UGC Approved Journal

IARJSET

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

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 8, August 2017

Performance Evaluation of Unsupervised Algorithms on Morpheme based Authorship Clustering

Dr. O Srinivasa Rao[1], Dr. N V Ganapathi Raju[2], Dr. Y Vijaya Latha[3], P. Vivek Varma[4]

Associate Professor, COE, JNTUK, Kakinada, Andhra Pradesh, India[1]

Professor, Dept of CSE, GRIET, Hyderabad, Telangana, India[2]

Professor, GRIET, Hyderabad, Telangana, India[3]

Dept. of IT, GRIET, Hyderabad, India[4]

Abstract: The aim of authorship attribution is to identify the author of an anonymous document. Earlier, many types of research used authorship attribution as a multi-class single labeled text classifier problem. However, in several applications, it is neither easy nor possible to find such labeled data so it is necessary to build unsupervised attribution models that are able to estimate similarities or differences in personal style of authors. The present paper experiments authorship clustering using morpheme-based N-gram on unsupervised clustering algorithms like K-means, Mini Batch K-means, and Ward Hierarchial clusterings. The performance of the clustering algorithms is evaluated using silhouette coefficient and calculated B-cubed F-score and found that K-means algorithm achieves better clustering performance on C 50 news groups data set.

Keywords: morphemes; authorship clustering; silhouette coefficient; BCubed F-score.

1. INTRODUCTION

Today, the expeditious growth of the electronic documents 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 reasons for the proliferation of internet technologies are, it’s faster,cheaper and usage of high performance digital tools opens up the possibility of profound tranformations. 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. In today’s technical world, 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, cyber crime, natural language processing, and attribution of authors on the Internet etc.

Authorship attribution is 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, the 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-classification task where various anonymous documents are categorized to the authentic author among many authors based on stylistic features using supervised machine learning algorithms.

DOI10.17148/IARJSET.2017.4808

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, 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] analyzed constructive unsupervised author clustering authorship linking model called SPATIUM-L1 and suggested a strategy that can be adapted without any problem to different genres by considering m most frequent terms of each text (m at most 200) and applying a simple distance measure to verify whether there is enough evidence that two texts was 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 uses them to cluster documents based on their authors. Then, we rank documents based on their cosine similarity using a new set of feature which is 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 achieved were 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 like average sentence length, function words ratio, type-token ratio and part of speech tags.

Verga, Patrick, et al. [6] adopted 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 showed 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, ClustDist method 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

The present paper utilized unsupervised clustering methods for authorship attribution. The paper considers K-Means, Mini Batch K-Means, and Hierarchical clustering on C 50 newsgroup data set of same genre for authorship clustering.

3.1 Algorithm for Authorship Clustering

The algorithm consists of five steps as given below.

Step 1: **Data collection Step**: The current paper uses C50 newsgroup's data set of the same genre for authorship clustering. That is implemented on 200 newsgroups' articles collected from four different authors (per author 50 documents).

Step 2: **Pre-processing Step**:

2.1 In step 2.1 numbers, special characters, commas and full stops are eliminated from the corpus.

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

Step 3: **Document Representation:** The corpus has been represented as Morpheme N-grams.

Where N=2 to 5 are considered for the experimentation purpose. A Morpheme is the smallest grammatical unit of a language where it is not identical to a word, and the principal difference between the morpheme and word but a can standalone, is that a morpheme may or may not stand alone.

The algorithm is implemented using Python 3.5. The morpheme representation of the C 50 newsgroups' data set is implemented using polyglot package.

For example :

```
w = “preprocessing”
w = Word(w, language="en")
```

morpheme = w.morphemes
print(morpheme)

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.
In the experimentation, scikit-learn package of python is used for generating tf-idf matrix and to cluster the documents using K-means algorithm. The parameters used to calculate Tfidf matrix is : tfidf_vect= TfidfVectorizer(max_features=500, use_idf=True, tokenizer=tokenizer.morpheme_gen, ngram_range=(2,5), analyzer="word") where max_features represents most frequent 500 features, ngram_range generates morpheme based word N-grams from 2 to 5.

