International Journal of Pure and Applied Mathematics
Volume 116 No. 8 2017, 341-346
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

ACADEMIC SOCIAL NETWORK DATASET APPLYING VARIOUS METRICS FOR MEASURING AUTHOR'S CONTRIBUTION

[1]C.Nalini, [2]G.Ayyappan
Professor, Department of CSE,BIST,BIHER,Bharath University
Asst.Professor, Department of IT, BIST,BIHER,Bharath University
[1]nalini.cse@bharathuniv.ac.in,[2]ayyappan.it@bharathuniv.ac.in

Abstract: Academic Social Networks are used by scholars for communication and research-related purposes. An emphasis on the use of metrics for assessing the research contribution of research work in conferences. Contribution for research we focus quantity which can be measured by number (or) size of the outputs in each researcher or author. The increasing tendency across scientific disciplines to write multi authored papers 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. This paper proposes the individual authors contribute their research work based on the quantity of management research work in researcher society. It is challenging to trace out the imprints of evolution of research in general and also to filter the quantified results for measuring the individual's contribution in their research career.

Keywords: Author, Researcher, Frequency, Metrics, Cartesian-Product, classification

1. Introduction

Academic Social Networks are used by scholars for communication and research-related purposes. Different from the earlier academic social networks like Mendeley [11], Zotero [17] and CiteULike [1-2] which were meant for uses as references and files sharing. The recent academic social networks, academia.edu and Researchgate.net came as full collaboration platforms. They allow users to communicate, collaborate, and follow or being followed, attracting millions of researchers providing better channels for scholarly communication. Researchgate. net, incepted in the same year (2008) by a physician, witnessed a viral expansion with more than 5 million members presently [3-4]. It has features similar to Academia.edu withother features borrowed from Twitter and Facebook, emphasizing discussions and collaboration.Researchers on Researchgate.net can create and modify their profiles, upload/download publications, view, comment, make/answer questions, follow or being followed by RSS service to keep current and up-to-date. The main advantages of Researchgate.net to researchers, is that it allows self-archiving, reputation building and informal exchange of publications, which would result in better publication visibility and knowledge sharing. It is possible to find papers from within Researchgate.net and, to search some external databases such as PubMed, Cite-Seer and arXiv through its efficient search engine. It is easy for a researcher to advertise on his profile different events such as meetings and workshops. The network provides a platform for a researcher to create a profile, publish his/her papers and communicate with other researchers, presenting a new way for scholarly communication. Resesearchgate.net is receiving good attention and becoming popular among the researchers community. Its popularity ranking is shown by Alexa.com, which calculates the global rank of a website using a combination of average daily visitors and page views on the site for the last 3 months.Since academic social networks over an attractive alternative to meet researchers' information needs in addition to other features where even negative results can be reported. When Researchgate.net reaches Satisfactory level of intake of academic faculty members, we think that it has the potential of becoming one of the evaluating tools of researchers' performance in the future. Here, the related work for this paper in section 2, we proposed method for finding individual frequency in section 3, the information about the dataset, we discussed result of various classifiers approaches in section 4 represents and finally we concluded the paper in section5.

2. Related work

We measure the individual authors contributions based on quantity frequency in the dataset. We find the solution based on these methods. Tailem Arif[9-10] proposed method 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 indicate that the proposed technique shows a lot of improvements over the existing methods.

ThiThanh Sang Nguyen et al.[2] proposed to represent the domain knowledge. The first model uses an ontology to represent the domain knowledge. The second model uses one automatically generated semantic network to represent domain terms, Web-pages, and the relations between them. Another new model, the conceptual prediction model, is proposed to automatically generate a semantic network of the semantic Web usage knowledge, which is the integration of domain knowledge and Web usage knowledge Arantxa Duque Barrachina et al.[11-12]proposed a Proof of Concept (PoC) end to end solution that utilises the Hadoop programming model, extended ecosystem and the Mahout Big Data Analytics library for categorising similar support calls for large technical support data sets. The proposed solution is evaluated on a VMware technical support dataset. Anand Kumar et al.[26]have developed a prototype of DCMS based on the Post gre SQL system and experiments using real MS data and workload show that DCMS significantly outperforms existing MS software systems. We also used it as a platform to test other data management issues such as security and compression. N K Nagwani [27]proposed technique is designed using semantic similarity based clustering and topic modeling using Latent Dirichlet Allocation (LDA) for summarizing the large text collection over Map Reduce framework. The summarization task is performed in four stages and provides a modular implementation of multiple documents summarization. The presented technique is evaluated in terms of scalability and various text summarization parameters namely[15-16], compression ratio, retention ratio, ROUGE and Pyramid score are also measured. The advantages of Map Reduce framework are clearly visible from the experiments and it is also demonstrated that Map Reduce provides a faster implementation of summarizing large text collections and is a powerful tool in Big Text Data analysis.

