Author Clustering using Hierarchical Clustering Analysis

Notebook for PAN at CLEF 2017

Helena Gómez-Adorno[1], Yuridiana Aleman[2], Darnes Vilariño[2],
Miguel A. Sanchez-Perez[1], David Pinto[2], and Grigori Sidorov[1]

[1]Instituto Politécnico Nacional (IPN),
Center for Computing Research (CIC), Mexico City, Mexico
[2]Benemérita Universidad Autónoma de Puebla (BUAP),
Faculty of Computer Science, Puebla, Mexico
helena.adorno@gmail.com, yuridiana.aleman@gmail.com,
dvilarinoayala@gmail.com
miguel.sanchez.nan@gmail.com, dpinto@cs.buap.mx,
sidorov@cic.ipn.mx

Abstract This paper presents our approach to the Author Clustering task at PAN 2017. We performed a hierarchical clustering analysis of different document features: typed and untyped character n-grams, and word n-grams. We experimented with two feature representation methods, log-entropy model, and tf-idf; while tuning minimum frequency threshold values to reduce the dimensionality. Our system was ranked 1st in both subtasks, author clustering and authorship-link ranking.

1 Introduction

Authorship attribution consists in identifying the author of a given document in a collection. There are several subtasks within the authorship attribution field such as author verification [18], author clustering [14], and plagiarism detection [15]. This paper describes our approach to the author clustering task at PAN 2017 [19,13]. Formally, the task is defined as follows: given a document collection, the task is to group documents written by the same author so that each cluster corresponds to a different author. This task can be also viewed as establishing authorship links between documents. Applications of this problem include automatic text processing in repositories (Web), retrieval of documents written by the same author, among others.

The number of distinct authors whose documents are included in the collection is not given. The corpus contains documents in three languages (English, Dutch, and Greek) and two genres (newspaper articles and reviews). Two application scenarios were analyzed:

1. Complete author clustering: We do a detailed analysis, where we need to identify the number k of different authors (clusters) in a collection and assign each document to exactly one of the k clusters.

2. Authorship-link ranking: In this scenario we explore the collection of documents as a retrieval task. We aim to establish "authorship" links between documents and provide a list of document pairs ranked by a confidence score.

We approached the first scenario using clustering techniques and extracting character n-grams and stylometric features in a bag of words representation for each document. The selected features are language- and genre-independent. For the second scenario we calculated the pairwise similarity between each pair of documents in each problem using the cosine similarity metric.

The structure of this paper is as follows: In Section 2, we give an overview of the literature in this research field. In Section 3, we describe our methodology for the Author Clustering. In Section 4, we present the results obtained in the two phases of the evaluation.

2 Related work

Author clustering began in PAN 2012 as part of the Author attribution task focusing on the paragraph-level instead of document-level. In PAN 2016, the task was extended by the addition of the authorship link ranking problem [14].

Bagnall [2] used a multi-headed recurrent neural network to train a character n-gram model with a softmax output for each text in all problems. Later, he applied a method to turn multiple softmax outputs into clustering decisions. As preprocessing, he removed special tokens and decomposed capital letters into an uppercase marker followed by the corresponding lowercase letter. Afterward, he deleted the low document frequency words (words that appear only in a document). He built a model for each language using all documents available in all problems along with randomly sampled texts from previous corpora (2014, 2015, 2016). The goal of the training phase is optimizing the F-Bcubed score. In this regard, the author applied four different strategies. First, by prioritizing the case where each document belongs to one cluster, where the F-Bcubed score is guaranteed to be larger than 0.5. The other strategies are based on constraining a single-linkage approach to avoid merging large clusters, a heuristic aiming to find anchor points in the F-Bcubed score landscape, and a cluster-aware approach with a programming error that punished any link that joined more than two documents. Bagnall's approach ranked first place with an F-score of 0.8223.

Kocher's system [4] was ranked second. The author proposed an unsupervised approach using simple features and a distance measure called SPATIUM-L1. The features extracted when computing the distance between a pair of documents correspond to the top m most frequent terms in the first document of the pair, hence the distance is asymmetric $\varDelta_{A,B} \neq \varDelta_{B,A}$. He considered two documents to be linked when the distance for that particular pair and the distance from the first document to the rest of the collection is larger than the average minus twice the standard deviation. To compute the links between documents he used single-linkage clustering. This approach obtained an F-score of 0.8218.

Sari & Stevenson [17] extracted two different features: word embeddings and character n-grams. Then, they applied clustering based on K-Means. The hyperparameter k was optimized using the Silhouette Coefficient for each of the samples, and the words embeddings were trained using Gensim word2vec implementation. The authors used the 5,000 most frequent character n-grams, which included n ranging from 3 to 8. Their system ranked third with an F-score of 0.7952.

Zmiycharov et.al. [20] performed a combination of classification and agglomerative clustering. The authors used a wide set of features such as average sentence length, function words ratio, type-token ratio, and part of speech tags. In the classification phase, they trained six different classifiers using an iterative SVM algorithm: one for each language/genre pair. This approach exceeded the baseline competition, but with lower results than the rest of the participants.

