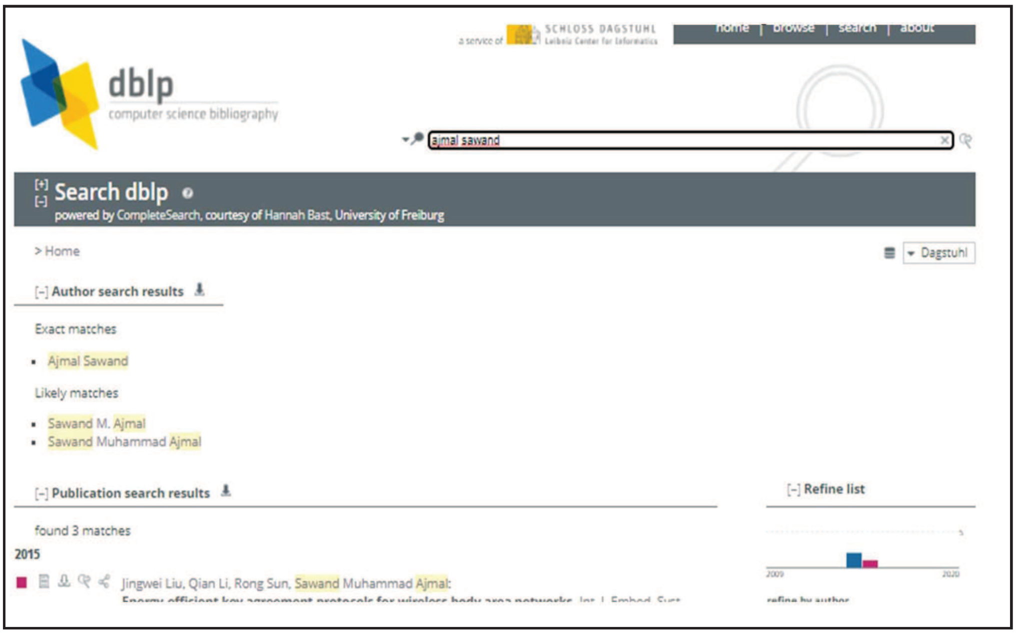
refine by search term

refine by type
☑ Journal Articles (only)
☑ Conference and Workshop Papers (only)
select all | deselect all

refine by coauthor
Junaid Baber (3)
Sumanta Guha (2)
Maheen Bakhtyar (2)
Asanee Kawtrakul (2)
M. Abdul Rehman (1)
Ahmad Waqas (1)
Shafaq Siddiqi (1)

refine by venue
ICDIM (2)
IEEE Access (1)
J. Digit. Inf. Manag. (1)
Multim. Tools Appl. (1)
Inf. Process. Manag. (1)

Figure 4. DBLP Display Records for ”Sher Muhamamd Doudpota”


a service of SCHLOSS DAGSTUHL Leibniz Center for Informatics
home | browse | search | about
dblp
computer science bibliography
[+]
[-]
▾ Dagstuhl
2009
2020

ajmal sawand

Search dblp
powered by CompleteSearch, courtesy of Hannah Bast, University of Freiburg

> Home

[–] Author search results

Exact matches

- Ajmal Sawand

Likely matches

- Sawand M. Ajmal
- Sawand Muhammad Ajmal

[–] Publication search results

found 3 matches

2015

■ Jingwei Liu, Qian Li, Rong Sun, Sawand Muhammad Ajmal:
Energy efficient key agreement protocols for wireless body area networks. Int. J. Embed. Syst.

[–] Refine list

refine by author

Figure 5. DBLP Display Records for Author name ”Ajmal Sawand”

To understand and overcome the problems of DBLP, it is very necessary to understand the structure and evolution of DBLP. Keeping in view, the fact that social media data is not often easy to mine, due to its huge volume and unstructured/semi-structured nature.

3.1. Graph-based Community Detection using Complex Network

The issue of name variation previously was also raised by [10]. The prescribes mechanism for automatically correcting author names to cope up with the inconsistencies. The corrections were divided into 4 categories – 1. Rename – To replace the mistaken (or spelling of the) name with correct one. 2. Merge – To replace all the variations of name and spelling with a correct version of name. 3. Split – To separate the identities of two author names with same spelling, but are different persons in the real world. 4. Distribute – To distinguish between a set of author names with same spelling, but all correspond to a different individual in real world and associated with different publications. A prototype of a complex network can originate from the co-author relationships in scientific research community [23]. Such type of graph based network evolution procedure is applied on the DBLP dataset by [24]. The "i-graph" tool is explored by [29] for the development of complex network from a given dataset, while [28] [30] have proposed some unique measures to identify the similarity between the two nodes. Unfortunately could not provide a foolproof method for resolve the issue of name consistencies.

