Automatic Co-authorship Network Extraction and Discovery of Central Authors

V. Umadevi
Department of Computer Science and Engineering
BMS College of Engineering
Bangalore 560019, Karnataka, India
umadevi.cse@bmsce.ac.in

ABSTRACT

Scholarly publications are made available online by individual researchers on their home pages. The list of publications reflects the research work carried by individuals collaborating with one or more other researchers. Automatic extraction of author's name from the list of publications is required to construct a co-authorship network for identification of prominent authors. This will aid the scientific community to identify the works carried out in collaboration among the researchers. This paper will discuss the framework developed to automatically extract the names of authors from the list of publications to construct a Co-authorship network. Further the centrality measure analysis was carried out on this Co-authorship network to discover key authors.

General Terms:

Framework, Author's name, Publication

Keywords:

Co-authorship Network, PERL, GEPHI, Automatic Extraction, Centrality Measure

1. INTRODUCTION

Analyzing co-authorship network to discover scientific collaboration among authors is a relatively novel research area.Two researchers are considered connected if they have co-authored a paper, and these types of connections between two scientists eventually constitute co-authorship networks [1]. Individual researchers increasingly choose to share their research findings by providing lists of their published works on their respective institute home page. Researchers create their own publication lists on the Web for many reasons, such as describing their research work and contributions. Construction of co-authorship network by automatically extracting names of the author from the publication list is a challenging task. The aim of extracting the names of the authors automatically from a list of publications is to generate an error free list of authors, effortlessly and in a prompt manner. In this paper, using PERL (Practical Extraction and Reporting Language) programming language, a framework developed to automatically extract names of author from the list of publications given in the personal home page of the individual faculties/researchers will be discussed. Further, the framework uses GEPHI, an open source network analysis tool for calculation of centrality measure analysis.
The objective of the proposed framework is to construct a co-authorship network by automatically extracting names of author from list of publications and carryout centrality measure analysis on the constructed co-authorship network to identify central authors. Identification of central authors will help in expert finding and contributors for conferences, journals, workshops etc. This will further be advantageous for institutions, students and funding agencies in locating prospective researchers for collaboration, project funding, guidance etc. The words researcher, author and scientist are used interchangeably in this paper.

2. RELATED WORK

A social network is a collection of individuals, each of whom is accustomed with some other subset of others by one or more different types of relations such as friendship, affinity and co-authorship. Co-authorship networks are an important class of social networks. An initial example of the analysis of co-authorship network is Erdos Number Project. Paul Erdos was an influential but nomadic Hungarian mathematician. He was one of the most prolific publishers of papers and published at least 1401 papers during his life, more than any other mathematician in history. In bibliographical terms, the Erdos number represents a mathematicians proximity to the great man. Those who published a paper along with Erdos have an Erdos number of 1. Those who published along with a co-author of Erdos have an Erdos number of 2, and so on. Paul Erdos himself has an Erdos number of 0. Einstein has Erdos number 2, since he wrote a paper with Ernst Straus, and Straus wrote many papers with Erdos. This number calculates the smallest number of co-authorship links between any individual researcher and Paul Erdos. Eventually, it reflects an individual researchers greatness in terms of being well connected with a prestigious scientific society centered by Paul Erdos. Many great scientists and noble prize winners have small Erdos numbers [1]. In literature co-authorship network has been analyzed extensively [2, 3, 4, 5] either by extracting manually or automatically names of authors from publication data. Automatic extraction of structured knowledge (names of authors) from colossal unstructured data (publication list) is a challenging work. Co-authorship network of faculty members from Department of Computer Science in IITs was analyzed by [6] to identify internal and

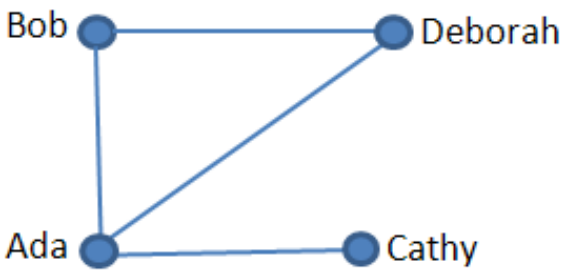

Fig. 1. Sample Co-authorship Network.

