Discovering Author Groups using a β-compact graph-based clustering.

Notebook for PAN at CLEF 2017

Yasmany García-Mondeja[1], Daniel Castro-Castro[1], Vania Lavielle-Castro[1], Rafael Muñoz[2]

[1]Desarrollo de Aplicaciones, Tecnología y Sistemas DATYS, Cuba

{yasmany, daniel.castro}@cerpamid.co.cu,
vania.lavielle@datys.cu

[2]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, España

rafael@dlsi.ua.es

Abstract. Identifying the authorship either of an anonymous or a doubtful document constitutes a cornerstone for automatic forensic applications. Moreover, it is a challenging task for both humans and computers. Clustering documents according to the linguistic style of the authors who wrote them has been a task little studied by the research community. In order to address this problem, PAN Evaluation Framework has become the first effort to promote the development of the author clustering. This article proposes a graph-based method, specifically β-compact clustering, for discovering the groups of documents written by the same author. The β-compact algorithm is based on the analysis of the similarity between documents and they belong to the same group as long as the similarity between them exceeds the threshold β and it is the maximum similarity with respect to other documents. In our proposal we evaluated different linguistic features and similarity measures presented in previous works of authorship analysis task. The training dataset was used to determine the best value of β parameter for each language. The result of the experiments was encouraging.

Keywords: Author clustering, β-compact clustering algorithm, linguistic features, similarity measures

1 Introduction

The documents clustering task, by author´s linguistic style, is of vital importance in forensic applications. A practical example would correspond to the identification in a computer of all the groups of documents written in this one and that each group of document has been written by a single author. Considering that this computer belongs to a public site.

In the evaluation framework the task is described as follows: "Given a collection of (up to 50) short documents (paragraphs extracted from larger documents), identify authorship links and groups of documents written by the same author. All documents are single-authored, in the same language, and belong to the same genre. However, the topic or text-length of documents may vary. The number of distinct authors whose documents are included in the collection is not given. "[1]

One of the most used strategies for documents representation in Text Mining (TM) applications, corresponds to the classic Bag of Words [4][9] and this will be the proposal used in our work. In different Authorship Analysis applications, complex methods involving several algorithms have been used in order to obtain the best results. In document clustering applications and other Artificial Intelligence (AI) tasks, ensembles of algorithms have also been employed. Despite this, the work presented by [7] is relevant, and they use a simple clustering algorithm and achieve encouraging results.

As a summary, in the last edition of authors clustering task, 6 papers were presented [1][3][7][8][12][13] and in general, the data of the documents collection set contained a high percentage of clusters composed of a single document, unlike what can be seen in the collection of this year released for training, where we observed several documents clusters with more than one document, although there are still few documents per group.

With our work, we want to propose and evaluate a clustering algorithm that we have used in topic document clustering tasks in our research center, and its purpose is to group objects with the condition that for each object of the group, at least there is an object with which the similarity between them is greater than a threshold of similarity and it´s the maximum similarity with an object of the collection.

It is important to emphasize aspects of the description of the author clustering problem, such as: short texts no longer than a paragraph; the texts corresponding to the same author are of the same genre but not necessarily the same topic or homogeneous length. In addition, as part of the task, we need to obtain a ranking of similarities between objects in the same clusters. Taking these details into consideration, in the next section we propose and describe our method considering binary linguistic features and a clustering algorithm based on compact clusters.

2 Implemented method

We propose the use of β-compact algorithm [10] for authorship clustering task, because it´s based on clustering objects with a similarity between them which is greater than a threshold of similarity previously adjusted with a training document collection, but only the greatest similarities are maintained.

In the following image (Figure 1) we expose the architecture of the implemented method and later we describe each one of the steps involved.

---

1 http://www.webis.de.

