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

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 10, October 2015

Document Clustering for Authorship Analysis

Pooja Khandelwal[1], Aishwarya Mujumdar[2], Nandita Lonkar[3], Ankita Magdum[4]

Computer Department, NBN Sinhgad School of Engineering, Ambegaon[1, 2, 3, 4]

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 automatic document organization, topic extraction and fast information retrieval.In technical publishing authorship of a work is 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.The methods include data collection, extracting features, document clustering, cluster evaluation Rand index

Keywords: Authorship Analysis, Rand index, Document clustering, Stylistic features.

I. INTRODUCTION

A)AUTHORSHIP ANALYSIS

Authorship analysis is the process of examining the characteristics of a piece of work in order to draw conclusions on its authorship. Authorshipanalysis has its roots in a linguistic research area called stylometry which refers to statistical analysis of literally style.[1]
Authorship analysis distinguishes text written by different authors by comparing some textual features like stylistic features.
Authorship analysis can be broadly classified as follows:

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

Major areas of applications where authorship analysis is used are analyzing texts in the literature, program codes and online messages.
Due to vast growth in web applications and social networks, authorship analysis is mostly focusing on online messages such as emails, blogs, forum, etc.
The features of Authorship analysis are:

(i) Lexical: Average word/Sentence length and vocabulary richness.
(ii) Syntactic: Frequency of Function words and use of punctuation.
(iii) Structural: Paragraph length and indentation.
(iv) Content-specific: Frequency of keywords. [1]

B) STYLISTIC FEATURES

Earlier authorship attribution researchers made use of a variety of statistical methods for identifying stylistic discriminators, Machine learning methods are applied to authorship attribution these days. For example Matthews & Merriam 1993, Holmes & Forsyth 1995, Stamatatos et al., 2001, De Vel et al., 2001.
Recent machine learning work and earlier Stylometric work mainly focused on candidate discriminator sets. Other features which used earlier were complexity based: average length of sentence, average length of word, token ratio and so forth. Use of Syntactic and quasi-syntactic features were facilitated due to recent technical advances in Part Of Speech (POS) tagging and automated parsing. [2]
The way human experts work on real life authorship attribution problems is different. Three classes of features are considered for the purpose of our experiments.
The stylistic features are as described below:

1. Lexical: We used a standard set of 480 function words. These words were filtered by using the information gain ranking on each training corpus and choosing the top 200 words.
2. Part-of-speech tags: To tag each word 59 POS tags were used.The feature of the frequencies of all POS bi-grams which appeared at least three times in the corpus.
3. Idiosyncratic Usage: Various types of idiosyncratic usage were considered-syntactic, formatting and spelling. For example, frequency of sentence fragments were checked, run on sentences, unbroken sequences of multiple question marks and other punctuation, words mentioned in caps and so on.

Finally, we can conclude that the use of the various features greatly enhances the accuracy of the results in comparison with methods which have generally been used in automated
Authorship attribution.[2]

C) DOCUMENT CLUSTERING

Document clustering is the application of converting cluster analysis to textual documents.
Clustering is an automatic learning technique aimed at grouping a set of objects into subsets or clusters.
Goal of document clustering is to create clusters that are coherent internally, but substantially different from each

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 10, October 2015

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

Algorithm

1. Cluster affinity Search Technique (CAST)
CAST is an algorithm which is proposed by Ben-Dor and Yakhini. It is used to cluster data according to the gene affinity.

The elements used in the CAST algorithm are:

1. gene-g
2. Cutoffparameter-t
3. C_{open}- cluster under consideration
4. Affinityfor a gene-a(g)

The input to the algorithm includes pair wise similarities of genes and t.
The cutoff parameter takes a real value between 0 and 1. C_{open} is the current cluster which is initially empty and genes are added to it.When a new cluster is created the initial affinity of all genes is 0.
A gene is said to have high affinity if a(g)>t mod $C_{open.}$
If it does not satisfy this measure it is said to have low affinity.
Genes with high affinity are added to the cluster and those with low affinity are removed from the cluster. This is done alternately.
When no more genes can be added or removed from the cluster C_{open} is closed.
The algorithm continues to run till every gene is assigned to a cluster and C_{open} is closed.