external collaboration work among faculties. This work does not discuss the method used to extract names of the authors from the publication list. A deterministic HTML Parser was developed to automatically generate large-scale Brazilian co-authorship networks, based on Lattes curricula belongings to researchers associated to specific academic areas [7]. Team-Beam algorithm has been developed by [8] to provide a flexible tool to extract a wide array of meta-data from scientific articles. But Team-Beam is a supervised machine learning algorithm, where large labeled training examples are required to learn a classification scheme for the individual text elements of an article.

Manual extraction of author names from publication list is time consuming and also it may result in erroneous data. Hence the objective of this work is to develop a framework for automatic extraction of author's names from the publication list for construction and analysis of co-authorship network.

3. CO-AUTHORSHIP NETWORK

Co-authorship is one of the most tangible and well documented forms of scientific collaboration [9]. Co-authorship network is represented as a graph. In such a graph, node represents an author or researcher. An edge exist between two nodes if two authors have co-authored a research paper together. For example say paper 1 was written by the authors Ada, Bob and Deborah; and paper 2 was written by Ada and Cathy, then the sample co-authorship network for these authors are as shown in Fig. 1. Some important applications, such as academic author ranking and expert recommendation can be extracted from co-authorship network [10].

4. PUBLICATION DATASET

Available online sources for publication databases are DBLP, CiteSeer, Google Scholar, etc., journal homepages, conference websites, homepages of individual researchers as well as organizational/institutional websites. In this work focus is on considering publication data of faculty members belonging to one single department of an institute. Institute home pages of faculty members is the trustworthy source of publication data. The publication data used in this study has been obtained from the individual faculty home pages working in the Department of Computer Science and Engineering, IIT Madras. The data used for the study was collected during the month of April, 2013. Total number of faculties in the department were twenty four [11]. Format of publication list maintained by faculty members are of various types. For the study considered two different formats of publication focusing on updated and complete list maintained by the faculty. Seven faculty members publication list was finally selected for study which were of two different formats. The sample publication of two different formats are as mentioned below.

```
Format 1:
V. K. Chaithanya Manam, V. Mahendran, and C. Siva
Ram Murthy, "Message-Driven Based Energy-Efficient
Routing in Heterogeneous Delay-Tolerant Networks,"
1st ACM Workshop on High Performance Mobile Oppor-
-tunistic Systems (MSWIM HP-MOSys), Paphos, Cyprus
Island, October 25, 2012.
Format 2:
Michalak, T. P., Aadithya, K. V., Szpzepaski, P. L.,
Ravindran, B., and Jennings, N. R. (2013) "Efficient
Computation of the Shapley Value for Game-Theoretic
Network Centrality". In the Journal of Artificial
Intelligence Research (JAIR), Vol 46, pp. 607-650.
March 2013. AAAI Press.
```

5. FRAMEWORK TO EXTRACT NAME OF THE AUTHOR, CONSTRUCT AND ANALYSE CO-AUTHORSHIP NETWORK

This section will discuss the framework developed for automatic extraction of author's name from the publication list to generate and analyse co-authorship network. PERL programming language [12] was used for extraction of author's names and from the extract co-authorship network was constructed. GEPHI, an open source network analysis tool was used for calculation of centrality measures. The work flow of the developed framework is shown in Fig. 2. Details of each module in the framework are discussed in detail in the following subsections.

5.1 Publication List Repository

Publication repository used in the framework was as discussed in section 4. Two different styles of the publication was used for parsing names of the authors. The two different formats are referred to as Format 1 and Format 2. Format Sample for the style of these two different formats are given in section 4. Total number of publications (which includes conference and journal publications) under Format 1 was 105 and for Format 2 was 719.

