Authorship Attribution using Unsupervised Clustering Algorithms on English C50 News Articles

Dr. O Srinivasa Rao¹, Dr. N V Ganapathi Raju², Dr. Y. Srilalitha³, Mrs. P. Bharathi⁴

¹Associate Professor, Dept of CSE, JNTUK, Kakinada, India
²Professor of CSE, GRIET, Hyderabad, India
³Assistant Professor of IT, GRIET, Hyderabad, India

Abstract: The aim of the authorship attribution is identifying the author of an unknown/anonymous document. Many earlier researches used authorship attribution as a multi class single labelled text classifier problem. However, in several applications it is not easy or even possible to find such labeled data and it is necessary to build unsupervised attribution models that are able to estimate similarities/differences in personal style of authors. The present paper experiments authorship attribution as a clustering task using various unsupervised clustering algorithms like K-means, Mini Batch K-means and Ward Hierarchical clusterings and our authorship clustering algorithm achieves 97% of clustering accuracy in clustering C50 English news groups articles.

Keywords: authorship clustering; unsupervised algorithms; C50 data set.

1. INTRODUCTION

Today the rapid growth of the electronic documents in internet in the form of emails, blogs, social networking, news groups, twitter, Facebook, etc. has created multitude ways to share information across the World Wide Web. The main reason for this is rapid development and proliferation of internet technologies at very low cost. This phenomenal growth of accessing information has created problems in author attribution, because some people circulate some of the articles and sometimes combine two or more articles in the social media. Hence authorship attribution has become an emerging research area in information retrieval research.

Authorship attribution has many applications in diverse areas including: intelligence, criminal law, civil law, computer forensics, in addition to the traditional application to literary research. Today, finding an anonymous author is not only the application of authorship attribution but it also finds a broad range of application, in areas such as, information retrieval, computational linguistics, cybercrime, natural language processing, and attribution of authors on the Internet etc.

Authorship attribution can be classified into Authorship Identification, Authorship Verification, Authorship Profiling and Authorship Clustering. Authorship identification means, Given a set of candidate authors for whom some texts of undisputed authorship exist, aim is to find the correct author. Authorship verification means, Given a set of documents by a single author and a questioned document, the authorship verification is to determine if the questioned document was written by that particular author or not [11, 12].

Author profiling distinguishes between classes of authors studying their socialist aspect, that is, how language is shared by people. This helps in identifying profiling aspects such as gender, age, native language, or personality type.

The author clustering task is more demanding than the classical authorship attribution problem. Given a document collection the task is to group documents written by the same author such that each cluster corresponds to a different author. The number of distinct authors whose documents are included is not given [7, 11].

The present paper is organized as follows. The literature is presented in section two. The section 3 and 4 describes the methodology and results and discussion. The conclusions are presented in section 5.

2. LITERATURE SURVEY

Generally, authorship attribution is a multi-class classification task where more an unknown/anonymous document is classified to the correct author among many authors based on stylistic features using supervised machine learning algorithms. Most of the earlier researchers treated authorship attribution as a classification task. However, there are multiple cases where authorship information of documents either does not exist or is not reliable. In such a case unsupervised authorship attribution should be applied where no labeled samples are available.
Douglas Bagnall [1] used recurrent neural networks for authorship clustering on very short and about disparate topics and observed statistically significant predictions regarding authorship and it is difficult to group documents into definite clusters with high accuracy.

Mirco Kocher [2] evaluated an effective unsupervised author clustering authorship linking model called SPATIUM-L1 and suggested strategy can be adapted without any problem to different in different genres by considering m most frequent terms of each text (m at most 200) and applying a simple distance measure to determine whether there is enough indication that two texts were written by the same author.

Mansoorizadeh, Muharram, et al. [3] proposed a two-step unsupervised method in order to perform author clustering. The approach combines different feature spaces and use them to cluster documents based on their authors. Then, we rank documents based on their cosine similarity using new set of feature which are different from the set we use for clustering.

Vartapetiance et al. [4] involved by generating clusters within larger sets of documents (n<=100) for an unknown number of distinct authors, where each set is in English, Dutch or Greek. The results that were achieved are not expected to be particularly remarkable due to substantial limitations on our time around the task.

Zmiycharov, Valentin, et al. [5] developed for the Authorship Link Ranking and Complete Author Clustering for a given a document collection with a combination of classification and agglomerative clustering with a rich set of features such as average sentence length, function words ratio, type-token ratio and part of speech tags.

Verga, Patrick, et al. [6] adapted the impostor method of authorship verification to authorship clustering using agglomerative clustering and, for efficiency, locality sensitive hashing and validated methods and shown on authorship clustering task, they shown that the impostor similarity method clearly outperforms other techniques on the blog corpus.

Sittar, Abdul, Hafiz et al. [8] proposed approach for author diarization task using various types of stylistic features which include lexical features, to uniquely identify an author. Furthermore, to find anomalous text within a single document, ClustDistmethod used, finally, clusters were generated by using simple k-means clustering algorithm. Experiments were performed both on training and testing data sets. It has been observed that by changing the text fragments length, promising results can be achieved.

Sari, Yunita et al. [9] presented Author Clustering task using simple character n-grams to represent the document collection and then ran K-Means clustering optimized using the Silhouette Coefficient. Their system yields competitive results and required only a short runtime. Character n-grams can capture a wide range of information, making them effective for authorship attribution.