3. Proposed work

The data has collected from Aminer Academic social network dataset. First data has selected and applied the preprocessing technique and implemented the text mining. After find the result of text mining, we have analysis author for individual frequency and statically analysis.

(i) Data extraction

The data which is obtained from various sources are always consist of combination of usable and irrelevant data. The data extraction is the first step in preprocessing of data. The data obtained from different sources are normally in unstructured, semi structured and structured type. The process of converting data into a usable form for processing further is called data extraction.

(ii) Text Mining

Text databases consist of huge collection of documents. They collect these information from several sources such as news articles, books, digital libraries[17-18], e-mail messages, web pages, etc. Due to increase in the amount of information, the text databases are growing rapidly. In many of the text databases, the data is semi-structured.

(iii) Basic Measures for Text Retrieval

We need to check the accuracy of a system when it retrieves a number of documents on the basis of user's input. Let the set of documents relevant to a query be denoted as {Relevant} and the set of retrieved document as {Retrieved}. The set of documents that are relevant and retrieved can be denoted as {Relevant} ∩ {Retrieved}.

There are three fundamental measures for assessing the quality of text retrieval –

- Precision
- Recall
- F-score

Precision

Precision is the percentage of retrieved documents that are in fact relevant to the query. Precision can be defined as –

$$\text{Precision} = |\{\text{Relevant}\} \cap \{\text{Retrieved}\}| \,/\, |\{\text{Retrieved}\}|$$

Recall

Recall is the percentage of documents that are relevant to the query and were in fact retrieved. Recall is defined as –

$$\text{Recall} = |\{\text{Relevant}\} \cap \{\text{Retrieved}\}| \,/\, |\{\text{Relevant}\}|$$

F-score

F-score is the commonly used trade-off. The information retrieval system often needs to trade-off for precision or vice versa. F-score is defined as harmonic mean of recall or precision as follows –

$$\text{F-score} = \text{recall} \times \text{precision} \,/\, (\text{recall} + \text{precision}) \,/\, 2$$

(iv) Author position

Individual Frequency (IF): the equal importance of all author positions in the article. Consider the total number of occurrences of all authors in research community based on the counts.

(v) Data Analysis and evaluation

The kind of analysis that can be performed on a set of data will be influenced by the goals identified at the outset, and the data actually gathered. Broadly speaking, you may take a qualitative analysis approach or a quantitative analysis approach, or a combination of qualitative and quantitative. The last of these is very common as it provides a more comprehensive account of the behavior being observed or performance being measured.

(vi) Model Evaluation

recommendation, association search, course search, academic performance evaluation, and topic modeling.

S.No	Tag (Starting of new attribute)	Attribute Name	Meaning of Attribute
1	#index	index id	index id of each paper
2	#*	paper title	Paper title
3	#@	Authors	Authors Name
4	#o	Affiliations	Affiliations of Authors
5	#t	Year	Year of conference
6	#c	publication venue	Publication venue
7	#%	the id of references of this paper	References articles ID
8	#!	Abstract	Abstract of each paper

4. Experimental setup and result

4.1 Source of data collection

In this experiment the dataset is downloaded fr**om https://aminer.org/aminernetwork.** The content of this data includes research article information, research article citation, author information and author collaboration. It contains **2,092,356** research articles and **8,024,869** citations among them.

The identification of data for the research was made on the various publications like as journals and conferences for studying Academic social network and Topic paper Author datasets in text format, to understand its authenticity and reliability Arnet Miner (AMiner) datasets are regarded as standard which was identified as data source.