5.2 Extraction of author name list

By using PERL programming language, a Parser was written to extract names of the authors from two different styles of publication list. Parser scans publication record one by one, from left to right, taking from the repository. The parser stops scanning at the stopword to extract the complete list of authors names from the record. The stop words used for Format1 and Format2 are as shown in Fig. 3 and 4 respectively. The parser uses separator character as shown in Fig. 3 and 4 for extraction of individual author's name. The total number of author's names extracted automatically by the parser was 736. Then, in the next module of the framework, one to one mapping on the extracted individual authors names will be carried out to a construct co-authorship network.

5.3 Construction of Co-authorship Networks

A co-authorship network describes research activities that have been carried out by a researcher. In this module, co-authorship network is constructed from the authors names generated by the previous module in the framework. This system uses a graph to represent the co-authorship among researchers based on their respective research publications. Each researcher is represented by a node. An

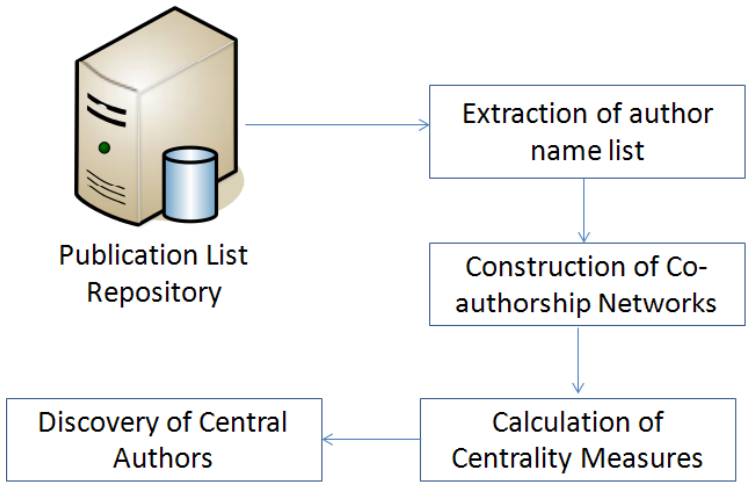

Fig. 2. Framework for automatic extraction of author's names to generate and analyse co-authorship network.

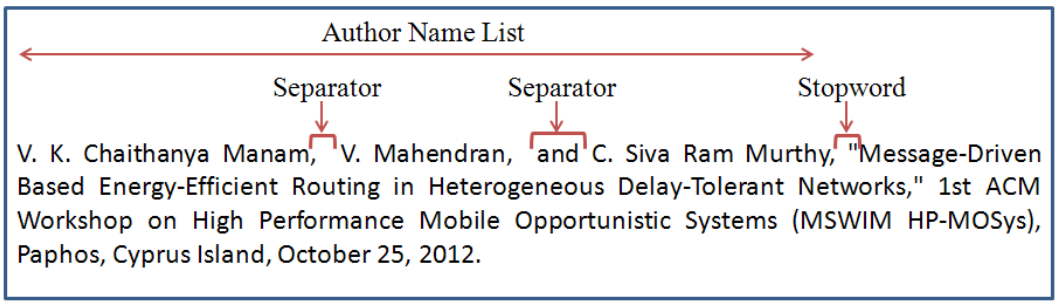

Fig. 3. Separator and Stopword for Publication Style of Format 1.

edge is created between a pair of nodes if two researchers have co-authored an research paper. For example, the name extracted was *Author1*, *Author2 and Author3*, then following edges will be created in the co-authorship network.

```
Author1 <-> Author2
Author1 <-> Author3
Author2 <-> Author3
```

Visualization of Co-authorship network constructed by this module is shown in Fig. 5 and part of this co-authorship network expanded is shown in Fig. 6. The average path length of the constructed co-authorship network was less than 6, which means that this network is like any other social network graph and is like a *small world*.

