IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

Discoveries of Research Genealogy from Large-Scale Academic Dataset: Issues, Challenges and Application

Sovan Bhattacharya[1*]

[1]Department of Computer Science & Engineering, National Institute of Technology, Durgapur, India

[*]*Corresponding Author: sovan.cse@gmail.com*

Available online at: www.ijcseonline.org

***Abstract*—** Genealogical research is the tracing of an individual's ancestral history using historical records, both official and unofficial. Challenges about genealogy problem like spelling names, legacy of a researcher can be measured not only in terms of his/her publications and scientific discoveries, in terms of the formation of other researchers. Now, research work is improving than oldest research. So population of researcher and scientist is increasing rapidly and it was more important now a days that to finding out who is better among all researcher. Author ranking can be solved this problem. Author ranking will not be perfect due to some causes, like naming disambiguation problem and uses of multiple name in paper. In Academic genealogy, is the relationship between advisor and advisee. Research area of advisor is more popular than his advisee research area may be good. From there we can do future prediction of an author. Another problem of author name disambiguity can be solved using genealogy tree hierarchy, as there are less chances of conflict in identifying an author based on his unique academic records. Another important challenge is that how much level (generation) we can visit from the genealogy tree. From the big dataset, we extract different metrics for an author. In this paper, we extract data of a particular author and from there we have analyze effects of an author rank.

***Keywords*—** Genealogy tree, Author name disambiguation, Citation

I. INTRODUCTION

A genealogy is a record or more specially table of the descent of a person and his family from an ancestor. When record will represent as tree then its call genealogy tree. Genealogy sometimes known as family history but some offer a slight difference in their definition. Genealogy is a work done to discover ancestors and descendants; it is the study of biological/ genetic decent. However, family history is the study of all aspects of a family's history. Therefore, in a single sentence genealogy expresses one's own aspects. Different challenges like spelling names, legacy of a researcher can be measured in genealogical research of publications and scientific discoveries. The h-index metric was proposed to measure a researcher's scientific output. This calculation is quite simple. Whereas it is calculated based on most cited publications and the number of citations they have received. A researcher has an h-index h means if he/she has at least h publications that have received at least h citations. A researcher has a genealogy h-index h means if he/she has at least h advisees and, at least one of them, has advised at least h advisees as well.

A professional genealogist can help you trace your ancestors. For example, a genealogist may be able to discover who your immigrant ancestors were and where they came from. A genealogist can research one of your family lines back to a specific time or individual. In philosophy, genealogy is a technique in which it can understood emergence of various philosophical and social beliefs. Genealogical research is the tracing of an individual's ancestral history using historical records, both official and unofficial.

In research, collaboration is teaching and mentoring relationship for students. As academic genealogy is formed as a tree or a table, it helps us to find out or trace the roots and leaves of the tree. Analyzing the academic graph or tree, we can find out multiple advisors and their advices in a simple application. Academic genealogy helps to create a tree and clarifies the relation between scholars and their mentors. While genealogy creates a family tree, academic genealogy is also a family tree of scholars according to mentoring relationships. More clearly, it is a sequence, which also can create a tree about advisor-advisee relationship. Now a day this advisor-advisee relationship is specifically seen among the Ph.D. Scholars and their mentors.

These technologies can be used in genealogy as an aid in finding and evaluating records and resources. This wiki article is a form of social media, due to its open edit nature and discussion feature. Knowing the purpose or benefits of each tool will help you to use them in your genealogy search. Social relationship between the researcher and his mentor are established from academic network. This making genealogy

© 2019, IJCSE All Rights Reserved

relational tree will resolve the problem of author name disambiguation. We categorised this relation into different part strongly, mediator and weakly connected. The root of the tree is strong then we can predict the strength of leaves will be strong. The same idea is applicable in the research community. In Academic genealogy, is the relationship between advisor and advisee. Research area of advisor is more popular than his advisee research area may be good. The level of the tree or generation is vital in this tree. If the level of tree increased then edges of the level will be decreased gradually. Therefore, the advisor and advisee they will have closest then it is strongly connected. After proper judgement of author's actual ranking then we can find out future prediction of an author.