The different systems presented in the Author Clustering task at PAN 2016 combined classification with clustering techniques, where the main differences are in preprocessing, feature extraction, and classification method.

3 Methodology

3.1 Complete author clustering

For the Author Clustering task at PAN 2017, we applied a Hierarchical Cluster Analysis (HCA) using an agglomerative [5] (bottom-up) approach. In this approach, each text starts in its own cluster and in each iteration we merged pairs of clusters.

To join clusters, we used an average linkage algorithm, where the average cosine distance between all the documents in the two considered clusters was used to decide if they were going to be merge. We used the Caliński Harabaz score [3] to evaluate the clustering model, where a higher Caliński-Harabaz score relates to a model with better defined clusters. So, in order to determine the number of clusters in each problem we performed the clustering process using a range of k values (with k varying from 1 to the number of samples in each problem) and chose the value of k with the highest Caliński Harabaz score. For k clusters, the Caliński Harabaz score is given as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

$$hc(k) = \frac{SS_B}{SS_W} \times \frac{N-k}{k-1}$$

where k is the number of clusters and N is the number of observations, SS_W is the overall within-cluster variance (equivalent to the total within sum of squares), and SS_B is the overall between-cluster variance. The total within sum of squares (SS_W) is calculated as follows:

$$SS_W = \sum_{i}^{k} \sum_{x \epsilon C_i} ||x - m_i||^2$$

where k denotes the number of clusters, x is the data point (document sample), C_i is the i$^{\text{th}}$ cluster, m_i is the centroid of the cluster i, and $||x - m_i||$ is the $L2$ norm (Euclidean distance) between the two vectors. The overall between-cluster variance is calculated using the total sum of squares (TSS) minus SS_W. The TSS is the squared distance of all the data points from the dataset's centroid; this measure is independent of the number of clusters.

SS_B measures the variance of all the cluster centroids from the dataset's grand centroid (when the centroids of each cluster are spread out and they are not too close to each other, the value of SS_B is larger). SS_W will keep on decreasing as the cluster size goes up. Therefore, for the Caliński-Harabasz score, the greatest ratio of $\frac{SS_B}{SS_W}$ indicates the optimal clustering size. In summary, this score is higher when clusters are dense and well separated, which means that different authors are probably well grouped in separate clusters.

Previous work on Authorship Attribution found that character n-grams are highly effective features, regardless of the language the texts are written in [9,11]. In our approach, we used a combination of typed character 3-grams, untyped character n-grams (with n varying between 2 and 8), and word n-grams (with n varying from 1 to 3). Typed character n-grams are character n-grams classified into ten categories based on affixes, words, and punctuation, and were introduced by Sapkota *et al.* [16].

The performance of each of the feature sets was evaluated separately and in combinations. The N most frequent terms in the vocabulary of each problem were selected based on a grid search and optimized based on the F-Bcubed score on the entire training set. We evaluated the N terms from 1 to 60,000 with a step of 50. We found that when selecting the most frequent 20,000 features we achieved the highest F-Bcubed score on the entire training set. Hence, we fixed this threshold for all the languages but selected the features separately for each problem.

Finally, we examined two feature representations based on a global weighting scheme: log-entropy and tf-idf on different clustering algorithms (k-means and hierarchical clustering). Global weighting functions measure the importance of a word across the entire collection of documents. Previous research on document similarity judgments [6,10] has shown that entropy-based global weighting is generally better than the tf-idf model. The log-entropy (le) weight is calculated as follows:

$$e_i = 1 + \sum_j \frac{p_{ij} \times \log p_{ij}}{\log n} \ \ where \ \ p_{ij} = \frac{tf_{ij}}{gf_i}$$

$$le_{ij} = e_i \times \log(tf_{ij} + 1)$$

where n is the number of documents, tf_{ij} is the frequency of the term i in document j, and gf_i is the frequency of term i in the hole collection. A term that appears once in every document, will have a weight of zero. A term that appears once in one document will have a weight of one. Any other combination of frequencies will assign a given term a weight between zero and one.

For the early bird submission, we used the k-means algorithm with tf-idf weighting scheme and the Silhouette Coefficient for choosing the number of clusters. In the final submission, we used a hierarchical clustering with log-entropy weighting scheme and the Caliński Harabaz score for choosing the number of clusters.

3.2 Authorship-link ranking

In order to establish the authorship links, we simply calculated the pairwise similarity between each pair of documents in each problem using the cosine similarity metric. The vector space model was built in the same manner as for the complete author clustering subtask, i.e., the same features and the same weighting scheme (log-entropy).

4 Results and Evaluation Measures

Two measures were used in order to estimate the performance of the submitted systems to the PAN CLEF 2017 campaign. The F-Bcubed score [1] was used to evaluate the clustering output. This measure corresponds to the harmonic mean between precision and recall. The Bcubed precision (P-Bcubed) represents the ratio of documents written by the same author in the same cluster. While the Bcubed recall (R-Bcubed) represents the ratio of documents written by an author that appear in its cluster. The Mean Average Precision (MAP) [7] is used to evaluate the authorship-link ranking. The MAP measures the average area under the precision-recall curve for a set of problems.

