International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp. 6667-6672
© Research India Publications. http://www.ripublication.com

Identifying the Research Specializations from the Publications using Text Mining and Linear Discriminant Analysis

Padmaja Ch V R[1], Lakshmi Narayana S[2] and Divakar Ch[3]

[1]Department of Computer Science & Engineering,
Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

[2]Senior Member, IEEE

[3]Professor, Department of Information Technology,
SRKR Engineering College, Bhimavaram, Andhra Pradesh, India

Abstract

In the past century data mining is one of the research area exercised by many scientists for delineating information from data and delivered good results. We attempted to find a scientific method to calculate few parameters like coefficients of Linear Discriminates and tabled to identify and range out reach of a scientist to the specializations. Reasonable conclusions are derived, and a good relationship is found from the data using LDA classification method. Coefficients of Linear Discriminates were computed from the key word percentages that are matched in 10 separate domains. The keywords were collected from the Microsoft Academic Search web site. From the independent datasets of author specializations in various domains, we found that the results are up to 80% to 90% accuracy.

Keywords: Text mining, domain specific keyword, domain specific word cloud, domain ratio table.

INTRODUCTION

Many scientists have tried to identify the author specialization using data mining techniques. Earlier a scientist himself indicates his specialization depending upon his area of working projects under take. An attempt was made to identify and range a particular scientist to his specializations using content based filtering [1] and statistical methods.

Text Mining[16] also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Text mining with R[2] applies analytic tools to learn from collections of text documents like books, newspapers, emails, etc. The goal is similar to humans learning by reading books. Using automated algorithms, we can learn from massive amounts of text, much more than a human can.

To experiment, around 1200 research publications of 50 professors were collected from their institute websites, which are publicly available. These publications are in different formats and we converted into text format to form the Corpus[2]. Pre-processing steps like removal of punctuations, numbers, stop words etc., were applied on Corpus. This corpus was then transformed into a Document Term Matrix(DTM) [2].

Initially, Word Clouds [3,4] are used to analyze these research publications. With word clouds, the analysis is fast and visually rich way to enable the readers to have some basic understanding of the data at hand. The Word clouds can be used as a tool for preliminary analysis and for validation of previous findings. However, they may not be accurate all the times as they require a lot of data pre-processing in order to reflect the actual summary of the document(s)/ publications(s).

A new approach, Domain Specific Word Cloud is proposed for generating more specific word clouds. In this new approach, words are identified which are related to a specific domain from all the research publications of an author along with their frequencies and formed a new DTM namely Domain DTM. These words along with their frequencies are then plotted as a Word cloud, termed as Domain Specific Word cloud. The Domain Specific Word cloud gives a better idea of the research area of an author in contrast to the normal Word cloud. The Domain Specific Word clouds can help us in understanding the main research area of an author. But, identifying various domains of an author through this Domain Specific Word cloud is not realistic.

A machine learning algorithm, KNN [5,6] was applied on the Domain DTM, to classify the documents and achieved 60% to 70% accuracy in results. To improve accuracy in classification, we applied another approach, where in various parameters of vectors for each publication of an author was derived. These parameters are keywords' percentage that match with each domain keywords. These vectors form the Domain Ratio Table which can be used for further study.

Statistical methods namely Principle Component Analysis(PCA) [7,8] and Linear Discernment Analysis(LDA) [9,10,11], were applied on Domain Ratio Tables to find a linear combination of features which characterizes or separates two or more classes and obtained 80% to 90% accuracy in results.

RELATED WORK

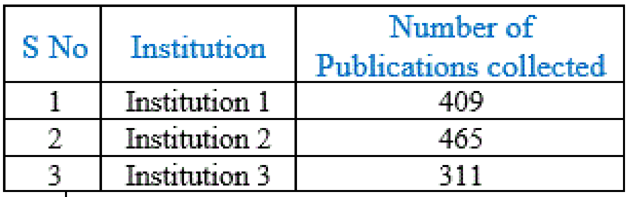
Earlier Mohamed Ali AlShaari [12] stated that every discipline has its own terminologies gives the ability to categorize any document, irrespective of discipline. He presented an algorithm that identifies the specified terminologies of a discipline and classified the documents depending on the ratio of these specified terminologies in each document.