Ambiguity is uncertainty of an object. Ambiguity is meaning in which several interpretations are there. It cannot be definitively resolved according to a rule or process with a finite number of steps. The concept of ambiguity is generally create problem. Context may play a role in resolving ambiguity. Different causes of ambiguity can produced. Lexical ambiguity can be addressed by algorithmic methods that finding the meaning with a word in context, this is called word sense disambiguation. Syntactic ambiguity is when a sentence can have two different meanings. It will modified as expression. Semantic ambiguity occurs when a sentence contains an ambiguous word or phrase word or phrase that has more than one meaning.

So many challenging problem occurs in ambiguity like resource management, information of knowledge production process, handling risk and uncertainty; it is also problem of business review. Ambiguity is also interesting because it marks a tension between the disciplines and the methodologies they employ. Ambiguous is important problem and challenges related to social media and ethics presented in this piece are implicitly focused on social work practice. Social media and social work: (1) It is used professionally with social media related like role of the student within the organization; (2) It related to personal and professional use of social media, and (3) All over implications for the social work profession related to risk management and ethical practice. Authors of scholarly documents they share names which makes it is too much problematic to distinguish each specific author's work.

In this project our focus are listed below.

1. Genealogy tree constructed from the large MAG dataset and we are finding different matrices with respect to features.
2. We have to established different relationship among scientist of academic research and after we have to see this relationship can be obliged how long means generation.
3. Is it possible that author name disambiguation problem solved by genealogy tree?
4. After resolved author, actual ranking then we look author future prospect prediction in research. Is the academic genealogy can resolved this problem.

Rest of the paper is organized as follows, Section I contains the introduction, Section II contain the related work of genealogy and author name disambiguation, Section III contain the some methodology, Section IV contain the results and discussion, section V explain the concludes research work with future directions.

II. RELATED WORK

2.1 Genealogy

This paper extract knowledge from graph by using different metrics, one of them is genealogical index, which can measure the relationship between advisor and advisee. This paper also established the correlation between this index metrics and bibliometric measurement like h-index, citation. It can influenced generation to generation and to measure knowledge transmission capacity. It can help to finding the origin and evolution of its disciplines in the present state of knowledge [6]. In this paper finding the set of metrics like identification, characterization, and classification of communities from there, it analysed mentoring relationship between mentors and scholar. This metrics also help to analyse a similarity graph and gather some knowledge like life cycle of the scientific topic and disciplines, emergence of specific area, lastly finds the pattern of a topic [5]. This project genealogical tree investigated how researchers determined the ownership of collaborative project and how they determined knowledge, skills, or resources from the collaborator could provide the order of authorship on collaborative nature [8]. This paper create genealogical research tree from their crawled data in NDLTD. Finding the different structure of academic formation and analyse the properties of the tree. Other new researcher can take other challenges from this genealogy tree [2].

2.2 Author name disambiguation

In this paper, evaluate the unique author ability by using regression model from the co-authorship network. This paper extract all the features statistically by measurement of quality of individual paper and to analyze author ability. This method cannot distinguish the seniority of an author increases when they are work together means inseparable co-authors [4]. NER model are introduced in author name disambiguation technique and all the features are extracted by text mining technique in Biomedical PubMed article collection. In this work have some limitation. In this proposed work did not compare with new training set and previously AND studies means semi-supervised and unsupervised learning. Another challenging task of author name disambiguation problem is same author have different name, can be justified [7]. In the large bibliographic dataset, author have find out about time based author name uses

© 2019, IJCSE All Rights Reserved

pattern like Rare, Swap, Co-occurrence are called synonym problem. It causes author name disambiguation mean it cannot separate author individually from this author set. From this dataset, naming pattern is extracted using supervised classification and then evaluate degree of an author from collaboration network structure. In this work have two-limitation first one frequently used two name only consider secondly it neglects complex name usage pattern [3]. This paper extracted data from webpages in structural way and applied graph based machine learning, and then formed clusters. Then observed F-score 0.95.This approach is had better perform for large dataset [1].