2. Hierarchical clustering
In this method clusters are analyzed by building a hierarchy.
There are 2 strategies for hierarchy namely agglomerated and divisive.
1. Agglomerative- This is a bottom up approach. Two clusters with the highest cluster similarity are merged. This process is continued till the desired numbers of clusters are produced.
2. K-means clustering
In this method there are n observations that are partitioned into k clusters.
K is given as an input to the algorithm.
Center of a cluster is the centroid.
Clustering results from average link are used to decide the initial centroids.
Each object is assigned to the cluster using the closest Euclidean distance.
Once all objects are assigned, centroids of k clusters are found.
This process is repeated till no objects are moved between clusters.

Similarity measures:-A similarity measure must be determined before clustering.

The measure tells degree of closeness or separation of objects and should correspond to characteristics that are responsible for creation of clusters.[3]

D) RAND INDEX

Rand index is a measure of similarity between two data clusterings. Rand index is related to the accuracy, but it is even applicable when class labels are not used. [3]

Definition

Given a set of n elements $S = \{o_1, \ldots, o_n\}$ and two partitions of S to compare $X = \{X_1, \ldots, X_r\}$, a partition of S into r subsets, and
$Y = \{Y_1, \ldots, Y_s\}$, a partition of S into s subsets, define the following:
a, the number of pairs of elements in S that are in the same set in X and in the same set in Y
b, the number of pairs of elements in S that are in different sets in X and in different sets in Y
c, the number of pairs of elements in S that are in the same set in X and in different sets in Y
d, the number of pairs of elements in S that are in different sets in X and in the same set in Y

$$R = \frac{a + b}{a + b + c + d}$$

$$R = \frac{a + b}{\frac{n}{2}}$$

Intuitively, a + b can be considered as the number of agreements between X and Y and c + d as the number of disagreements between X and Y.
The rand Index has a value between 0 to 1, with 0 indicating that two clusters do not agree on any pair of points and 1 indicating that data clusters are exactly the same. [4]

II. LITERATURE SURVEY

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. To achieve generic authorship identification we use “content-free” features.

For example features such as sentence length were used by Yule in 1938 and vocabulary richness was used by Yule in 1944. Later, Burrows (1987) developed a set of words having more than 50 high-frequency words, which were then tested on Federalist papers. Holmes (1998) analyzed the use of “shorter” words and “vowel words”. Such word-based and character-based features require intensive efforts in selecting the most appropriate set of words that best distinguish a given set of authors (Holmes & Forsyth, 1995), and sometimes those features are not reliable discriminators when applied to a wide range of applications.[5]

Function-word usage determines how to syntactically form a sentence. Rooted from linguistic research, part of speech (POS) and punctuation usage are other important 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. This method is mostly used in online documentation which has less information content.

The different techniques used were:

1. Statistical analysis: This method was used for calculating document statistics based on metrics to examine similarity between various pieces of work done by using characteristics of the author.
2. Machine Learning: Classification methods are used to predict the author of a piece of work based on set of metrics.[5]

III. CONCLUSION

The paper describes authorship analysis using measures like rand index, algorithms for document clustering, similarity metrics and stylistic features for analyzing technical documents. The study of features extracted from technical documents including stylistic features and number of words, characters and other patterns is done. Study about lexical, syntactic and structural features is made. The methods for identification of the real author to resolve ambiguity is explained.

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] Moshe Koppel and Jonathan Schler, "Exploiting Stylistic Idiosyncrasies for Authorship Attribution", IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, 2003.
[3] Ka Yee Yeung and Walter L. Ruzzo, "Am empirical study on principal component analysis for clustering gene expression data", USA, May 3, 2001.
[4] 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.
[5] 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, 57(3):378-393, 2006.