An architecture for text mining called DISCOTEX, *i.e.,* Discovery from Text Extraction which used a learned information extraction system to convert text into data which is more structured for mining interesting relationships was proposed by T. Lalitha and S. Meenakshi [13]. It combines the information extraction module with standard rule induction module to improve recall of underlying system. E. Alan Calvillo [14] explained a better classification of research papers by providing an architecture that works with a knowledge database related topics of databases, operating system and programming. This architecture uses k-means algorithm for clustering of research papers. N. Arunachalam et.al [15] proposed an idea of ontology based text mining approach to cluster research proposals on the basis of similarities in research area. They used ontology which is a knowledge repository containing concepts and terms and relationships between the concepts which makes the task of searching similar pattern of text effective, efficient and interactive.

METHODS AND IMPLEMENTATION

R[18] is a programming language and software environment for statistical computing and graphics supported by the R foundation for Statistical Computing. Rstudio[19] is a free and open-source integrated development environment (IDE) for R. It was founded by JJ Allaire, creator of the programming language ColdFusion. The proposed work is implemented completely in Rstudio.

We collected around 1200 research publications of 50 professors, from their institute websites, which are publicly available. Table1 represents the statistics of the publications collection from three reputed national institutions.



S No	Institution	Number of Publications collected
1	Institution 1	409
2	Institution 2	465
3	Institution 3	311

Table 1 : Publications collected for Analysis

The publications of different authors are collected in different formats, later converted into text files to generate Corpus. To this Corpus, various text mining pre-processing techniques are applied and identified frequent terms. The frequent terms are visualized through Word Clouds.

A document is usually categorized by its few key words. These words specify to which topic the document is related and covering, and do not necessarily appear more number of times within the document. Hence total term frequency is not necessarily indicative of a term's information content. These frequent terms were compared with the domain specific keywords which are collected from Microsoft Academic Search web site. The domain matching ratios are tabulated for applying some statistical methods. The main objective of this work is to classify the publications of different domains so that, the specializations of the author can be found. The architecture model of the proposed system is explained in Figure 1.