III. METHODOLOGY

In our work related on academic social network, where collected dataset are large amount. Actually this work about on author respective where some problem will arise now a days naming disambiguation problem. It is a big challenging problem. Another aspect for prediction on author's future. After survey of author name disambiguation problem, we have seen maximum cases they have applied supervised and unsupervised machine learning algorithm. However, I am trying to solve this two problem using graphical methods.

Actually this paper is represented the graphical relationship among authors each other's. Graph is the collection of set of vertex and edges where each vertex is represented as author and edges is the relationship among authors that is denoted by weight means how many times they have relationship each and others with respect to their published paper just like paper citation. We find out ranking index of individual author. For example, when we say ranking index of an author is R means R number of ancestor of author have minimum R ranking. We only consider the ancestor author is determined by its starting research period is earlier than victim author starting research period. Another issue, in this graphical representation we have to consider how much level is called generation.

In this paper, I will describe three problem i) Formation of different types of genealogy tree and finding genealogical index in generation to generation. ii) Author name disambiguation problem will resolved by author-genealogy tree. iii) Genealogical analysis will be predict author future.

Objective

1. To define a similarity profile that captures multiple aspects of similarity between author profiles.
2. Then we established relation among his co-author set we established mentoring relation among author.
3. From this mentoring relation we finding the origin vertex in this genealogy graph.
4. Find out the level up tree we can consider.

IV. RESULTS AND DISCUSSION

Citation network is formed by considering papers, authors and venues as node and relationship among them as edge. Different kinds of relationship occur among them, in the same time numbers of nodes can also be numerous. Therefore, we consider it as a complex network and try to justify our problem represented as genealogy tree structure. From this complex network we have extract different features tree. One of them is co-author genealogy tree that can help to solve different problem. We are finding genealogical index of the entire related author have relation with victim author. This victim author have faced naming ambiguity problem be resolved by this tree.

4.1 Data Collection

Over the last decade, there is an impressive growth in computational and storage advances. Open on line citation data sets are easily available. We perform all our proposed research on Computer Science bibliographic data sets. Most frequently used databases in the research community include Microsoft Academic Search (MAS), Arnetminer data set, DBLP, Web of Science, Scopus, Google Scholar etc. Various citation network dataset is available online to explore, among the arnetminer is quite popular for its open access but comparatively smaller than Microsoft Academic Search Bibliography dataset.

We downloaded two large academic graphs: Microsoft Academic graph (MAG) and AMiner data set, which was freely available .Total size MAG, papers 104 GB and AMiner papers 39 GB in zipped format on dated 22/03/2017. It consists of total 166,192,182 papers from MAG and 154,771,162 papers from AMiner publications.

Table 1. Feature Identification

Field	Field content
Paper id	Text string
Paper Title	Text string
Author	Text string
Affiliation	Text string
Year	Integers
Field of study	Text string
Venue	Text string
References	Text string

4.2 Feature Extraction and Network Modelling

Bibliographic data set plays an elemental role in undergoing our research study and finding appropriate solution for problems existing in the literature. With presence of diverse key entities such as papers, authors, venue, field of study, affiliation of an author, research groups, cities in our dataset; a colossal citation network is formed which conclusively could not be represented as a directed graph but forms a

© 2019, IJCSE All Rights Reserved

dynamic complex network. The relationship between several entities mentioned above are changing dynamically over time. We propose to do an extensive study of such a complex network built-layered modelling and extracting citation relationships to form genealogy tree between co-authors, venue and field of study.

From the large dataset we extract dictionary of author Zev Rosenwaks has total 862 publication in the year range of 1977 to 2017. This author work with 1010 number of co-author among them 579 co-author are single paper author. Figure 1 show that author Zev Rosenwaks have work with less number of co-author. All this co-author have greater than 50 number of publication.

