IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 5, May 2016

Document Clustering for Authorship Analysis

Pooja Khandelwal[1], Aishwarya Mujumdar[2], Nandita Lonkar[3], Ankita Magdum[4], Dr. Rajesh S. Prasad[5]

Dept of Computer Engineering, NBN Sinhgad School of Engineering, Savitribai Phule Pune University[1,2,3,4]

Head of Dept of Computer Engineering, NBN Sinhgad School of Engineering, Savitribai Phule Pune University[5]

Abstract: The widespread use of computers and the advent of the internet has made it easier to plagiarize the work of others. Most cases of plagiarism are found in academia where documents are typically essays or reports. Detection of plagiarism can be manual or software assisted. Software assisted detection and analysis allows vast collections of documents to be compared to each other making accurate and successful detection.Document clustering is the application of cluster analysis to textual documents. It has applications in the automatic document organization, topic extraction and fast information retrieval. In technical publishing authorship of a work are claimed by those making intellectual contributions to the completion of the research described in the work. Analysis of this work is termed as authorship analysis.

Keywords: Clustering, Author identification, k-means.

I. INTRODUCTION

Data mining and text mining trends are increasing day by day,and now-a-days everything is available on the internet which in turn leads to the malicious and criminal activities.So there is need of the identity of the content which is available.[1] Authorship Analysis is basically the study of computational characteristics and features of documents written by an author. Various different methods for authorship analysis and identification of an author of a document is described in the paper. Authorship analysis is defined as " it is the process of identifying the characteristics of a work in order to identify the author".

Authorship analysis helps to distinguish the text written by various authors by comparing some of the textual features. Authorship analysis is applicable in many areas such as texts analysis in literature, online messages, program codes etc. It is broadly classified as below:

1.Authorship profiling or characterization determines the author's characteristics that produced a given piece of work . Characteristics included are educational background, gender etc.
2.Similarity detection collects different pieces of work and compares using plagiarism,whether they are composed by same author or not.

Due to the rapid increase in web applications and huge amount of text data it is necessary to focus on the study of authorship analysis.[2][8]

II. RELATED WORK

Earlier,researchers studied the word usage of different authors to identify authors, but the efficiency of this work is however limited as the word usage mostly depends on the topic of the article. Content free features were used to achieve generic authorship identification. For example, features such as length of sentences were used by Yule in 1938 and vocabulary richness was also considered by Yule in 1944. Afterwards, Burrows(1987) developed a word set having more than 50 high frequency words, which were then tested on federalist papers. Holmes in 1998 analyzed the use of short words and vowels. Such features based on words and characters requires intensive efforts in selecting the most appropriate set of words that best distinguish a given set of authors (Holmes & Forsyth,1995), and sometimes, when applied to a wide range of applications those features are not reliable discriminators. Function word usage determines how to syntactically form a sentence. Rooted from linguistic research, part of speech (POS) and punctuation usage are other syntactic features which have been applied to authorship research. The different features type like structural features attracted more attentions. While writing any document people have different habits like paragraph length,use of indentation,,use of signature, can be strong authorial evidence of personal writing style.[3]

By using statistical methods, accurate calculations can be performed and this has helped to successfully deduce author identity in the past . This paper consisted of 146 essays based on politics written by a number of authors like Alexander Hamilton, James Madison, John Jay. This was undisputedly the best contribution in Author Identification. Their approach for Author Identification was based on Bayesian statistical analysis of the occurrences of prepositions and conjunctions like or, to, but, and, etc. Thus, it was helpful in classifying the authors accordingly. The research in Author Identification then saw a tremendous attention and speed up rapidly.[3] [8]

III. METHODS

A.Data Collection

Data Collection is the process of gathering information related to a particular topic, which helps an individual to answer questions related to that topic. The Appropriate Data Collection is essential to maintain its integrity.