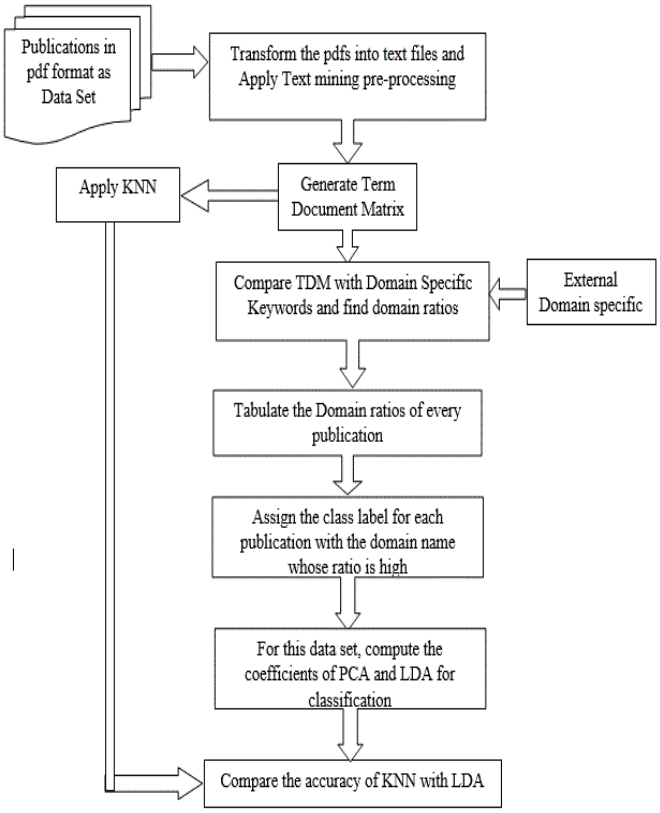

Publications in pdf format as Data Set
Transform the pdfs into text files and Apply Text mining pre-processing
Apply KNN
Generate Term Document Matrix
Compare TDM with Domain Specific Keywords and find domain ratios
External Domain specific
Tabulate the Domain ratios of every publication
Assign the class label for each publication with the domain name whose ratio is high
For this data set, compute the coefficients of PCA and LDA for classification
Compare the accuracy of KNN with LDA

Figure 1: Architecture of Proposed Method

A. *Pre-processing the documents*

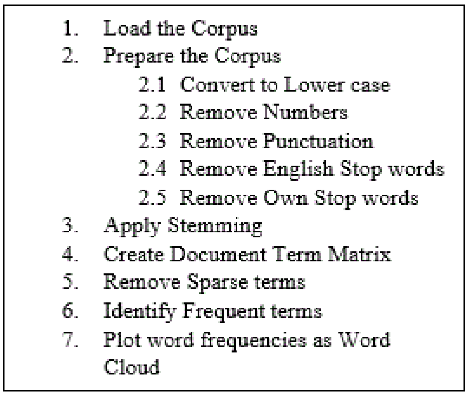
In this process, to study the title of this article, research publications of various authors from reputed national institutions, were collected. All these publications which are in different formats are converted into text files for generating the Corpus [2]. A Corpus is a collection of texts, usually stored electronically, and from which we perform our analysis. The pre-processing steps of text mining like, removing Numbers, converting to lowercase, removing punctuations et., are applied to this Corpus. Later, general stop words of English are removed from these documents.

A stop word has no extra sense other than it concern to grammar. Now, the Corpus is transformed into Document Term Matrix, a matrix with documents as rows and terms as the columns and a count of the frequency of words as the cells of the matrix. While generating the Document Term Matrix sparse terms (which are infrequent) are eliminated.

B. Word Clouds

A Word Cloud is a graphical representation of text / words in a two-dimensional space, in which the words are accentuated by occupying more prominence in the representation. The pseudo code for generating Word Clouds is explained in Figure 2.

Word Clouds gives an idea about the most frequent terms in the Corpus. Here the terms/words with frequency greater than or equal to 1000 are identified. They are: algorithm, can, data, database, optim, page, perform, plan, process, queri, result, set, system, time, transact, use, valu etc., To provide a quick visual overview of the frequent terms, a Word Cloud can be generated. This representation clearly shows the most frequently occurring words by hiding the infrequent terms or words as shown in the figure 3(a). So, the word cloud gives an idea of frequent words in a piece of text.



```
1. Load the Corpus
2. Prepare the Corpus
    2.1 Convert to Lower case
    2.2 Remove Numbers
    2.3 Remove Punctuation
    2.4 Remove English Stop words
    2.5 Remove Own Stop words
3. Apply Stemming
4. Create Document Term Matrix
5. Remove Sparse terms
6. Identify Frequent terms
7. Plot word frequencies as Word
   Cloud
```

Figure 2: Pseudo Code for generating Word Cloud

These word clouds can help us to draw some basic idea of area of research of each author. Just by seeing the diagrams, one can understand that his main area of research is Databases. But there are so many words in the figure like may, however, figure, will etc., which may not help us to form a statement regarding his area of specialization. So, it will be a better if we can form the cloud with the words from this document(s) which are related to some specific domain. For this we need the keywords of various domains.

