Computing Frequency of the Resembled Words of a Articles Automatically as a Trend to Researchers

Arpitha T[1], Akshatha K[2], Ariptha V[3], Amrutha A[4], Manasa DR[5]
[1,2,3,4,5] *Vivekananda Institute of Technology*

(E-mail: [1]arpitha811@gmail.com, [2]gowdaakshatha5@gmail.com, [3]arpitha284@gmail.com, [4]amrutha28021996@gmail.com,[5] manasadr1997@gmail.com)

Abstract—Bibliometric analysis provides a variety of tools for exploring publication data, but often involves manual effort. This paper presents an automatic method for extracting and examining key research themes by using natural language processing to parse a large collection of papers. The method was applied to over 8,000 papers published in the software engineering field over the past 20 years. Key research themes were identified and visualized, so that trends could be highlighted. Some research fields that are in decline are identified, along with newly popular research topics such as fuzzy set membership, cloud computing, feature selection and agile development teams.

Keywords—*Bibliometric Analysis; Natural Language Processing; Information Extraction; Software Engineering;*

INTRODUCTION

Bibliometric analysis is used to study academic publications to better understand the evolution of research in various fields [1]. A variety of metrics have been used including citation graphs, annual paper growth rates, and using data about authorship, geographic location, institution, collaborations etc [2][3]. The information gained from Bibliometric studies is useful for several reasons such as managing and organizing research proposals [4], predicting the likely impact of research plans and visualizing research themes [5]. Most of these existing techniques are quantitative, applying statistical analysis to numerical data about publications. There have however been some qualitative studies that use content analysis to investigate the content of the publications [6]. One study investigated the journal "Pain", manually assessing each paper published over a 32 year period to extract the different types of pain being investigated [7]. Rather than manually assessing the content of each publication, an alternative is to use text mining, such as in one study where the abstracts of a collection of papers on M-Learning, or mobile learning, were automatically parsed to examine academic trends [8]. Another study investigated the bibliographic records for articles about ceramics collected from articles that contained the word 'ceramic' included in the Scopus database [9]. The limitation of these studies is that they only parsed the abstract of the paper. The study discussed in this paper uses the complete text, rather than just the abstract, of much larger collection of papers.

Natural Language Processing (NLP) techniques have been applied in several areas both for generating and understanding natural language. Text mining is used to extract meaning automatically from written text, and this research is concerned with statistical, or probabilistic NLP for information extraction [10]. In this study a collection of papers published over the past 2 decades has been parsed to extract the key features of each paper, clustering similar papers together to identify recurring themes.

There are several reasons for parsing and clustering related documents, including for indexing and organizing documents, segmenting customer profiles or for collaborative filtering [11]. The objective of this research is to extract and visualize the themes of a document, so that related documents can be clustered. This paper investigates how clusters have evolved over the past couple of decades. There are various approaches to extracting the key features of a document, in this study the log likelihood of each word's occurrence is calculated. The similarity of the features of documents can also be compared in different ways, such as Euclidean distance, or cosine similarity [12], or as this project uses, the RV coefficient is calculated. Clusters are then created by grouping related documents.

The NLP methods used in this research were applied to a collection of academic papers within the field of software engineering. Software engineering emerged from the software crisis of the 1960s, 70s, and 80s, to study the application of engineering techniques to the design, production and maintenance of software [13] with the goal of reducing the number of projects that ran over their budgets or schedules. One related research project investigated the CAiSE community, identifying and examining the relationships between authors in the CAiSE community [14].

A variety of approaches to Software Engineering have been devised and investigated over the past few decades, for instance, formal methods, object oriented approaches, the use and reuse of components, process analysis such as with the capability maturity model, etc. Whilst none of these approaches have singularly solved the challenges of software engineering, they have become important research themes. In the most recent editions of software engineering textbooks, topics such as agile methods, Scrum, aspect oriented and service oriented engineering, have been added [13]. This paper investigates how the popularity of software engineering

research themes has evolved, identifying themes that are growing in their popularity.