DOI 10.17148/IARJSET.2016.3510

The Data collection here, in this case is to collect papers or documents authored by different authors and store it in database of the system.[4]

Documents were collected from different authors. As expected in cooperative work, some authors wrote many papers while others wrote few. Documents were stored in the database such that each author can upload maximum 10 files. If 11th file is uploaded by the author, system will not accept the document and will display the message that only 10 files can be uploaded and no more files can be uploaded. Documents uploaded must be in text format i.e. ".txt" format. In this system there is a limit on size of file. File having size more than 5000bytes cannot be uploaded.[4]

B.Cluster Evaluation

Cluster Evaluation is the process of grouping a set of objects in such a way that objects within a same group are more similar to each other than the objects belonging to different groups.[5]

C.Min-Max K-means algorithm:

We propose min-max k-means algorithm to minimize the maximum intra-cluster variance objective. Weights are assigned to the clusters relative to their intra-cluster variance. Our methods prevent the occurrence of clusters with large intra-cluster variances in the solution. This method reveals high quality solutions, regardless of the initialization Min-Max k-means constitutes a sound approach for initializing k-Means.[5]

D.Document Clustering

Document Clustering is the application of converting cluster analysis to textual documents.[6]

Clustering is referred as an automated learning technique whose goal is to group a set of objects into subsets or clusters.

Document in same cluster should be as similar as possible, while documents in a cluster should be as dissimilar as possible from documents in the other clusters.

Document Clustering has its applications topic extraction, automatic document organization and fast information retrieval or filtering. Here in this system, documents are clustered based on User ID. Each user on uploading a document, gets an id of a particular author. So documents are clustered according to the User ID.[6]

E.Algorithm

Min Max K means:

K means is used to minimize the distance between clusters, but the problem with this algorithm is that if the initialization is not correct the algorithm may produce inaccurate or even faulty results. We therefore use a part of the Min Max algorithm to overcome the initialization drawback of K-Means.[7]

The K means algorithm:

Step 1: Choose the number of clusters.

Step 2: Set the initial partition, and the initial mean vectors for each cluster.

Step 3: For each remaining individual...

Step 4: Get averages for comparison to the Cluster 1: Add individual's A value to the sum of A values of the individuals in Cluster 1, then divide by the total number of scores that were summed. Add individual's B value to the sum of B values of the individuals in Cluster 1, then divide by the total number of scores that were summed.

Step 5: Get averages for comparison to the Cluster 2: Add individual's A value to the sum of A values of the individuals in Cluster 2, then divide by the total number of scores that were summed. Add individual's B value to the sum of B values of the individuals in Cluster 2, then divide by the total number of scores that were summed.

Step: 6 If the averages found in Step 4 are closer to the mean values of Cluster 1, then this individual belongs to Cluster 1, and the averages found now become the new mean vectors for Cluster 1. If closer to Cluster 2, then it goes to Cluster 2, along with the averages as new mean vectors.

Step 7: If there are more individual's to process, continue again with Step 4. Otherwise go to Step 8.

Step 8: Now compare each individual's distance to its own cluster's mean vector, and to that of the opposite cluster. The distance to its cluster's mean vector should be smaller than it distance to the other vector. If not, relocate the individual to the opposite cluster.

Step 9: If any relocations occurred in Step 8, the algorithm must continue again with Step 3, using all individuals and the new mean vectors. If no relocations occurred, stop. Clustering is complete. Again, in case the algorithm never settles on a final solution, it may be a good idea to implement a maximum number of iterations check.[7]

Proposed algorithm

1. Initialize the min and the max coordinate.
2. Pass the No of clusters as a parameter.
3. Create Points.
4. Calculate the distance between 2 points.
5. Set random centroids.
6. Use the K-means algorithm
7. Add in new data one at a time recalculating centroids with each new data.
8. Clear the state of the cluster.
9. Assign points to the closest cluster.
10. Calculate the new centroid.
11. Calculate the total distance between the new and the old centroids.