In this work, keywords from various domains are collected from the Microsoft Academic Research website. These domains include Algorithms and Design(ad), Artificial Intelligence(ai), Bioinformatics(bio), Computer Graphics(cg), Databases(db), Data mining(dm), Networks(nt), Operating Systems(os), Software Engineering(se), Security & Privacy(sec). 30 - 50 keywords from each domain considered for this work. All pre-processing steps of text mining like removing Numbers, converting to lowercase, removing punctuations etc., are applied for these keywords and generated Keyword Corpus and then obtained the Keyword Document Term Matrix.

Most frequent keywords of each publication is identified by comparing the Document Term Matrix with Keyword Document Term Matrix. With these frequent keywords, a Domain Specific Word Cloud is generated as shown in Figure 3(b) for identifying the research area of the author in a better way.

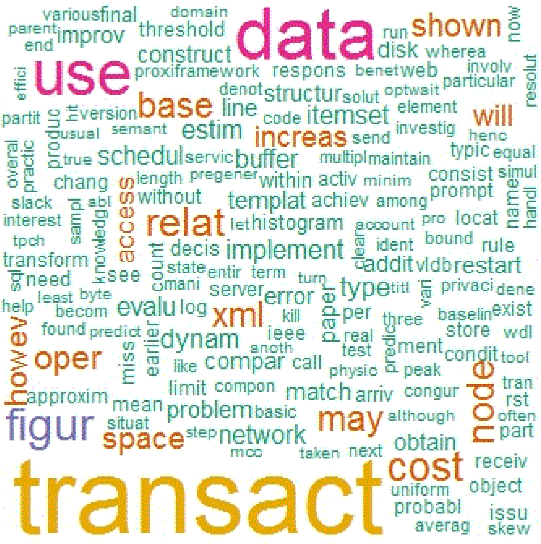


Figure 3(a): Word Cloud represents the frequent words of all Publications of Author 5

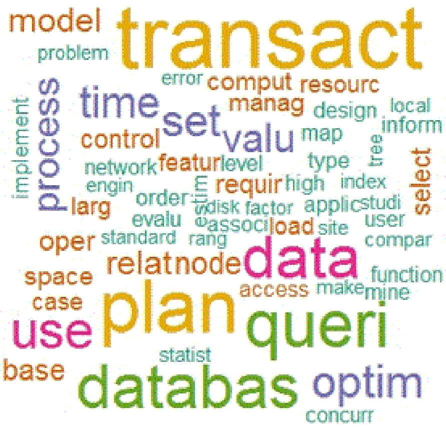


Figure 3(b): Word Could represents the frequent keywords |of all Publications of Author 5

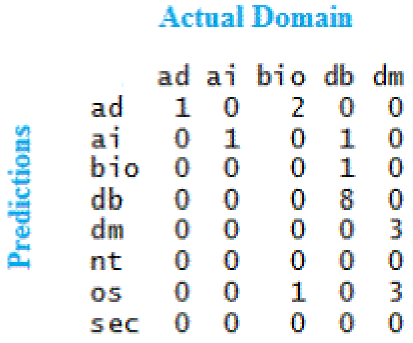
The most interesting research domain of an author can be identified through Domain Specific Word clouds. But in reality, the author may have worked on different domains. Now our research problem is how to identify the various research domains of any author on which they worked? For this we need to analyze each and every document. Representing each document as a Word cloud may not be possible and visualizing all together may become impractical.

C. K-Nearest Neighbor (KNN) Analysis

K-Nearest Neighbor classification (KNN) [5,6] is applied for document classification. It is a non-parametric method used for both classification and regression. In both the cases, it uses k nearest examples to estimate the outcome. In KNN classification, the outcome is a class membership. An object is classified by majority vote of its neighbors. In KNN regression, the outcome is the property value for the object. This value is the average values of its k nearest neighbors.

The KNN algorithm is applied on frequent terms of each publication, present in the Keyword Document Term Matrix. A confusion matrix[17] displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The confusion matrix is calculated and shown in the figure 3(a).

The accuracy[17] is the proportion of the total number of predictions that were correct. The accuracy of KNN algorithm for classifying research papers is up to 60% – 70%. To improve the accuracy, statistical methods PCA and LDA applied on domain ratio table, as discussed in next sections.