4.3 Preliminary Results

Genealogical research field tree is building of relation between different fields that will extracted from co-authorship layer. In here, we say that which we field is more popular in research field and progressive today and can be predicted for future prospect.

In the same way we can formed venue-centric genealogy tree will more precise build relationship between venues. The citation can be help to forming this genealogical tree formation.

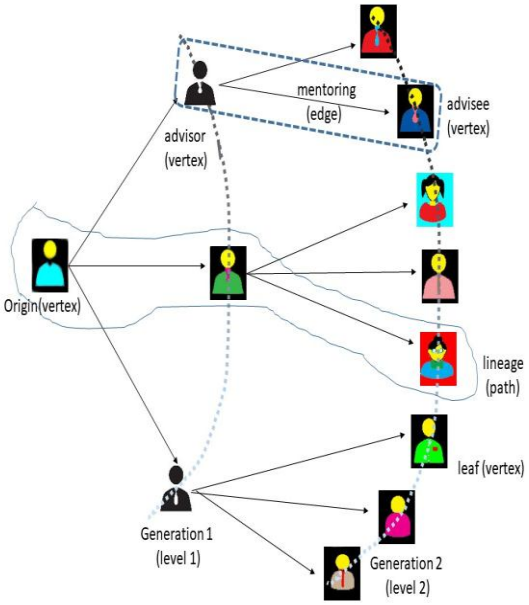

Figure 4: Co-Authorship Genealogy tree

After formation of all types of genealogy tree, we can predict the author's future prospect. For author future prediction, we have required three part, first one we have finding genealogical index of co-author in victim authors from co-authors genealogy tree. Second part, to find genealogical index of research field in victim authors. If these fields is popular and progressive then it will be good for an author otherwise not. The last one is venue centric genealogy that can help to say authors past publication records indexing is good for future prediction of victim author. After analysis all aspects then we easily say that victim author will be good for future.

4.4 DISCUSSION

4.4.1 Different types of Academic Genealogy tree Formation and Indexing

In the large amount of dataset, we have extracted some of the features like set of author, set of venues, and set of research fields and authors affiliation. When we established the relationship among author set and represented it in a graphical then it will form an academic genealogical graph. From this graph, we build relationship between advisor and advisee then it will form tree structure called genealogy tree. At each level of the tree is called generation to generation. Formally this graph represented as directed graph G is a pair of (V,E), Where V is a finite set of authors and E is the edges between vertices, edge is the relationship between authors,

When this edges are mentoring relation then this form graph to tree. In this tree have source vertex have level 0 and like this way next vertex is called generation 1, generation 2…., generation k. Mentoring edge is directed where advisor vertex directed to advisee vertex. Formally, length of victim vertex means from source to victim vertex length all are connected and make a path. When this vertex will be venue and edges means citation then this formation of tree is called venue-centric genealogy tree. Like this way, we have to create research field centric genealogy tree.

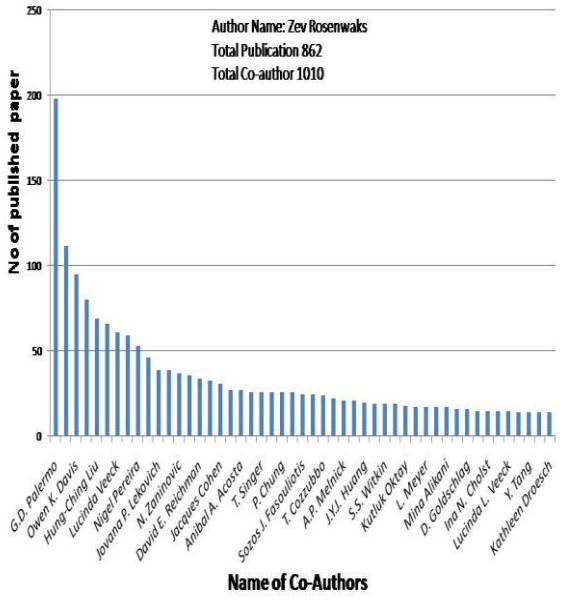

Fig 1: Author Zev Rosenwaks co-author vs no of publication

In figure 1: show that the author Zev Rosenwaks works individually co-authors wise no of publications