5.4 Calculation of Centrality Measures

Using GEPHI, an open source network analysis tool centrality measures were calculated for co-authorship constructed by the previous module. Centrality is regarded as one of the most important and commonly used conceptual tools for exploring actor roles in social networks [13]. The four centrality measures calculated for the constructed co-authorship network was Degree, Betweenness, Eigenvector and Closeness. Short description of these four centrality measures [14] are as given below.

Degree Centrality: An important node is involved in large number of interactions.

Closeness Centrality: An important node is typically close to, and can communicate quickly with the other nodes in the network.

Betweenness Centrality: An important node will lie on a high proportion of paths between other nodes in the network.

Eigenvector Centrality: An important node is connected to important neighbors.

Authors with higher degree are more central to the structure and tend to have a greater capacity to influence others. Closeness centrality, focuses on how close a author is to all the other authors in a network. Closeness can be regarded as a measure of how long it will take information to spread from a given vertex to others in the network [15]. Closeness centrality describes the extent of influence of an author on the network. Betweenness centrality is based on the number of shortest paths passing through a vertex. Vertices with a high betweenness play the role of connecting different groups. Authors with highest betweeenness are the pivots in the network knowledge flow. Eigenvector centrality takes into account the centrality value of the neighbors of a node to assign a centrality value to it. A node that has a high eigenvector score is adjacent to nodes that are themselves of high scorers. Thus eigenvector centrality is an influence measure, that depends both on the number and quality of its connections [16].

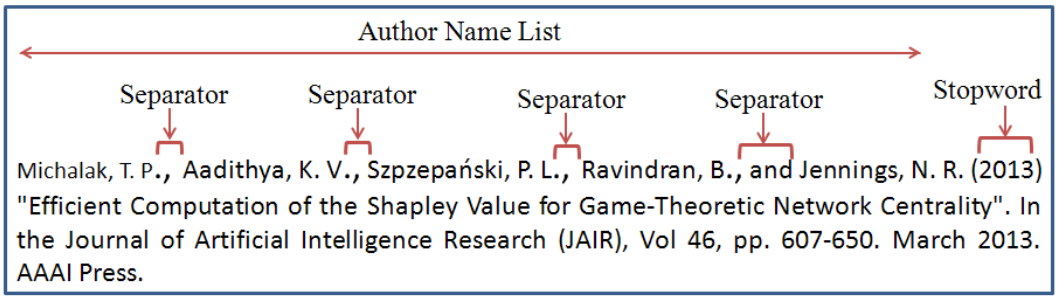

Fig. 4. Separator and Stopword for Publication Style of Format 2.

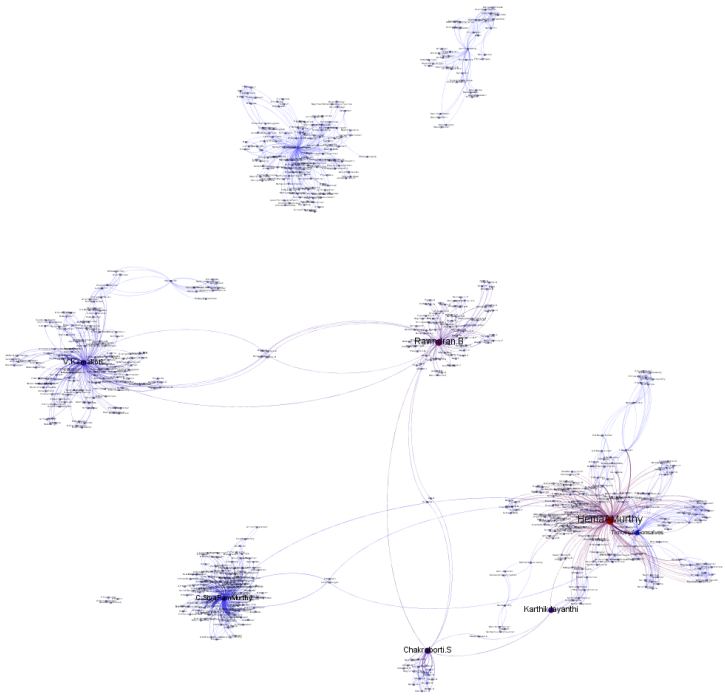

Fig. 5. Generated Co-authorship Network.