Layton et al. [13] named their methodology NUANCE, for n-gram Unsupervised Automated Natural Cluster Ensemble and testing indicates that the derived clusters have a strong correlation to the true authorship of unseen documents.

3. METHODOLOGY


3.1 Algorithm for Authorship Clustering

The algorithm consists of five steps as given below.

Step 1: Data collection Step: The present paper uses C50 news group’s data set of same genre for authorship clustering. The present paper is implemented on 200 news groups’ articles collected from four different authors (per author 50 documents).

Example the document considered is: - “After arriving home from the movies one night, I decided that I was not going to be a moviegoer anymore. I was tired of the problems involved in getting to the movies and dealing with the theater itself and some of the patrons. The next day I arranged to have cable TV service installed in my home. I may now see movies a bit later than other people, but I'll be more relaxed watching box office hits in the comfort of my own living room.”

Step 2: Pre-processing Step:

2.1 In this step, corpus is converted to UTF-8 Unicode format.

2.2 In this step numbers, special characters, commas and full stops are eliminated from the corpus.

2.3 Removed stop words from the corpus but did not use stemming method on the corpus.

Example: After preprocessing the document converted is: “arriving home movies one night decided not going moviegoer anymore tired problems involved getting movies dealing theater patrons next day arranged cable TV service installed home may now see movies bit later than other people but more relaxed watching box office hits comfort own living room “
Step 3: **Document Representation:** The corpus has been represented as Word N-grams. N=1, 2, 3 are considered for the experimentation purpose.

Step 4: **Vector Space Model Representation:** Calculate Term Frequency (TF) and Inverse Document Frequency (IDF) for every document from Step 3 and represent all the documents of the authors as Vector Space Model.

Example: - Sample N-grams for the text considered and their tf-idf scores in descending order.

<table>
<thead>
<tr>
<th>N-gram</th>
<th>tf-idf</th>
</tr>
</thead>
<tbody>
<tr>
<td>movies</td>
<td>0.27975144247209416</td>
</tr>
<tr>
<td>home</td>
<td>0.1865096164806276</td>
</tr>
<tr>
<td>home movies</td>
<td>0.1865096164806276</td>
</tr>
<tr>
<td>anymore</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>anymore tired</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>anymore tired problems</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>arranged</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>arranged cable</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>arranged cable tv</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>arriving</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>arriving home</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>arriving home movies</td>
<td>0.09325048082403138</td>
</tr>
<tr>
<td>bit</td>
<td>0.09325048082403138</td>
</tr>
</tbody>
</table>

Step 5: **Clustering Step:** Use Unsupervised classification algorithms for K-means, Mini Batch K-means and Ward Hierarchical Clustering for the Authorship Clustering. The scores for calculating % of clustering rate can be calculated as follows:-

\[ \text{Score} = \frac{N_1}{N_2} \times 100 \]

**Algorithm for Authorship Clustering**

4. RESULTS & ANALYSIS

The corpus of C 50 news groups articles are collected from the internet. Totally 50 news groups articles are collected from each author consists of Aaron Pressman, Benjamin Kang Lim, David Lawder and Darren Schuettler respectively. The implementation of authorship clustering has been done with a Python language using Scikit learn module.

- The function used for K-means is
  
  Syntax: - KMeans(n_clusters= 4)

  Where n_clusters represents the number of clusters to form as well as the number of centroids to generate.

- The function used for Mini Batch K-means is

  Syntax: - MiniBatchKMeans(n_clusters= num, batch_size=bsize)Where n_clusters represents the number of clusters to form as well as the number of centroids to generate and batch_size represents Size of the mini batches.

- The function used for Ward Hierarchical Clustering is

  Syntax: - AgglomerativeClustering (n_clusters= num, linkage="ward")

Where n_clusters represents the number of clusters to form as well as the number of centroids to generate and Linkage represents the linkage criterion determines which distance to use between sets of observation. “Ward” minimizes the variance of the clusters being merged.

Table 1 shows the percentage of cluster rate for various clustering algorithms and visualization of the clustered documents of various authors are shown for the K-means algorithm is shown in Fig.1.

<table>
<thead>
<tr>
<th>Author Name</th>
<th>% of Cluster Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K-means</td>
</tr>
<tr>
<td>Aaron Pressman</td>
<td>100</td>
</tr>
<tr>
<td>Benjamin Kang Lim</td>
<td>100</td>
</tr>
<tr>
<td>Darren Schuettler</td>
<td>96</td>
</tr>
<tr>
<td>David Lawder</td>
<td>92</td>
</tr>
</tbody>
</table>
The average cluster rate of the K-means, Mini Batch K-means and Ward Hierarchical algorithms are shown in table 2.

Table 2: Average performance of unsupervised clustering algorithms

<table>
<thead>
<tr>
<th>Clustering algorithms</th>
<th>K-means</th>
<th>Mini Batch K-means</th>
<th>Ward Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>97</td>
<td>85</td>
<td>85</td>
</tr>
</tbody>
</table>

From the above table 2 it is observed that the average cluster rate of the K-means clustering outperforms Mini Batch K-means and Ward Hierarchical clustering algorithms.

5. CONCLUSION & FUTURE WORK

This paper explores authorship clustering on English news groups articles using three unsupervised clustering algorithms. It is observed that the proposed approach showed best results with 97 percent of accuracy using K-Means clustering. This motivates us to further study clustering techniques and attain domain independent accuracy. In our
future work we try to explore Semantic knowledge and study the different author ship Machine learning techniques on Large Scale.

REFERENCES