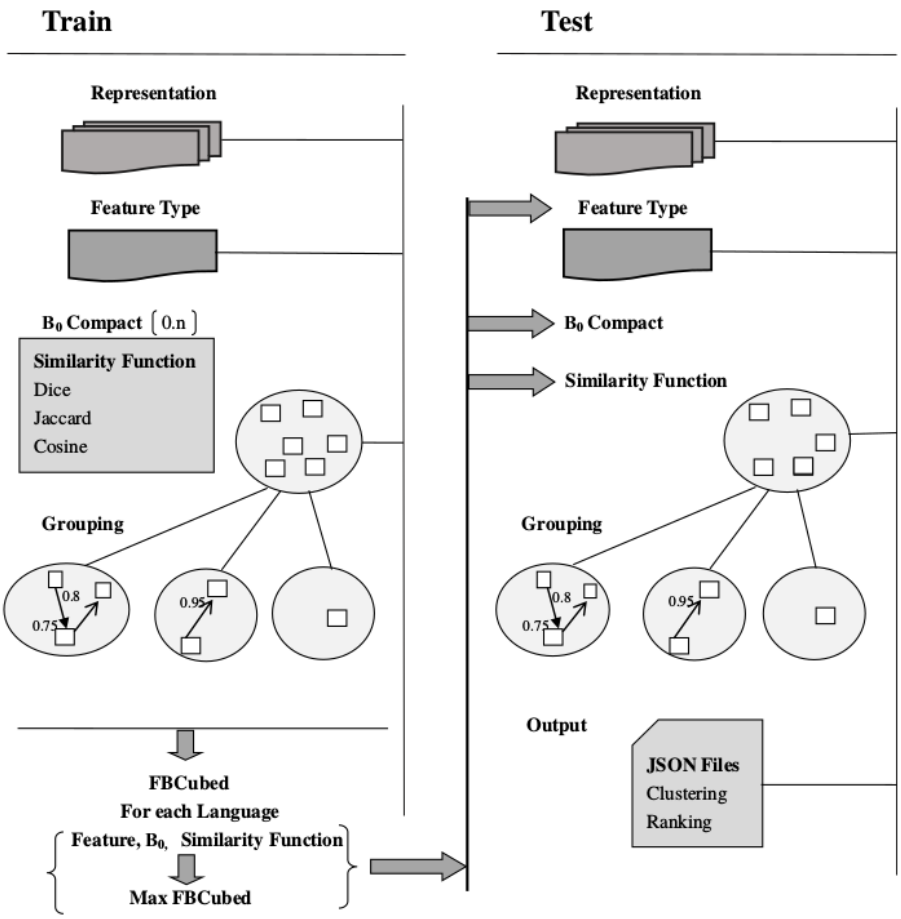

Fig. 1. Scheme of the proposed clustering task using β-compact algorithm. Training and test phase.

Both in the training and test stage, collections of documents are received and the final purpose is to obtain groups of documents, where all the documents of a group belong to the same author.

The algorithm proposed, to obtain the groups, requires the representation of each of the documents, a comparison function that allows to evaluate the similarity between a pair of documents and a threshold β to decide when two documents must belong to the same cluster.

For documents representation, we used the classic Bag of Word, and with the training dataset we tried different types of features [2]. We experimented with 3 similarity functions to analyze the similarity between documents. We used the Dice, Jaccard and Cosine functions [6], using only binary features, that is, we did not compute the frequency of each of the features, only their appearance in the document. The idea of

considering only binary features is due to the short extension of the documents, up to one paragraph.

The β-compact clustering algorithm is described in the next pseudo program code. First we need to define the concept "Graph of Maximum β similarity: It´s an oriented graph in which the vertices are the objects and exist an arista between two vertices O_i and O_j if O_j is β-similar with O_i and O_j is the most similar of all the rest of objects [10].

```
In: U - universe of documents
Out: Cluster – Several groups of documents

Cluster = Ø
G = BuildGraphMaximum_β_Similarity(U)
Cluster = SearchConexedComponentsIgnoringOrientation(G)
```

We performed different runs, in which the comparison function and the representation were varied, as well as the value of β from 0 to 0.5. When the β was greater than 0.5, there were no changes in the clusters obtained. For each language, the best result is determined by analyzing the clusters obtained by comparing them with the training data using the FBcubed [5] measure.

Finally, for each language, we have a configuration of the method, with a binary features representation to be calculated, a similarity function and a threshold β.

It is important to note that, due to the nature of the β-compact algorithm, two documents can belong to the same group, although the similarity between them not necessarily exceeds the defined β, because the only condition is that each one of them has a similarity greater than β with some document of their group. This characteristic may lead to non-necessarily spherical clusters. It can be observed in the outputs of the method in the ranking file, where similarities between documents of the same cluster will appear that are below the threshold β and therefore will appear little ranked in this list.

For the similarity ranking construction, we took into account all similarities among objects of a group and order them. The similarities that are above the threshold β were distributed in a scale of 0.5 to 1, corresponding to 0.5 the similarities that are equal to β and close to 1 the greater ones. Similarly, was realized a distribution of the similarities that did not exceed the β, which in this case were distributed from 0 to 0.5.

3 Experimental results

With the documents of the training dataset released and explained in [11], we performed different runs of the method, evaluating in all cases the results with the groups proposed by specialists, using the FBCubed measure suggested in the competition. Table 1 shows a summary of the parameter configurations and features representation used in the training phase. In Table 2, we present the final configurations used by the method for the final evaluation.