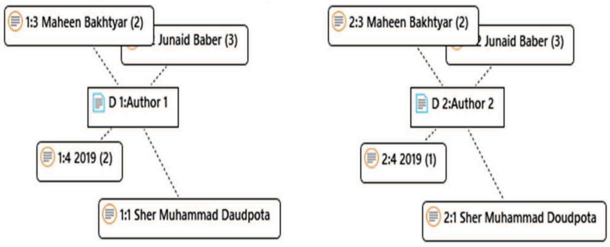


Figure 6. Node Similarity between Author1 and Author2

3.2. Random Walk Algorithm

An algorithm is proposed in [9] to mine the social network information available at DBLP. The proposed study implements graph based community detection approach, and named it as random walk, to extract the information about relevant authors, conferences and research topics with respect to an author. An application named 'DBConnect' was also developed, not only to implement the community mining but also to extend the semi-structured nature of DBLP data to the relational structure to be used by other applications. This application requires a run time evolution of a complex network that may be time consuming and requires a lot of resources across the internet. Furthermore, for some unknown reasons the proposed strategy is still in applicable to DBLP that this online bibliography service still suffers from such type of inconsistencies as shown in Figure 3 and 4. In [3] and [2] algorithms are proposed to improve the search using DBLP filter and to find expert in a given field respectively targeting the DBLP dataset.

3.3. Time-constrained Probabilistic Factor Graph Model (TPFG)

Another interesting algorithm proposed by [5] to mine the advisor-advisee relationship from the DBLP dataset. The mining was performed by implementing Time-constrained Probabilistic Factor Graph Model (TPFG) on the DBLP data. This model generates a tentative advisor-advisee level relationship graph with maximum possibilities. Then [5] had developed a leaning algorithm to optimize the results of TPFG and delivers the advisor-advisee relationship graph with increased efficiency. The problem with this approach is that there is any existing dataset available to evaluate the accuracy of the proposed system. Therefore the tentative relationships acquired by this system can be be generalized and implied as generalized information.

4. Proposed Solution

The recent form of DBLP has evolved from a very small XML file, which has now reached to a size of 2GB with more than 3.6 million records [8]. The contents of this file are as follows –

1) Article – The root element of the DBLP.xml file, corresponding to each publication.

a) Key – The first attribute of ¡article¿ which is a unique identifier of each record.

b) Mdate – The second attribute of ¡article¿ with last modification of the article file.

2) Author – This element is repeated 1 to 5 times within the ¡article¿. The name of each author is enclosed in the ¡author¿ element

3) Title – Gives the title of each publication

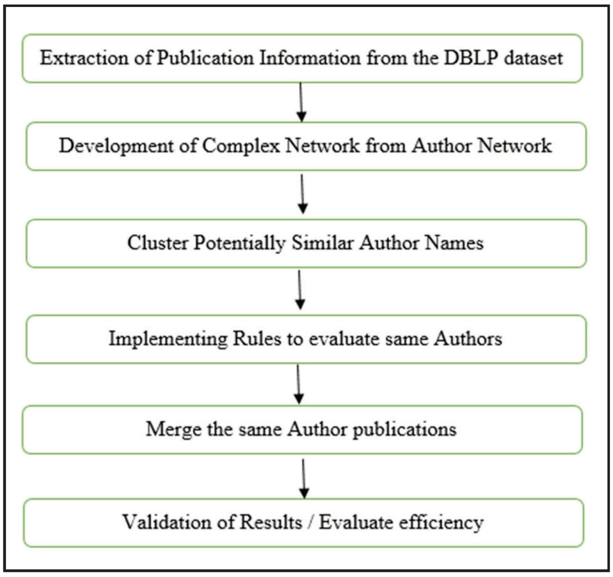

Extraction of Publication Information from the DBLP dataset
Development of Complex Network from Author Network
Cluster Potentially Similar Author Names
Implementing Rules to evaluate same Authors
Merge the same Author publications
Validation of Results / Evaluate efficiency

Figure 7. DBLP Improvement through Complex Network Analysis Flow Methodology

4) Pages – The number of pages the publication covers. The format is like starting page number – ending page number (i.e. 513-526).

5) Year – This field is crucial in case of conferences, sometimes the year when the conference was hosted is different from the year when the proceedings of the conference were published. This field holds a numeric value referring to the year when the paper was published in the conference.

6) URL & EE – The location at the internet where the publication is available.

The DBLP dataset provides comprehensive information about the co-author network. To handle the issue of name inconsistencies, the first step is to extract the required features (as shown in Table 2) of each author to be represent as node in a complex network. A complex network is a graph with non-trivial features, and used to model a real world systems. Such type of graph have some random as well as regular connections, which makes it different from a traditional graph. To remove the mentioned imperfections of DBLP and the objectives of this study, The node similarity will be calculated where each author in DBLP represented with a unique in the complex network. For clear understanding, let's consider the an author represented by DBLP, which is by 2 different node due to variation in spelling of name as discussed in Section 1. The node similarity calculation is illustrated in equation. 2 and 1. Where, N_1 and N_2 is the object of nodes representing $Author_1$ and $Author_2$ respectively. F_1, F_2 and F_3 are the features of author as shown in Table. 2.