Arnetminer is designed to search and perform data mining operations against academic publications on the Internet, using social network analysis to identify connection between researchers, conferences, and publications. This allows it to provide services such as expert finding [19-20], geographic search, reviewer Due to limitations of memory and processing capabilities, we proposed random subset approach by filtering this Arnetminer was created as a research project in social influence analysis, social network ranking, and social network extraction. A number of peer-reviewed papers have been published arising from the development of the system. It has been in operation for more than three years, and has indexed 1,300,000 researchers and more than three million publications. The research was funded by the Chinese National High-tech R&D Program and the National Science Foundation of China.

Arnetminer is commonly used in academia to identify relationships between and draw statistical correlations about research and researchers. It has attracted 2,766,356 independent IP accesses from 220 countries. The product has been used in Elsevier's SciVerse platform and academic conferences such as SIGKDD, ICDM, PKDD, WSDM.

5. Result-Text mining

The dataset is organized in the following table with the attributes and their interpretations

massive real time dataset to 6000 records from **2,092,356 records/**research articles. We divided 6000 records into

three different folders labeled as 'large', 'medium', and 'small', by text mining the reference part of the record guarded by the tag '#%' as shown in the table 1. As we are looking for unbiased or balanced dataset each folder is assumed to have uniform sized text records.

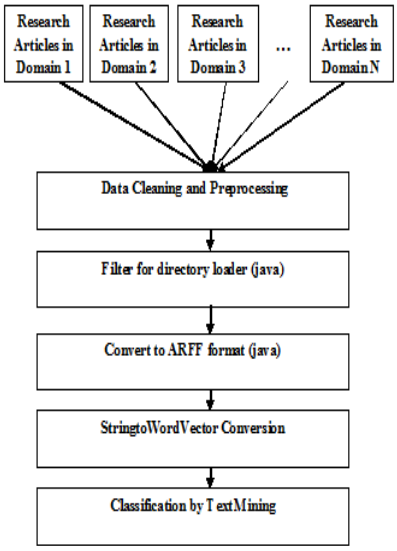

Figure 1. Proposed Architecture for classification of SN Articles by Text mining process

Each folder contains 2000 individual records in text file format for the purpose of text mining. Then, we apply the code in to weka simple Command Line Instruction (CLI) in this command for converting one major folder it contains three sub individual folders with labels small, medium, large folders into Attribute -Relation File Format. This is implemented by the java class especially by the command,
javaweka.core.converters.TextDirectoryLoader -dir D:\inputfolder > D:\output.arff .

5.1 Result – various classification Technique

In this experimental exercise we select four categories of classifications namely Bayes types, Ensemble (Meta) types, Decision Tree types and Decision rule types several methods BayesNet, NaiveBayes from Bayes classification[21-22], Attribute Selected Classifier, Dagging from Meta classification, DecisionStump, J48 from Trees classification and JRip, ZeroR from Rules classification. We trained the input dataset with these classifiers and tested using cross validation procedure using 10 folds. The results obtained with eight classifiers are tabulated as shown in table 2.

S. No	Name of the Classifiers	Category of the Classifiers	Accuracy
1	BayesNet	Bayes	83.98 %
2	Complement	Bayes	82.18 %
3	AttributeSelectedClassifier	Meta	84.41 %
4	Dagging	Meta	88.35 %
5	Jrip	Rules	87.58 %
6	ZeroR	Rules	33.33 %
7	DecisionStump	Trees	64.65 %
8	J48	Trees	66.76 %

In Bayes classifier, Bayes Net accuracy has 83.98% and Bayes. Complement accuracy has 82.18%. In Meta classifier, Attribute Selected Classifier accuracy is 84.41% and Dagging accuracy is 88.41%.In Rules classifier Jrip accuracy has 87.58% and ZeroR accuracy has 33.33%.In Trees classifier, Decision stump accuracy has 64.65% and J48 accuracy has 66.76%.

In this experiment mainly focuses on text mining the data from Academic social networks In this experiment Meta Category and Bayes Classification methods and Rules one of the Jripclasifier shows the maximum accuracy for the academic social network dataset. The Dagging classifier is the best classification method belongs to the Meta classification method compare than other classifiers. We observe Dagging is the optimal classifier while comparing with the performance of other classifiers selected for this experiment.