© 2019, IJCSE All Rights Reserved

4.4.2 Author Name disambiguation problem Resolved by genealogy tree

Authors of scholarly documents they have to share names, which makes it is too much problematic to distinguish each specific author's work. Author name disambiguation, also called as personal disambiguation. It is a one type of disambiguation and record linkage has applied to the names of individual people. An editor may apply the process to scholarly documents where the goal is to find all mentions of the same author and cluster them together. Authors of scholarly documents often share names, which makes it hard to distinguish each author's work. With the rapid growth of scientific research literature, day to day author name ambiguities has become a big challenging problem. There have multiple reasons that cause author names to be ambiguous, among them: from the different field like name multiple names have spelling, misspelling, name change due to marriage, or the use of middle names and initials due causes of ambiguation. Some author cannot maintain any universal standard for naming properly. There are two aspects of name ambiguity; one of them is synonymous name that means a single author with multiple name representation, another aspect is name homonym that is multiple authors sharing the same name representation.

Author name disambiguation has information about the authors such as their affiliations, email addresses, year of publication, co-authors, topic information to distinguish between authors. Effects on name ambiguity It not work to compares different similarity measures for publications or proposes new similarity measures. Most of Author Name Disambiguation approaches fall in different types of machine learning. To estimate author name similarity in a different style is applied.

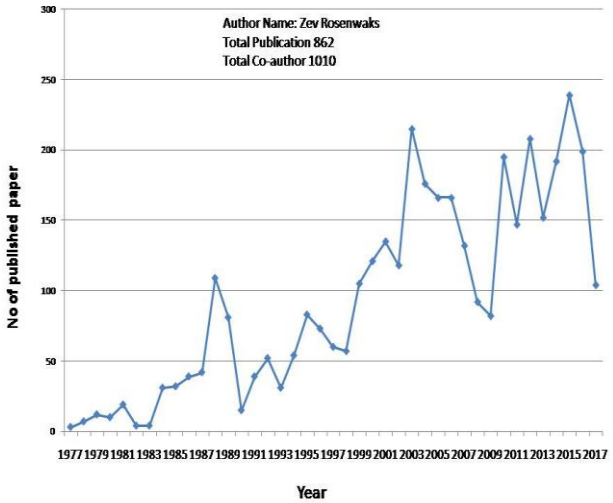

Fig 2: Author Zev Rosenwaks year vs no of publication

In figure 2: show that the author Zev Rosenwaks works from 1977 to 2017. Before 1987, this author has less 30 number of paper in per year but after 1987 per year publication increase gradually. However, it is never possible that author publication 200 in year. So, all the publication of the author is not same author that why the naming ambiguity occur in this publication.

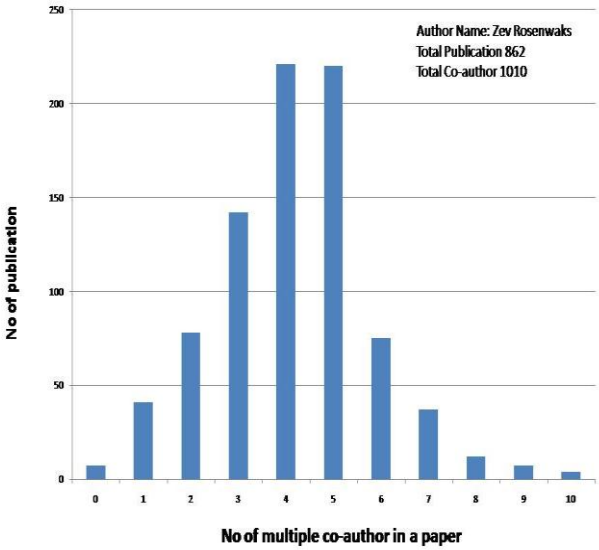

Fig 3: Author Zev Rosenwaks No of multiple co-author vs no of publication

In the figure 3 author Zev Rosenwaks works with how many co-author we have seen 25% of 4 or 5 multiple co-author paper are published. Less number of paper are single and large multi co-author paper. So maximum author have tendency that they are work medium number of co-author paper.