Actual Domain

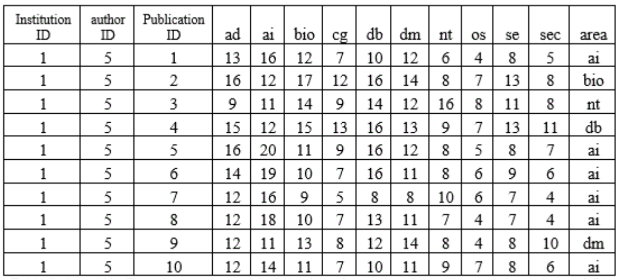
Predictions

	ad	ai	bio	db	dm
ad	1	0	2	0	0
ai	0	1	0	1	0
bio	0	0	0	1	0
db	0	0	0	8	0
dm	0	0	0	0	3
nt	0	0	0	0	0
os	0	0	1	0	3
sec	0	0	0	0	0

Table 2: KNN Confusion Matrix

D. Domain Ratio Table

Instead of representing each publication as a Word cloud, we extracted the proportion of keywords which will decide the area/domain of the publication. For each publication of an author, the percentages of frequent words match with respect to the given 10 domain keywords are calculated. The Domain Ratio Table, shown in Table 3, gives the percentage of keywords match in various domains for every publication of individual author. This table gives information only for 10 publications of Author5, at Institute 1. The actual number of publications are around 71.

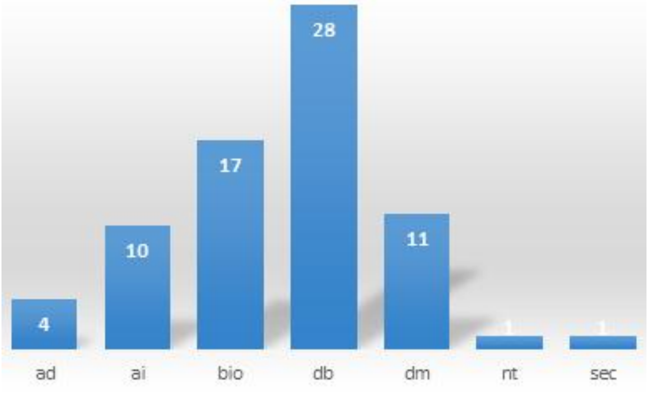


Institution ID	author ID	Publication ID	ad	ai	bio	cg	db	dm	nt	os	se	sec	area
1	5	1	13	16	12	7	10	12	6	4	8	5	ai
1	5	2	16	12	17	12	16	14	8	7	13	8	bio
1	5	3	9	11	14	9	14	12	16	8	11	8	nt
1	5	4	15	12	15	13	16	13	9	7	13	11	db
1	5	5	16	20	11	9	16	12	8	5	8	7	ai
1	5	6	14	19	10	7	16	11	8	6	9	6	ai
1	5	7	12	16	9	5	8	8	10	6	7	4	ai
1	5	8	12	18	10	7	13	11	7	4	7	4	ai
1	5	9	12	11	13	8	12	14	8	4	8	10	dm
1	5	10	12	14	11	7	10	11	9	7	8	6	ai

Table 3: Attaching Domain name to each publication of an author depending on keywords match %

An attempt is made to give brief information regarding the number of publications in each domain of Author 5 at Institution 1 with the help of a graph, shown in Graph 1. From this graph, it can be easily observed that this author has major contribution for the domains like Databases, Bioinformatics, Data mining, Artificial Intelligence, and also worked in Algorithms and Design, Networks, Security & Privacy.

We applied some statistical methods on this Domain Ratio Table for classifying the publications, through which we may identify the research specialization of an author.