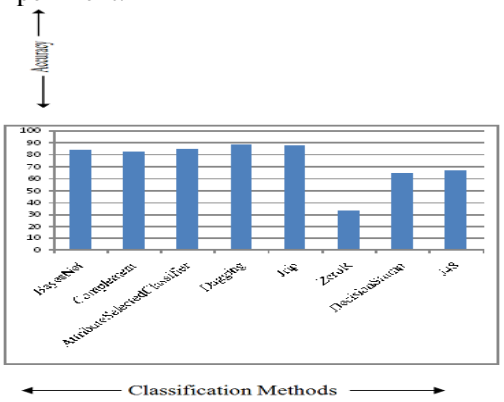

Figure 2. Graphical representation between selected classification methods and accuracy

The accuracies obtained from the selected classifiers are shown in Fig.6.1.2. In this figure represents comparison of all categories of classifiers. In Bayes classifier method, sBayes Net has high accuracy compare than Bayes Complement. In Meta classification, Dagging has high accuracy compare than Attribute Selected Classifier. In Rules classification method, Jrip has high accuracy compare than ZeroR. In Trees classification J48 has high accuracy compare than Decision Stump. The Meta category, Bayes Category and Jrip in Rules category have dominates to ZeroR classifier in Rules category and Tree category. The above diagram represents high accuracy and similar tendency are having Dagging classifier ,Jrip classifier, Attribute Selected Classifier, Bayes Net classifier and Bayes Complement classifier and then low accuracy are having Trees category(J48 and Decision Stump) and ZeroR in Rules category.

5.2 Result - Individual Author contributions

We apply Cartesian product or cross product concepts in Topic-Paper-Author dataset we split in to each and every author. The length of authors' position from first author position to thirty second authors position. We avoid redundancy and find there are 22875 individual authors contributions and find top 20 individual authors contributions on conferences from the period of 1990 to 2005 and five topics.

e fix the threshold below 50 to 30 for the count of individual frequency in Topic Paper Author dataset. We find top 79 authors of individual frequency who are all contribute the conference in between 1990 to 2005 in five topics.

6. Conclusion

Social Networks like 'AMiner' helps the miners to go for such direction and they obtain results to throw more results. There are types of researchers initiating (originating, by authoring), collaborating (by coauthoring), and supporting (by reference or citations). In this paper we aim at novel estimation for the absolute frequency of individual contribution and relative weighted frequency with respect to their collaboration also. In future, we estimate and evaluate authors contributions based on H index and I index of each authors using social networks.

References

[1]TasleemArif,Exploring , The Use Of Hybrid Similarity Measure For Author Name Disambiguation,INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 12, DECEMBER 2015 ISSN 2277-8616 ,pp:171-175.

[2] ThiThanh Sang Nguyen, Hai Yan Lu, and JieLu,Web-Page Recommendation Based on Web Usage and Domain Knowledge, IEEE Transactions on Knowledge and Data Engineering (Volume: 26, Issue: 10, Oct. 2014),Page(s): 2574 - 2587,Page(s): 2574 - 2587Print ISSN: 1041-4347.

[3] JieTang,LiminYao,DuoZhang,and Jing Zhang,A Combination Approach to Web User Profiling ACM Transactions on Knowledge Discovery from Data, Vol. V, No. N, March 2010, Pages 1–38.

[4] Arantxa Duque BarrachinaandAislingO'Driscoll,A big data methodology for categorising technical support requests using Hadoop and Mahout,DuqueBarrachina and O'Driscoll Journal of Big Data 2014, 1:1 http://www.journalofbigdata.com/content/1/1/1

[5] Anand Kumar, Vladimir Grupcev, MeryemBerrada, Joseph C Fogarty, Yi-Cheng Tu1,Xingquan Zhu, Sagar A PanditandYuniXia,DCMS: A data analytics and management system for molecular simulation,Journal of Big Data 2014, 1:9 http://www.journalofbigdata.com/content/1/1/9.

[6]Udayakumar R., Kaliyamurthie K.P., Khanaa, Thooyamani K.P., Data mining a boon: Predictive system for university topper women in academia, World Applied Sciences Journal, v-29, i-14, pp-86-90, 2014.

[7]Kaliyamurthie K.P., Parameswari D., Udayakumar R., QOS aware privacy preserving location monitoring in wireless sensor network, Indian Journal of Science and Technology, v-6, i-SUPPL5, pp-4648-4652, 2013.

[8]BrinthaRajakumari S., Nalini C., An efficient cost model for data storage with horizontal layout in the cloud, Indian Journal of Science and Technology, v-7, i-, pp-45-46, 2014.