METHODOLOGY

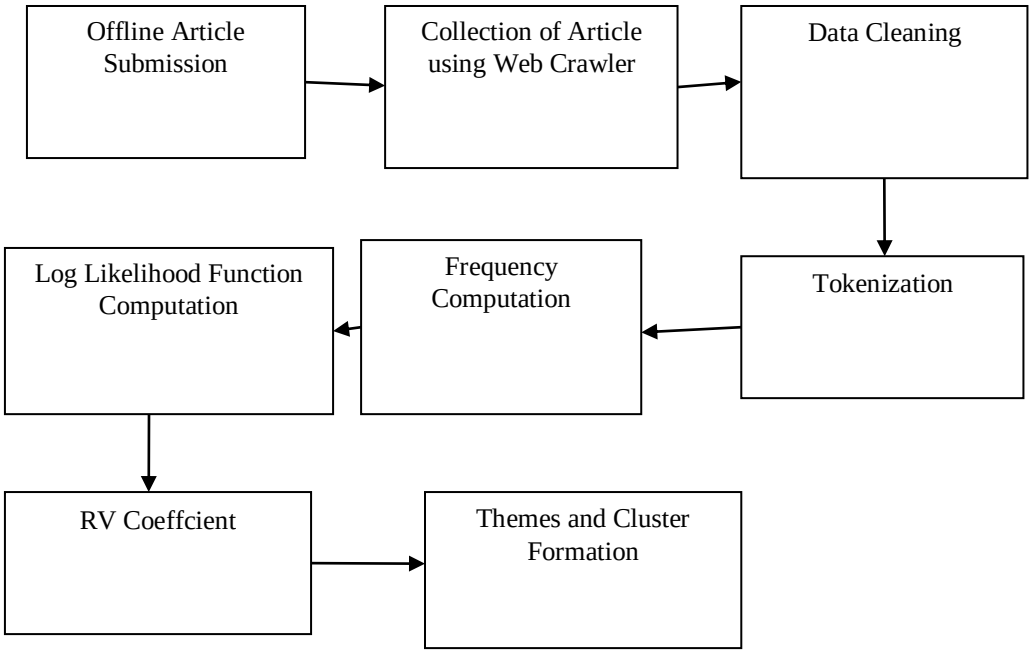

Fig: Methodology of Project

The Data Cleaning algorithm is responsible for removal of stop words. Each of tweet is cleaned by removing the stop words from tweet.

These are the set of words which do not have any specific meaning. The data mining forum has defined set of keywords. Stop words are words which are filtered out before or after processing of natural language data (text). There is not one definite list of stop words which all tools use and such a filter is not always used. The list of stop words used in the algorithm are as follows

a,able,about,across,after,all,almost,also,am,among,an,and,any,are,as,at,be,because,been,but,by,can,cannot,could,dear,did,do,does,either,else,ever,every,for,from,get,got,had,has,have,he,her,hers,him,his,how,however,i,if,in,into,is,it,its,just,least,let,like,likely,may,me,might,most,must,my,neither,no,nor,not,of,off,often,on,only,or,other,our,own,rather,said,say,says,she,should,since,so,some,than,that,the,their,them,then,there,these,they,this,tis,to,too,twas,us,wants,was,we,were,what,when,where,which,while,who,whom,why,will,with,would,yet,you,your

Data Cleaning is used for removing the stop words from each of the tweets and clean them. After the data cleaning process is completed the clean data can be represented as a set *(CleanId, CleanData, ArticleId). CleanId* is the unique Id associated with the Tweet, *Clean Data* is the clean data after removal of clean data and Article Id is the unique Id associated with the article.