Graph 1: Publications of author 5 after assigning the class label

E. Principal Component Analysis

As a part of unsupervised learning, we applied PCA for this data set. In unsupervised learning, we won’t use any class label. Principal component analysis (PCA) is a statistical method which uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principle components can’t be more than number of original variables. By this transformation, the first principal component has the largest possible variance and each succeeding component has the height variance possible under the constraint that is orthogonal to the preceding components. As the principal components are eigenvectors of the covariance matrix which is symmetric, they are orthogonal.

Applying PCA to Domain Ratio Table of Author5, the publications can be classified as shown in the Figure 4(a). From this figure, it is observed that the classifications are not identifying the area of specializations. Hence this experiment has been extended with supervised learning method.

F. Linear Discriminant analysis

Both Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation techniques that are commonly used for dimensionality reduction. PCA can be described as an “unsupervised” algorithm, since it “ignores” class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset. In contrast to PCA, LDA is “supervised” and computes the directions (“linear discriminants”) that will represent the axes that that maximize the separation between multiple classes.

In supervised leaning, we use class label for classification. As a part of supervised learning, we implemented a statistical method, Linear Discriminant Analysis (LDA). With LDA, a linear combination of features are found which characterizes or separates two or more classes. The following Figure 4(b) 5 is obtained by applying LDA on the data set given Table 3.

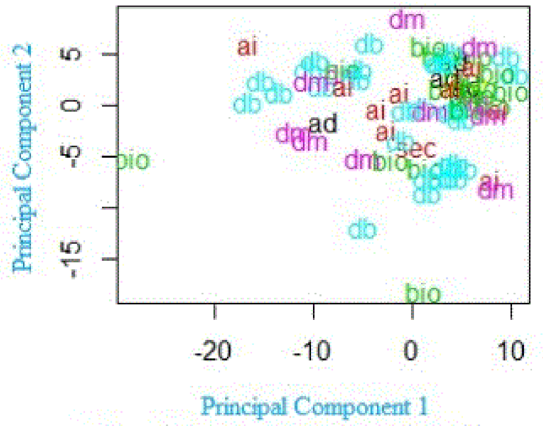


Figure 4(a): Author 5 Publications Plot using PCA

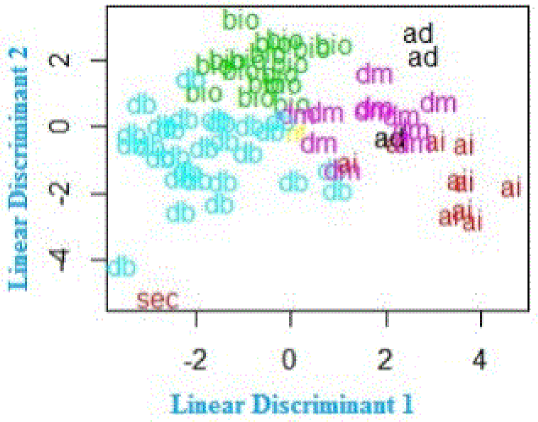


Figure 4(b): Author 5 Publications Plot using LDA

RESULTS

We observed that LDA classified better way in comparison to PCA. To check the classification error, we spited the data into parts and trained the classifier, and other part that will be used to test the classifier. For this process, we tried an 80:20 split. After applying LDA on both Training and Testing data, the results are shown in the following Figure 5(a) and Figure 5(b).

These figures, shows a reduction in misclassification error[17] when compared to PCA. The misclassification error is the proportion of the total number of predictions that were incorrect. Here we classified the publications of one single author. When we cross verified manually with his area of interest as given in the web site, we found this approach works fine. We tried the same approach for remaining authors. We got encouraging results as shown in the following Table 4.