5.5 Discovery of Central Authors

Author Hema A. Murthy is having highest betweenness as well as highest eigenvector centrality; it means that this author is occupying the central position in terms of information flow and also highly connected in the network. Second highest betweenness value was for the author Ravindran. B, but his eigenvector centrality was less when compared to that of author V. Kamakoti, Timothy A. Gonsalves, C. SivaRamMurthy and Krishna M. Sivalingam. This indicates that Ravindran. B is less connected in the network when compared to authors V. Kamakoti, Timothy A. Gonsalves, and C. SivaRamMurthy and Krishna M. Sivalingam. Degree, Betweenness and Eigenvector centrality measure for the top eight authors are given in the table 1. Among the seven faculty members publications considered for analysis, only five faculty members were in the top eight ranked authors. The other two faculty members were found to be junior faculty within the department when compared to the remaining five faculty members. Closeness centrality value for the authors Krishna M. Sivalingam and John Augustine was of value one. Among the publication list of the seven faculty members considered for analysis, Timothy A. Gonsalves publication list was not obtained from his home page. But he has co-authored with Hema A. Murthy, hence he is found to be among the top eight ranked authors as seen from table 1 of centrality measures.

6. CONCLUSION

Researchers present their publication list on their home pages which reflects their research work. Motivation of this paper work was the development of framework to automatically extract names of the authors from the list of publications for constructing co-authorship network and then identification of central authors. Pro-

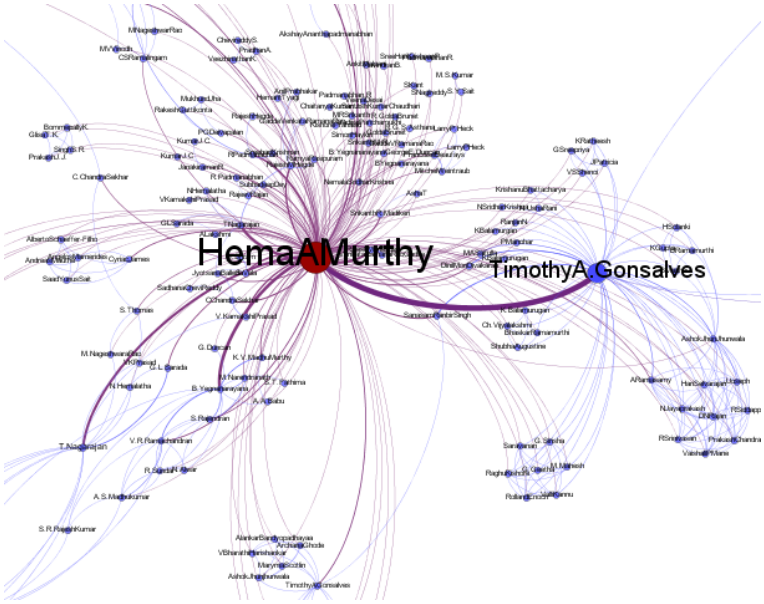

Fig. 6. Expanded: Part of Co-authorship Network

Table 1. Top eight authors based on centrality measures

Rank	Author Name	Degree Centrality	Author Name	Betweenness Centrality	Author Name	Eigenvector Centrality
1	Hema A. Murthy	187	Hema A. Murthy	110903	Hema A. Murthy	1.00
2	V. Kamakoti	152	Ravindran. B	91591	V. Kamakoti	0.54
3	C. SivaRamMurthy	131	Chakraborti. S	82781	Timothy A. Gonsalves	0.44
4	Krishna M. Sivalingam	128	Karthik Jayanthi	79182	C. SivaRamMurthy	0.39
5	Ravindran. B	84	V. Kamakoti	72169	Krishna M. Sivalingam	0.36
6	Timothy A. Gonsalves	55	C. SivaRamMurthy	63932	Ravindran. B	0.23
7	John Augustine	38	Timothy A. Gonsalves	57688	Ashok Jhun Jhunwala	0.19
8	B. S. Manoj	25	Krishna M. Sivalingam	7867	A. Ramasamy	0.17