Tokenization

Tokenization is a process of converting the clean data into a set of words known as tokens. Each of the token can be represented as *Token Id, Token Name and Article ID*

Token ID	Article ID	Token Name

Frequency Computation

This is a process in which the frequency computation is performed. For each of the articles the frequency is computed. Frequency is number of times a i^{th} token appears in article j^{th}. The frequency matrix is computed in the following format

Freq ID	Article ID	Token Name	Frequency

Log Likelihood Function

The log likelihood is determined for each of the tokens in the articles and is given by the equation

$$-2\ln\lambda = 2\sum_{i=1}^{l} O_i \ln(\frac{O_i}{E_i})$$

Where,

$O_i = Number\ of\ times\ other\ i^{th}\ word\ appearing\ in\ an$

$N_i = frequeny\ of\ i^{th}\ word$

$E_i = \exp ected\ \ frequency\ word$

The expected Frequency is computed in the following way for all the words of articles

$$E_i = \frac{N_i \sum_{i=1}^{N_{wordsother}} O_i}{\sum_{i=1}^{N_{wordsin\ article}} N_i}$$

The data is stored in the following format

Freq ID	Article ID	Token Name	Frequency	Log Likelihood Function

RV Coefficient Similarity

Select the two articles to be compared. Find the List of unique words in Article A1. Find the List of unique words in Article A2, Find the intersection set between the two lists Compute the mean of the values of frequency for the words in article1, Compute the mean of the values of frequency for the words in article 2, Compute the standard deviation for the words in article1, Compute the standard deviation for the words in article2, Compute the RV Coefficient value for the 2 articles using

$$\varphi = \frac{1}{N_{values} - 1} \frac{\sum_{i=1}^{N_{values}} (x_i - \bar{x})(y_i - \bar{y})}{\bar{\bar{x}}\,\bar{\bar{y}}}$$

where,

$x_i = word\ \ frequency\ of\ i^{th}\ word\ in\ article\ x$

$y_i = word\ \ frequency\ of\ i^{th}\ \ word\ in\ article\ y$

$\bar{x} = mean\ of\ \ words\ \ for\ article\ x$

$\bar{y} = mean\ of\ \ words\ for\ article\ y$

$\bar{\bar{x}} = s\tan dard\ deviation\ of\ words\ \ for\ article\ x$

$\bar{\bar{y}} = s\tan dard\ deviation\ of\ words\ \ for\ article\ y$

$N_{values} = Number\ of\ \ unique\ values$

Clustering Process

Clustering is an approach by which each of the articles are compared and then if they the value is higher than a certain threshold then they are grouped into one cluster. The process is repeated until all articles are scanned and grouping is performed.

Themes

These are set of words whose frequency is highest.

References

[1] M. Thelwall, "Bibliometrics to webometrics". Journal of Information Science 2008 34(4), pp 605-621.
[2] B. Cronin, C. Sugimoto, "Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact", MIT Press 2014.
[3] G. Keshaval, M. Gowda, "ACM transaction on information systems (1989-2006): A bibliometric study." Information Studies 2008, 14(4), pp 223-234.
[4] J. Ma, W. Xu, Y. Sun, E. Turban, S. Wang, O. Liu, An "Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection." IEEE Transactions onSystems, Man and Cybernetics – Part A: Systems and Humans 2012, Vol. 42, No. 3.
[5] M. Lee, T. Chen, "Revealing research themes and trends in knowledge management from 1995 to 2010" Knowledge Based Systems 2012 28, pp 47-58.
[6] M. Zitt, E. Bassecoulard, "Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis" Scientometrics, 1994, Volume 30, No.1, pp 333-351.
[7] S. Mogil, K. Simmonds, J. Simmonds, "Pain Research from 1975 to 2007: A categorical and bibliometric meta-trend analysis of every research paper published in the journal Pain" Pain 2009, 142, pp 48-58.
[8] J. Hung, K. Zhang, "Examining mobile learning trends 2003- 2008: a categorical meta-trend analysis using text mining techniques" Journal of Computing in Higher Education 2012, Volume 24, Issue 1, pp 1-17.
[9] S. Deville, A. Stevenson, "Mapping ceramics research and its evolution" Journal of the American Ceramic Society 2015 98(8) 2324-2332.