Step 5: **Clustering Step**: Use Unsupervised classification algorithms for K-means, Mini Batch K-means and Ward Hierarchical Clustering for the Authorship Clustering.
Step 6: **Performance Evaluation**: To evaluate the performance of the authorship clustering Silhouette Coefficient and B-cubed F-Score are calculated.
The Silhouette Coefficient is defined for each sample and is composed of two scores a and b: where a: The mean distance between the first score 'a' and all other points in the same class and b: The mean distance between the second score 'b' and all other points in the next nearest cluster.The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

B-cubed F-Score is calculated in the following way-
N1 = Number of documents justified to be of topic T in cluster X
N2 = Number of documents in cluster X
N3 = Number of documents justified to be of topic T in entire hierarchy

$$\text{Precision}(X, T) = N\,l / N2$$

$$\text{Recall}(X, T) = Nl / N3$$

$$F = \frac{2PR}{(P + R)}$$

Where F=F-Measure, P=precision, and R=recall.

The overall F Measure is the average F measure of all clusters.

Algorithm for Authorship Clustering

4. RESULTS & ANALYSIS

The corpus of C 50 newsgroups' articles is collected from the internet. Totally, 50 newsgroups' 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 by 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 form along with 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 along with 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 along with the number of centroids to generate and Linkage represents the linkage criterion to determine, which distance to use between sets of observation. "Ward" minimizes the variance of the clusters being merged.

The Silhouette coefficient for clusters is calculated using the equation X. The Silhouette coefficient is calculated for clusters ranging from 4 to 8 because the number of authors considered for the experimentation is 4.
For example when K-means for four to seven clusters is calculated, Silhouette score are
n= 4 is 0.307872420139, n= 5 is 0.314046546105, n= 6 is 0.304324826355,
n= 7 is 0.315737979835. Then average silhousse score is : 0.310495443108. The number of clusters is choosen to be 5 because its silhousse score is closer to the average score.

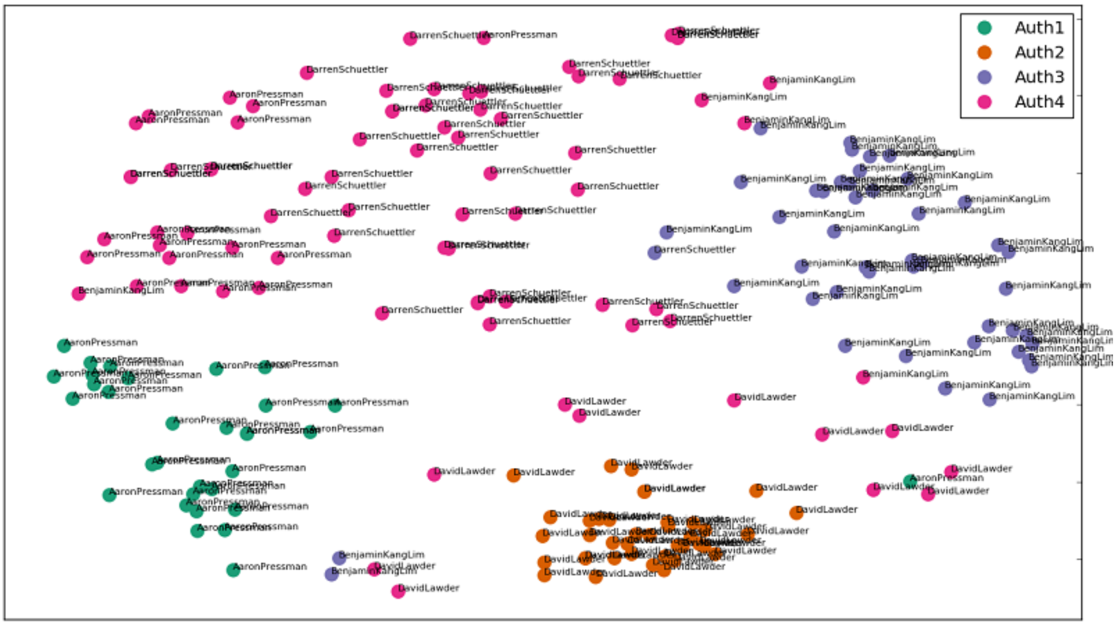
DOI10.17148/IARJSET.2017.4808

The Tables 1,2 and 3 shows precision, recall and BCusbed F-score values for each clustering algorithm. The Fig.1 shows visualization of clustering of the authors.