From the large dataset we extract dictionary of author Zev Rosenwaks. Author has total 862 publication in the year range of 1977 to 2017. This author work with 1010 number of co-author among them 579 co-author are single paper author. Figure 1 show that author Zev Rosenwaks have work with less number of co-author. All this co-author have greater than 50 number of publication.

V. CONCLUSION AND FUTURE SCOPE

Main objective of this research work is to forming different categories of genealogy tree and established their relation means edges. In the large amount dataset MAG and Aminer to be extracted and filter different field wise study like co-author set, field of study, and venue. From the co-author set we have finding author name disambiguation problem.

© 2019, IJCSE All Rights Reserved

1. To define a similarity profile that captures multiple aspects of similarity between author profiles.
2. Then we established relation among his co-author set we established mentoring relation among author.
3. From this mentoring relation we finding the origin vertex in this genealogy graph.

In our work related on academic social network, where collected dataset are large amount. Actually this work about on author respective where some problem will arise now a days naming disambiguation problem. It is a big challenging problem. Another aspect for prediction on author future. After survey of author name disambiguation problem, We have seen maximum cases they have applied supervised and unsupervised machine learning algorithm. However, I am trying to solve this two problem using graphical methods. In this paper, I will describe three problem

i) Formation of different types of genealogy tree and finding genealogical index in different generation.
ii) Author name disambiguation problem resolved by author-genealogy tree.
iii) Genealogical analysis will be predict author future.

REFERENCES

[1] Mehmet Ali Abdulhayoglu and Bart Thijs. 2017. Use of ResearchGate and Google CSE for author name disambiguation. Scientometrics 111, 3 (2017), 1965–1985.

[2] Wellington Dores, Fabrício Benevenuto, and Alberto HF Laender. 2016. Extracting academic genealogy trees from the networked digital library of theses and dissertations. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. ACM, 163–166.

[3] Janaína Gomide, Hugo Kling, and Daniel Figueiredo. 2017. Name usage pattern in the synonym ambiguity problem in bibliographic data. Scientometrics 112, 2 (2017), 747–766.

[4] Rasmus AX Persson. 2017. Bibliometric author evaluation through linear regression on the coauthor network. Journal of Informetrics 11, 1 (2017), 299–306.

[5] Luciano Rossi, Rafael JP Damaceno, Igor L Freire, Etelvino JH Bechara, and Jesús P Mena-Chalco. 2018. Topological metrics in academic genealogy graphs. Journal of Informetrics 12, 4 (2018), 1042–1058.

[6] Luciano Rossi, Igor L Freire, and Jesús P Mena-Chalco. 2017. Genealogical index: A metric to analyze advisor–advisee relationships. Journal of Informetrics 11, 2 (2017), 564–582.

[7] Min Song, Erin Hea-Jin Kim, and Ha Jin Kim. 2015. Exploring author name disambiguation on PubMed-scale. Journal of informetrics 9, 4 (2015), 924–941.

[8] Besiki Stvilia, Charles C Hinnant, Shuheng Wu, Adam Worrall, Dong Joon Lee, Kathleen Burnett, Gary Burnett, Michelle M Kazmer, and Paul F Marty. 2017. Toward collaborator selection and determination of data ownership and publication authorship in research collaborations. Library & Information Science Research 39, 2 (2017), 85–97.

Authors Profile

Mr. SOVAN BHATTACHARYA completed M.Tech from JADAVPUR UNIVERSITY in year 2013. He is currently pursuing Ph.D in Computer Science & Engineering from NIT Durgapur since 2016. He is a member of ACM since 2018. He has published three research papers in reputed international journals and conferences. His main research work focuses on Network Security, Data Science, Data Mining, IoT and Citation Network based education. He has 4 years of teaching experience and 2 years of Research Experience.

© 2019, IJCSE All Rights Reserved