$$\sum_{i=1}^{3} F_i = Similarity(F_i Author1, F_i Author2) \qquad (1)$$

$$D(N_1, N_2) = F_1 + F_2 + F_3 \qquad (2)$$

In the above case, we have some interesting information about author's names along with the spellings of name, which can be utilized to evaluate the similarity between any two nodes. For example number of co-authors, and date of publishing of the article. A suitable data mining model will be proposed, to track the similarity automatically by the DBLP, resulting with increased efficiency in the overall system. The similarity between the spellings of name is equal to 0.95 (on basis of string matching) while the similarity on basis of number of co-authors of node 1 and node 6 is 0.75, as 2 of the three authors are same. The similarity on basis of the time period in which the two authors have made the publication is almost same and is in the same discipline of computer science. All these types of similarities, when aggregated gives the similarity *index* = 0.84. This

Co-authors of Sher Muhammad Daudpota	Co-authors of Sher Muhammad Doudpota
Junaid Baber (3)	Junaid Baber (3)
Faheem Akhtar Rajpoot (2)	Sumanta Guha (2)
Muhammad Azeem (2)	Asanee Kawtrakul (2)
Maheen Bakhtyar (2)	Maheen Bakhtyar (2)
Atta Muhammad (2)	M. Abdul Rehman (1)
Rakhi Batra (1)	Shafaq Siddiqi (1)
Tariq Mahmood (1)	Ahmad Waqas (1)
Varsha Devi (1)	
Kamal Badar (1)	
Zenun Kastrati (1)	
Khalil ur Rehman (1)	
Jingsha He (1)	
Azhar Imran Mudassir (1)	
Ihsan Ullah (1)	
Mohammad Nurunnabi (1)	
Allah Ditta (1)	
Ali Shariq Imran (1)	
Irum Sindhu (1)	
Muhammad Haroon (1)	
Publish years of Sher Muhammad Daudpota	**Publish years of Sher Muhammad Daudpota**
2020 (3)	2019 (1)
2019 (2)	2014 (1)
2018 (1)	2013 (1)
2016 (1)	2012 (3)

Table 1. Author

Feature id	Required Features
F_1	Author Name
F_2	List of Co-authors
F_3	List of Years of

Table 2. Feature Extraction from DBLP Record

value is close the 1 suggesting that the two nodes are very similar. The minor difference can be also be removed if the DBLP dataset includes another important attribute, which it is currently missing. This attribute is author affiliation, if DBLP adds this attribute the efficiency of DBLP can be improved to a significant level. the node similarity is plotted in Figure. 6.

5. Conclusion

In this study the objective was to understand the social network of authors, by utilizing the records available in DBLP. This will help in recognizing the same author appearing with variation in names. The DBLP provides a comprehensive dataset of author information but doesn't utilizes it effectively and sometimes, maintains two different records of same author just because of minor variation in the spelling of names. Because of this problem DBLP is unable to produce correct information about the research index and publications of an author. Mining the social network of authors, rather than considering the name spelling will help DBLP to recognize different names are referring to the same author. Thus, this paper not only intend to improve overall efficiency of DBLP as Web Application, but it will also help authors to be recognised for their scientific contributions in a better way.

References

[1] Burrows, Roger. (2012). Living with the h-index? metric assemblages in the contemporary academy. *The sociological review*, 60 (2), 355–372, 2012.

[2] Deng, Hongbo., King, Irwin., Lyu, Michael R. (2008). Formal models for expert finding on dblp bibliography data. *In 2008 Eighth IEEE International Conference on Data Mining*, pages 163–172. IEEE.

[3] Du, Jiang., Jin, Peiquan., Zheng, Lizhou., Wan, Shouhong., Yue, Lihua. (2014). Dblp-filter: effectively search on the dblp bibliography. *In*: Proceedings of the 23rd International Conference on World Wide Web, pages 255–256.

[4] Harzing, Anne-Wil., Wal, Ron Van der. (2009). A google scholar h-index for journals: An alternative metriLihua c to measure journal impact in economics and business. *Journal of the American Society for Information Science and technology*, 60 (1), 41–46.

[5] Hazelkorn, Ellen. (2015). Rankings and the reshaping of higher education: The battle for world-class excellence. Springer.

[6] Ley, Michael. (2008). Dblp computer science bibliography. http://dblp. uni-trier.de/.

[7] Lindgren, Lena. (2011). If robert merton said it, it must be true: A citation analysis in the field of performance measurement. *Evaluation*, 17 (1), 7– 19, 2011.

[8] Reitz, Florian., Hoffmann, Oliver. (2013). Learning from the past: An analysis of person name corrections in the dblp collection and social network properties of affected entities. *In*: The influence of technology on social network analysis and mining, pages 427–453. Springer.

[9] Wang, Chi., Han, Jiawei., Jia, Yuntao., Tang, Jie., Zhang, Duo., Yu, Yintao., Guo, Jingyi. (2010). Mining advisor-advisee relationships from research publication networks. *In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 203–212.

[10] Zaiane, Osmar R., Chen, Jiyang., Goebel, Randy. (2007). Dbconnect: mining research community on dblp data. *In*: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 74–81.