DOI 10.17148/IARJSET.2016.3510

12.iterate and repeat the steps from 7 till no data changes are possible and no new data can be added to a new cluster.[7]

IV. EXPERIMENTS AND RESULT

Following are the steps which are to be followed:
i.The user creates a new account by filling all the necessary information in the registration table and then log in.(If the user is already registered then directly click on login)

ii.Once the user is logged in he can upload the documents, for uploading documents the user has to give the following things as input: File name, Author name/names and then select the text document to be uploaded and click on add button. The uploaded files can be seen below.

iii.In the back-end the features are extracted from the documents, the features we have taken into consideration, here are stop-words, comma, double quoted words, colon, semicolon etc. The count of total numbers of features is taken and saved and then the files are divided into clusters we have created here three clusters.

iv.The user can add more files by clicking on add more files. As soon as the files are uploaded at the back-end features are extracted and clusters of documents are formed.

iv.If the user wants to check author of any unknown documents, then he can click on check author which will take him to the new page. There the user needs to upload the text document of the unknown author and then click on find author.

v.The search for author works in the following way:
The features are extracted from unknown documents and the total count is taken, then the count is compared with the counts of other documents in the database. We have considered the variation in count with +/- 15. So all the documents with this range are only considered in serach and others are skipped. The content matching of documents is done line by line.

v.If the author is present in the database, then it will display the author name with the content matching percentage. Else he will display no author found for this document.

Below shown are some test cases tested for application created by us.

Table 1: Test Cases

Test Id	Test case	Input	Output	Pass/ Fail
1	To check whether the user can register/login.	Register as new user filing all necessary information and if already registered login.	You are successfully logged in.	Pass
2	To check if files are getting uploaded successfully.	File name, author name and select path of text document.	The file is successfully added and can be seen.	Pass
3	To check only text files is uploaded.	Input any pdf file.	Sorry. Only text files can be uploaded.	Pass
4	To check file more than 5000 bytes is not uploaded.	File more than 5000 bytes is selected.	Please upload a file size less than 5000 bytes.	Pass
5	To check if a single user cannot upload more than 10 files.	Uploading the 11th file.	More than 10 files cannot be uploaded.	Pass
6	To find the correct author of the document.	Upload the file whose author is to be found.	Content matched. Author name and matching percentile displayed	Pass

V. CONCLUSION

A lot of study has been done in the field of text mining.We have constructed a system for author identification. Although this system has some limitations, we have put all the efforts to minimize those limitation and increase its efficiency. The system will give outputs in the form of author name and the content matching percentage. The author whose documents are previously stored in the database and whose matching percentage is high will be displayed in the output. Thus we can determine by this author of document using document clustering.

REFERENCES

[1] Sara El Manar El Bounanai and Ismail Kassou, "Authorship Analysis Studies: A Survey", International journal of Computer Applications (0978887), Volume 86-12, January 2014.

[2] Rong Zheng and Jiexun Li and Hsinchun Chen and Zan Huang, "A Framework for Authorship Identification of Online Messages:Writing-Style: Features and Classification Techniques," Journal of the American Society for Information Science and Technology.

[3] Moshe Koppel and Jonathan Schler, "Exploiting Stylistic Idiosyncrasies for Authorship Attribution", IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, 2003.

[4] Vishal Chandani, Ninad Deshmane, Kshitij Buva, Suvrat Apte, Dr. R.S. Prasad, "Author Identification Method using Hybrid Technique", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 4, Issue 04,April-2015.

[5] Ka Yee Yeung and Walter L. Ruzzo, "Am empirical study on principal component analysis for clustering gene expression data",USA, May 3, 2001.

[6] Daniel Berry and Edward Sazonow, "Clustering Technical Documents by Stylistic Features for Authorship Analysis," Proceedings of IEEE Southeast Con 2015, April 9-12, 2015,Florida.

[7] Baolan Yuan,Wanjun Zhang and Yubo Yuan," A Max-Min Clustering Method For k-Means Algorithm Of Data Clustering",Journal Of Industrial And Management Optimization,Vol. 8,Number 3,August 2012.

[8] R. S. Prasad, U. V. Kulkarni, J. R. Prasad, "A Novel Evolutionary Connectionist Text Summarizer (ECTS),", 2009, IEEE Xplore.