Table 1 presents the results of our early bird submission obtained on the PAN Author Clustering 2017 test dataset evaluated on the TIRA platform [12]. In this submission, we had a problem with our authorship-link ranking module, for this reason the MAP evaluation measure is not available.

Table 1. Early bird submission results in the Author Clustering subtask.

Language	F-Bcubed	R-Bcubed	P-Bcubed
English	0.5868	0.6858	0.5914
Greek	0.5372	0.6306	0.5461
Dutch	0.5372	0.6306	0.5461
Average	0.5483	0.6630	0.5479

Table 2 presents the results of our final submission obtained on the PAN Author Clustering 2017 test dataset. Our final system increased the performance of our early bird submission by 2.5% in terms of the mean F-Bcubed score. We also observed a similar improvement on the training set, were the final configuration of the system achieved 3% more than our baseline system in terms of the mean F-Bcubed score. Our system was ranked 1st in both subtasks, author clustering (evaluated with the mean F-Bcubed score) and authorship-link ranking (evaluated with the MAP score).

Table 2. Results on the Author Clustering 2017 test dataset.

Language	F-Bcubed	R-Bcubed	P-Bcubed	MAP
English	0.5913	0.6175	0.6483	0.5211
Greek	0.5517	0.5743	0.6222	0.4220
Dutch	0,5765	0.7204	0.5508	0.4224
Average	0.5733	0.6379	0.6069	0.4554

5 Conclusions

We presented our system submitted to the Author Clustering task at PAN 2017. We carried out experiments using different features: typed and untyped character n-grams, and word n-grams. Our final submission implemented log-entropy weighting scheme on the combination of the 20,000 most frequent terms with hierarchical clustering. We optimized the number of clusters in each problem using the Caliński Harabaz score.

In future research, we would like to adapt the feature set for each language (subcorpus), as described in [8], in order to improve system performance for each of the languages individually.

Acknowledgments

This work was partially supported by the Mexican Government (CONACYT projects 240844, SNI, COFAA-IPN, SIP-IPN 20171813, 20171344, and 20172008).

References

1. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval 12(4), 461–486 (2009)
2. Bagnall, D.: Authorship Clustering Using Multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal (2016)
3. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics 3(1), 1–27 (1974)
4. Kocher, M.: UniNE at CLEF 2016: Author Clustering—Notebook for PAN at CLEF 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal (2016)
5. Layton, R., Watters, P., Dazeley, R.: Automated unsupervised authorship analysis using evidence accumulation clustering. Natural Language Engineering 19, 95–101 (2013)
6. Lee, M.D., Navarro, D.J., Nikkerud, H.: An empirical evaluation of models of text document similarity. In: Proceedings of the Cognitive Science Society. vol. 27 (2005)
7. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
8. Markov, I., Gómez-Adorno, H., Sidorov, G.: Language- and subtask-dependent feature selection and classifier parameter tuning for author profiling. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017)
9. Markov, I., Stamatatos, E., Sidorov, G.: Improving cross-topic authorship attribution: The role of pre-processing. In: Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing 2017, Springer (2017)
10. Pincombe, B.: Comparison of human and latent semantic analysis (lsa) judgements of pairwise document similarities for a news corpus. Tech. rep., DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION SALISBURY (AUSTRALIA) INFO SCIENCES LAB (2004)

11. Posadas-Durán, J., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., Chanona-Hernández, L.: Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing 21, 627–639 (2016)
12. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative. pp. 268–299. CLEF'14 (2014)
13. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative. CLEF'17, Berlin Heidelberg New York (2017)
14. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In: Fuhr, N., Quaresma, P., Larsen, B., Gonçalves, T., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16). Berlin Heidelberg New York (2016)
15. Sanchez-Perez, M.A., Gelbukh, A., Sidorov, G.: Adaptive algorithm for plagiarism detection: The best-performing approach at PAN 2014 text alignment competition. In: Proceedings of the 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8–11, 2015. vol. 9283, pp. 402–413. Springer (2015)
16. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies. pp. 93–102. NAACL-HLT '15, Association for Computational Linguistics (2015)
17. Sari, Y., Stevenson, M.: Exploring Word Embeddings and Character N-Grams for Author Clustering—Notebook for PAN at CLEF 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal (2016)
18. Stamatatos, E., amd Ben Verhoeven, W.D., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the Author Identification Task at PAN 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers (2015)
19. Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN 2017: Style Breach Detection and Author Clustering. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings (2017)
20. Zmiycharov, V., Alexandrov, D., Georgiev, H., Kiprov, Y., Georgiev, G., Koychev, I., Nakov, P.: Experiments in Authorship-Link Ranking and Complete Author Clustering—Notebook for PAN at CLEF 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal (2016)