Table 1. Parameters configuration of the method.

Parameters				
Features	Character	words	lemma	POS-Tagging
N-gram	1-5	1-5	1-5	1-5
Language	en, gr, du	en, gr, du	en	en
β	0.05-0.5	0.05-0.5	0.05-0.5	0.05-0.5

The features used in the documents representation are those presented in the Features row, we test using n-grams character representations (n = 1, ..., 5), for English (en), Greek (gr) and Dutch (du). Representations of n-grams words (n = 1, ..., 5) for the three languages mentioned. For the representations of Lemma and Part of Speech Tagging (POS-Tagging) we only processed the English texts. For each of the configurations, evaluations were performed varying the β from 0.05 to 0.5.

Table 2. Parameters configurations for the final evaluation.

English				
Article	Lemma	0.235	N = 1	Dice
Review	Character	0.21	N = 3	Dice
Greek				
Article	Character	0.175	N = 3	Dice
Review	Character	0.455	N = 2	Dice
Dutch				
Article	Character	0.23	N = 3	Dice
Review	words	0.27	N = 1	Dice

Table 2 shows the configurations of the parameters including the textual genre of the documents offered in the collection.

The general results achieved in the competition [11] are shown table 3.

Table 3. Results of PAN 2017 Clustering Task

User	**Mean Average Precision**	**Mean F-score**	**Runtime**
aleman17	0.455154	0.573159	00:02:05
kocher17	0.395054	0.551680	00:00:41
castrocastro17a	**0.379999**	**0.564703**	**00:15:49**
halvani17	0.139425	0.548751	00:12:25
spiewak17	0.125229	0.466319	00:00:26
alberts17	0.041608	0.527642	00:01:45

These are general results and they are an average between al the languages processed. To analyze the results in each language please consult the PAN overview [11].

4 Conclusions and future work

As conclusions, we must emphasize that one of the essential aspects in our work is the features identification for the documents representation, and in this we could try other ideas presented in the literature. The algorithm usually obtains small groups, and we

could evaluate in a future work the differences that can be reached when we use other variants such as the β-connected and β-strongly compact algorithms. The β-connected would obtain larger groups, while the strongly compact would have smaller and more compact groups than those proposed in our work.

We propose to evaluate different features weighing strategies and other comparison functions proposed in the literature. Using the results achieved in this work, take into consideration comparisons with ensemble clustering algorithms.

5 References

1. Anna Vartapetiance, Lee Gillam: A Big Increase in Known Unknowns: from Author Verification to Author Clustering. CLEF (Working Notes) 2016: 1008-1013
2. Daniel Castro, Yaritza Adame, María Peláez Brioso, Rafael Muñoz: Authorship Verification, combining Linguistic Features and Different Similarity Functions. CLEF (Working Notes) 2015
3. Douglas Bagnall: Authorship Clustering using Multi-headed Recurrent Neural Networks. CLEF (Working Notes) 2016: 791-804
4. 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
5. Efstathios Stamatatos, Michael Tschuggnall, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, Martin Potthast: Clustering by Authorship Within and Across Documents. CLEF (Working Notes) 2016: 691-715
6. Gomaa, W. and A. Fahmy. A Survey of Text Similarity Approaches. International Journal of Computer Applications (0975 – 8887) Volume 68– No.13. 2013
7. Mirco Kocher: UniNE at CLEF 2016: Author Clustering. CLEF (Working Notes) 2016: 895-902
8. Muharram Mansoorizadeh, Mohammad Aminian, Taher Rahgooy, Mehdy Eskandari: Multi Feature Space Combination for Authorship Clustering. CLEF (Working Notes) 2016: 932-938
9. Patrick Juola. Authorship Attribution. In Foundations and Trends in Information Retrieval, Volume 1, Issue 3, March 2008
10. Reynaldo Gil-García, José Manuel Badía-Contelles, Aurora Pons-Porrata: A Parallel Algorithm for Incremental Compact Clustering. Euro-Par 2003: 310-317
11. 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: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
12. Valentin Zmiycharov, Dimitar Alexandrov, Hristo Georgiev, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, Preslav Nakov: Experiments in Authorship-Link Ranking and Complete Author Clustering. CLEF (Working Notes) 2016: 1018-1023
13. Yunita Sari, Mark Stevenson: Exploring Word Embeddings and Character N-Grams for Author Clustering. CLEF (Working Notes) 2016: 984-991