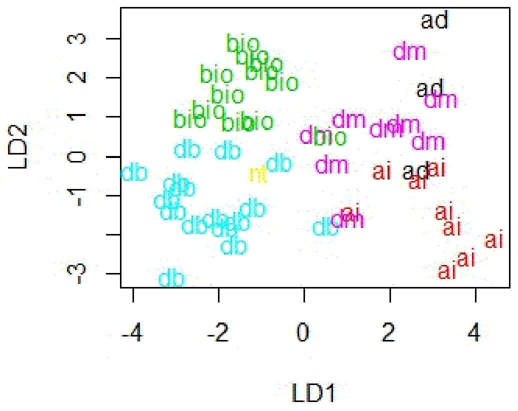


Figure 5(a): Author 5 - Training Set Plot using LDA

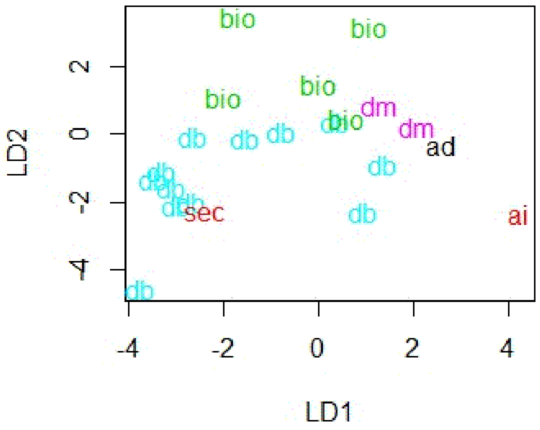
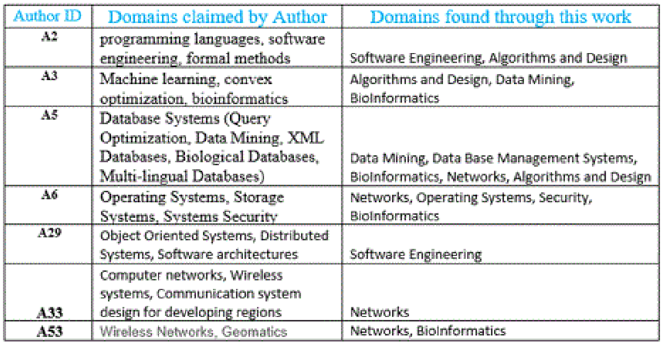


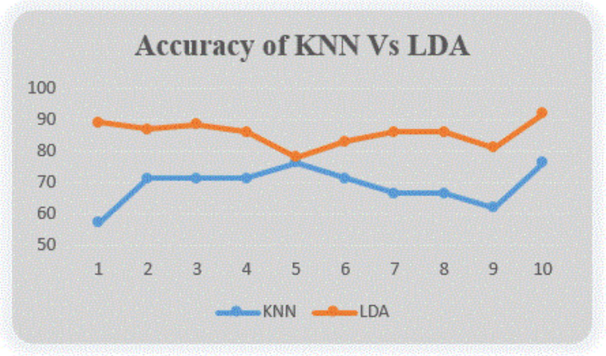
Figure 5(b): Author 5 - Test Set Plot using LDA



Author ID	Domains claimed by Author	Domains found through this work
A2	programming languages, software engineering, formal methods	Software Engineering, Algorithms and Design
A3	Machine learning, convex optimization, bioinformatics	Algorithms and Design, Data Mining, BioInformatics
A5	Database Systems (Query Optimization, Data Mining, XML Databases, Biological Databases, Multi-lingual Databases)	Data Mining, Data Base Management Systems, BioInformatics, Networks, Algorithms and Design
A6	Operating Systems, Storage Systems, Systems Security	Networks, Operating Systems, Security, BioInformatics
A29	Object Oriented Systems, Distributed Systems, Software architectures	Software Engineering
A33	Computer networks, Wireless systems, Communication system design for developing regions	Networks
A53	Wireless Networks, Geomatics	Networks, BioInformatics

Table 4: Domains Claimed Vs Domains Found

The following graph 2 shows the accuracy of KNN algorithm applied on Term Document Matrix Vs LDA on Domain Ratio Table. It clearly shows Linear Discriminate Analysis method gives better results compared to KNN algorithm.