[10] P. Jackson, I. Moulinier, "Natural Languge Processing for Online Applications, Text Retrieval, Extraction and Categorization" John Benjamins Publishing Company 2002.
[11] C. Aggarwal, C. Zhai, "A Survey of Text Clustering Algorithms" Mining Text Data 2012 Springer, pp77-129.
[12] A. Huang, "Similarity Measures for Text Document Clustering" New Zealand Computer Science Research Student Conference 2008, pp 49-56.
[13] I. Sommerville, "Software Engineering" 10th Ed., Pearson Education, 2015.
[14] M. Jarke, M. Pham, R. Klamma, "Evolution of the CAiSE Author Community: A Social Network Analysis" Seminal Contributions to Information Systems Engineering, 2013, 15-33.
[15] K. Cosh, "Towards Automatically Retrieving Discoveries and Generating Ontologies" In Kim, K. (eds) Information Science and Applications. Lecture Notes in Electrical Engineering, 2015, vol. 339, Springer.
[16] P. Rayson, R. Garside, "Comparing Corpora using Frequency Profiling". Workshop on Comparing Corpora held in conjunction with the 38th annual meeting of the Association for Computational Linguistics, ACL 2000.
[17] K. Cosh, R. Burns, T. Daniel, "Content Clouds, Classifying Content in Web 2.0" Library Review 2008 Vol. 57, Issue 9, pp 722-729.
[18] H. Do, "Prioritizing JUnit Test Cases: An Empirical Assessment and Cost-Benefits Analysis" Empirical Software Engineering, 2006 Vol. 11, Issue. 1.
[19] P. Robert, Y. Escoufier, "A Unifying Tool for Linear Multivariate Statistical Methods: the RV-Coefficient" Journal of the Royal Statistical Society 1976, Vol. 25, No. 3.
[20] P. Agarwal, M. Alam, R. Biswas, "Issues, Challenges and Tools of Clustering Algorithms", IJCSI International Journal of Computer Sciences 2011, Vol. 8, Issue 3, No.2.
[21] Y. Zhao, G. Karypis, "*Evaluation of hierarchical clustering algorithms for document datasets*", Proceedings of the International Conference on Information and Knowledge Management, 2002.
[22] Milne, O. Medelyan, I. H. Witten, "*Mining domain-specific thesauri from wikipedia: A case study*", In Proc. of the International Conference on Web Intelligence (IEEE/WIC/ACM WI'2006), *2006.*
[23] R.Z. Osmar, S. Alexander, "*Finding similar queries to satisfy searches based on query traces*", Workshops of the 8th International Conference on Object-Oriented Information Systems, pp. 207-216, September 2002.
[24] G. Salton, A. Wong, C. S., "*A Vector Space Model For Automatic Indexing*" in , Cornell University.
[25] K. P. Supreethi E.V. Prasad,"*Web Document Clustering Technique Using Case Grammar Structure"*, International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) 13-15 Dec. 2007
[26] K. Hammouda and M. Kamel, "*Phrase-Based Document Similarity Based on an Index Graph Model*," Proc. IEEE Int'l Conf. 2002
[27] J.L. Fagan, "*Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-Syntactic Methods*," PhD thesis, Dept. of Computer Science, Cornell Univ., Sept. 1987.
[28] Maitri P. Naik ; Harshadkumar B. Prajapati ; Vipul K. Dabhi,"*A survey on semantic document clustering*", 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT),5-7 March 2015
[29] Tar et al., "Enhancing Traditional Text Documents Clustering based on Ontology", Int. J of Comput. Applicat., vol. 33.10, pp. 38-42, 2011.