posed framework was applied on the publication list of faculties belonging to the Department of Computer Science and Engineering, Indian Institute of Technology Madras. Here, the framework was adopted only for two different styles of publication list. The style of publication list maintained by individual faculties differ. The focus of future work is to train machine learning algorithm to adopt to all different styles of publications which makes automatic extraction of the list of authors possible, irrespective of the publication format.

7. REFERENCES

[1] S. Uddin, L. Hossain, A. Abbasi, and K. Rasmussen. Trend and efficiency analysis of co-authorship network. *Scientometrics*, 20:687–699, February 2012.

[2] M. E. J. Newman. Co-authorship networks and patterns of scientific collaboration. In *Proc. of the National Academy of Sciences*, pages 5200–5205, 2004.

[3] Wolfgang Reinhardt, Christian Meier, Hendrik Drachsler, and Peter Sloep. Analyzing 5 years of ec-tel proceedings. In *Proceedings of the 6th European conference on Technology enhanced learning: towards ubiquitous learning*, pages 531–536, Springer-Verlag Berlin, 2011.

[4] M. E. J. Newman. Scientific collaboration networks - 2 shortest paths, weighted networks, and centrality. *Physical Review E*, 64:016132–1–7, 2001.

[5] M. E. J. Newman. The structure of scientific collaboration networks. In *Proc. of the National Academy of Sciences*, pages 404–409, 2001.

[6] Tasleem Arif, Rashid Ali, and M. Asger. Scientific co-authorship social networks: A case study of computer science scenario in india. *International Journal of Computer Applications*, 52:38–45, August 2012.

[7] Jess P. Mena-Chalco and Roberto Marcondes Cesar Junior. Towards automatic discovery of co-authorship networks in the brazilian academic areas. In *IEEE Seventh International Conference on e-Science Workshops*, pages 53–60, Stockholm, 2011.

[8] Roman Kern, Kris Jack, Maya Hristakeva, and Michael Granitzer. Teambeam meta-data extraction from scientific literature. *D-Lib Magazine*, 18, 2012.

[9] Wolfgang Glnzel and Andrs Schubert. *Analysing scientific networks through co-authorship*. Handbook of Quantitative Science and Technology Research, Kluwer Academic Publishers, Netherlands, 2005.

[10] Y. Han, B. Zhou, J. Pei, and Y. Jia. Understanding importance of collaborations in co-authorship networks. In *SIAM International Conference on Data Mining*, pages 1112–1123, 2009.

[11] Department of Computer Science & Engineering. Iit madras. `http://www.cse.iitm.ac.in/faculty/`, 2013. [Online; accessed April-2013].

[12] PERL. The perl programming language. `http://www.perl.org/`, 2013. [Online; accessed March-2013].

[13] Chaoqun Ni, Cassidy R. Sugimoto, and Jiepu Jiang. Degree, closeness, and betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically. In *Proceedings of ISSI*, pages 1–13, Durban, 2011.

[14] Centrality Measures. An introduction. `https://sites.google.com/site/networkanalysisacourse/schedule/an-introduction-to-centrality-measures`, 2013. [Online; accessed May-2013].

[15] M. E. J. Newman. A measure of betweenness centrality based on random walks. *Cornell University Library arXiv:cond-mat/0309045*, pages 1–15, 2003.

[16] R. J. DSouza and Johny Jose. Significance of eigenvector centrality for routing in a delay tolerant network. *Journal of Computations & Modelling*, 1:91–100, 2011.