Graph 2: Classification Accuracy

CONCLUSION AND FUTURE STUDIES

With the Domain Specific Word clouds, we can draw the major area of research of any author but, may not identify the other domains of the author on which he worked. In this experiment the unsupervised learning method PCA did not deliver good results. The supervised learning method LDA is classified better way compared with PCA. LDA method has few limitations like number of publications to classify. The results are satisfactory when the publication range is 20 to 100.

Our goal is to identify the specializations of an author and while we extended this to all publications, then we easily identify area changes in research in the past few years. We can also identify the emerging stream or research area with the number of publications came up in a specific research area. So, the we can raise or demand funds for that stream.

REFERENCES

[1] Christina V, Karpagavalli S, Suganya G, “A Study on Email Spam Filtering Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 12– No.1, December 2010

[2] Graham Williams, “Data Science with R Text Mining”, 9th June 2014, http://onepager.togaware.com/ for more OnePageR's.

[3] Quim Castellà, Charles Sutton, “Word Storms: Multiples of Word Clouds for Visual Comparison of Documents”, WWW’14, April 7–11, 2014, Seoul, Korea. ACM 978-1-4503-2744-2/14/04.

[4] J. Clark. “Clustered word clouds”, Neoformix, April 2008. URL http://www.neoformix.com/.

[5] Bang, S. L., Yang, J. D., & Yang, H. J., “Hierarchical document categorization with k-NN and concept-based thesauri. Information Processing and Management”, pp. 397–406, 2006.

[6] Thiago S.Guzella, Walimir M. Caminhas “ A Review of machine Learning Approches to Spam Filtering”,*Elsever* , Expert System with Applications-2009.

[7] Mardia, K., Kent, J. and Bibby, J. “Multivariate Analysis”, Academic Press,1979.

[8] A. Murua, W. Stuetzle, J. Tantrum, S. Sieberts, “Model Based Document Classification and Clustering”, International Journal of Tomography & Statistics”, Winter 2008, Vol. 8, No. W08;

[9] R. O. Duda, et al. (2001). Pattern Classification. John Wiley & Sons, Inc.

[10] A. Johnson & D. W. Wichern (1988). Applied Multivariate Statistical Analysis. Prentice Hall.

[11] S. Mika, Y.-H. Hu, J. Larsen, E. Wilson, & S. Douglas, “Fisher Discriminant Analysis with Kernels” Neural Networks for Signal Processing IX, pp. 41–48. IEEE.

[12] Mohamed Ali AlShaari, “Text Documents Classification Using Word Intersections”, IACSIT International Journal of Engineering and Technology, Vol. 6, No. 2, April 2014

[13] T. Lalitha and S. Meenakshi, “Text Mining Algorithm Discotex (Dis-Covery from Text Extraction) With Information Extraction”, Journal of Theoretical & Applied Information Technology, vol. 64, no. 2, **(2014)**.

[14] E. A. Calvillo, P. Alejandro, M. Jaime, P. Julio and F. T. Jesualdo, “Searching research papers using clustering and text mining”, Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on IEEE, **(2013)**.

[15] N. Arunachalam, E. Sathya, B. S. Hismath and M. M. Uma, “An Ontology Based Text Mining Framework for R&D Project Selection”, International Journal of Computer Science & Information Technology (IJCSIT), vol. 5, **(2013)**.

[16] V. Gupta and G. S. Lehal, “A survey of text mining techniques and applications”, Journal of emerging technologies in web intelligence, vol. 1, no. 1, **(2009)**, pp. 60-76.

[17] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction to Data Mining”, Addison-Wesley, 2005. ISBN : 0321321367.

[18] Hornik, Kurt (November 26, 2015). *"R FAQ"*. The Comprehensive R Archive Network. 2.1 What is R? *Retrieved 06-12-2015.*

[19] “Why Rstudio?”, *Rstudio.com., Retrieved 15-12-2015.*