Cluster Label	Precision	Recall	F-Score
Benjamin Kang Lim	1	0.6	0.7499
David Lawder	1	0.78	0.8764
Darren Schuettler	0.6024	1	0.7519
Aaron Pressman	1	0.64	0.7804
Benjamin Kang Lim	1	0.32	0.4848

Cluster Label	Precision	Recall	F-Score
Benjamin Kang Lim	0.9782609	0.9	0.9375
David Lawder	1	0.78	0.8764
Darren Schuettler	0.5903614	0.98	0.736842
Aaron Pressman	1	0.64	0.7804

Cluster Label	Precision	Recall	F-Score
Benjamin Kang Lim	0.9375	0.6	0.7317
David Lawder	1	0.78	0.8764
Darren Schuettler	0.5925	0.96	0.7328
Aaron Pressman	1	0.64	0.7804
Benjamin Kang Lim	1	0.32	0.4848



5. CONCLUSION & FUTURE WORK

Earlier researches used authorship attribution as a classifier problem. However, in most of the 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 current paper experimets authorship clustering using morpheme based Ngram on unsupervised clustering algorithms like K-means, Mini Batch K-means and Ward Hierarchial clusterings. The performance of the clustering algorithms are evaluated using silhouette coefficient and calculated BCubed F-score and found that K-means algorithm achieves better clustering performance on C50 news groups data set. The experiment shows Mini Batch K-means shows higher cluster accuracy compared with other algorithms. The future scope of the work need to be implement authorship linking.

REFERENCES

1. Bagnall, Douglas. "Authorship clustering using multi-headed recurrent neural networks." arXiv preprint arXiv: 1608.04485 (2016).
2. Kocher, Mirco. "UniNE at CLEF 2016: Author Clustering." CLEF, 2016.
3. Mansoorizadeh, Muharram, et al. "Multi feature space combination for authorship clustering." CLEF, 2016.
4. Vartapetiance, Anna, and Lee Gillam. "A Big Increase in Known Unknowns: from Author Verification to Author Clustering-Notebook for PAN at CLEF 2016." Working Notes of CLEF 2016-Conference and Labs of the Evaluation forum,\'vora, Portugal, 5-8 September, 2016.
5. Zmiycharov, Valentin, et al. "Experiments in Authorship-Link Ranking and Complete Author Clustering." CLEF, 2016.
6. Verga, Patrick, et al. "Efficient Unsupervised Authorship Clustering Using Impostor Similarity."
7. Stamatatos, Efstathios, et al. "Clustering by authorship within and across documents." Working Notes Papers of the CLEF (2016).
8. Sittar, Abdul, Hafiz Rizwan Iqbal, and A. Nawab. "Author Diarization Using Cluster-Distance Approach." Working Notes Papers of the CLEF (2016).
9. Sari, Yunita, and Mark Stevenson. "Exploring Word Embeddings and Character N-Grams for Author Clustering." CLEF, 2016.
10. Khandelwal, Pooja, et al. "Document clustering for authorship analysis." International advanced research journal in science Engineering and technology 2.10 (2015): 205.
11. Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for Information Science and Technology, Volume 60 Issue 3, Pages 538-556, March 2009.
12. Stamatatos, Efstathios, et al. "Overview of the Author Identification Task at PAN 2014." CLEF (Working Notes). 2014.
13. Layton, Robert, Paul Watters, and Richard Dazeley. "Automated unsupervised authorship analysis using evidence accumulation clustering." Natural Language Engineering 19.01 (2013): 95-120.
14. http://scikit-learn.org/stable/modules/clustering.html#k-means
15. http://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

DOI10.17148/IARJSET.2017.4808