# A Big Increase in Known Unknowns: from Author Verification to Author Clustering

## Notebook for PAN at CLEF 2016

Anna Vartapetiance and Lee Gillam

Department of Computing, University of Surrey, UK
{a.vartapetiance, l.gillam}@surrey.ac.uk

**Abstract.** Previous PAN workshops have afforded evaluation of our approaches to author verification/identification based on stopword co-occurrence patterns. Problems have tended to involve comparing one document to a small set of documents (n<=5) of known authorship. This paper discusses the adaptation of one of our approaches to a PAN 2016 problem of author clustering, which involves 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. We describe our previous approaches as the background to the approach taken to this task and briefly overview the results that were achieved, which are not expected to be particularly remarkable due to substantial limitations on our time around the task.

## 1 Introduction

In previous years of the International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN – for example, [1]), we have tested our ideas on co-occurrence patterns of stopwords [2], explored 3 variations of such an approach [3], and created a means to adapt for topic-specific term length [4]. These previous years of PAN were geared towards a classification task, deciding the degree to which a document belongs to a class comprised of other documents. In PAN2016 this has changed from a classification task to a categorization task, with an unknown number of categories less than or equal to the number of documents [5]. Where previous tasks involved small numbers of documents in the existing class (n<=5), this task involves generating clusters from larger sets of documents (n<=100), where each set of documents is in English, Dutch or Greek. This author clustering task could be considered as offering a more rigorous version of the classification task, as a kind of attribution given noise, which could also help to bring a more objective evaluation of authorship, in particular, by reducing the so-called "halo effect" of expert testimony.

In this paper, we discuss the simple adaptations made to our existing approach in order to address this task. Due to significant limitations on our time, we were unable to evaluate our approach with any real form of rigour beyond a limited brute force determination of category formation thresholds, and so results obtained reflect more a desire to continue our involvement in PAN and are not expected to be remarkable.

In section 2, we briefly discuss the previous approaches we have used for author verification. Section 3 explains the modifications made to address PAN 2016. Section 4 offers results and evaluation, and Section 5 concludes the paper.

## 2 Previous methods applied

For PAN2012, we used a mean-variance framework for author 'attribution', analysing co-occurrence a small set (up to 10 of the most frequent) of stopwords within a specified maximum word window [2], and extended this approach to Greek and Spanish texts PAN2013 simply by using language-specific stopword lists. PAN2014 required another stopword list, this time for Dutch (see Table 1, below, for full set of stopwords used across these tasks), and allowed us to explore two further approaches based on (i) an aggregate cosine comparison of positional frequencies and (ii) a single cosine comparison [3]. In PAN2015, we created a means to adapt for topic-specific term length [4] and account for positional variation due to this. These approaches provided fertile ground for a PAN2016 effort.

**Table 1:** List of stopwords for all four languages

| *Language* | *Stopwords* |
|---|---|
| Dutch | De Van Een Het En In Is Dat Op Te |
| English | The Be To Of And A In That Have I |
| Greek | Και Το Να Τον Η Της Με Που Την Από |
| Spanish | De La Que El En Y A Los Del Se |

## 3 PAN 2016

For this year's task, the shift from classification to categorization curtailed deeper explorations into the effectiveness of the previous approaches and focused activity entirely on adaptation to the new task. Many possible categorization approaches exist, but where text is involved these tend to focus towards topics and involve feature selection approaches geared towards discrimination of topic-specific keywords, with similarity evaluation based on these features, for example with similarity measures over vector space models. Since, according to the task description, the text topic may vary, topic-specific approaches were ruled out. For this reason, we quickly fall back to our previously evaluated methods; also because our previous approaches involve determining similarity over, essentially, vector spaces.

Our approach operates, in general, as follows: we consider a maximum window distance, $w$, as a distance between any two stopwords in a stoplist of length $l$. For each document, we produce a matrix ($w$ by $l^2$) representing the separation distances between pairs of stopwords. The variants of our approach relate to different ways in which then to treat the data in this matrix, and a number of further alternatives that we are yet to explore exist here also. In terms of matching, this approach carries statistical advantages – since stopwords are very hard for an author to avoid, in contrast to keywords, sparsity of such a matrix should be less of an issue – whilst

disproportionality may be indicative of individual preferences and factor out structural (grammatical) restrictions – for example, 'of the' but typically not 'the of', similar to the 'bread and butter' but typically not 'butter and bread' of [6]. Our adaptation for topic-specific term length attempts to address differences in separation distances in respect to a propensity for, for example, long compound nominals in certain topics compared to others ('chiral single-walled carbon nanotubes'; 'twin-engine tandem rotor heavy-lift helicopter'). We address this through the notion of a 'topic cost', which we determine by counting the number of terms between the stopwords of interest and the length of these terms, and using the difference between these two values to re-distribute a given position score. This requires, for each language, an additional resource - also a stoplist, albeit rather more comprehensive than those identified above - to be able to expose the terms. For PAN2015, we could then investigate similarities between one unknown document and any number (although n<=5) of known documents, and also between all known documents, to establish expectations on similarity. A document could be considered as being by the same author if the highest similarity values obtained in comparing the known document to the unknown documents – with comparison made pattern-wise based on cosine similarity – average higher than a certain threshold; 0.30, 0.40, 0.60 and 0.80 respectively for Dutch, English, Greek and Spanish languages.

By contrast, PAN2016 necessitates this comparison of all possible pairs of documents (optimizations may subsequently be identified) in order to create the unknown number of clusters representing the unknown number of authors per problem for the known number of documents. Our similarity scores between document pairs can be used for the ranking part of the task, with a threshold used to be selective over those which appear in ranking. Pairs which remain in the rankings are all above threshold and reported as clusters, with a minimum of 2 members, created by collecting and flattening ranked pairs with documents in common (e.g. [a,b], [b,c], [b,d], [a,e] → [a, b, c, d, e]); those that are not ranked are reported as singletons (clusters with a single member). Following some bug-fixing of the clustering, and a small number of trials with the training dataset, we used 0.5 as the threshold for all three languages. Due to lack of time we were unable to evaluate in any significant ways the results that could be achieved by refining this threshold, applying our prior – or new – approaches, or evaluating feature set *reduction* as tried successfully in previous years, and these would all offer potential for future experimental work.

## 4 Results

Results for each of the training and test collection of PAN 2016 are shown in the tables below based on the evaluation metrics being used. We present these only for the purpose of documenting the results, and reserve interpretation due to the absence of knowledge of performance of other systems as would assist us in contextualization.

**Table 2:** Results from the Evaluator based on the Training Corpus and thresholds for clusters of en=0.5, nl=0.5, gr=0.5

| *Problem (#docs)* | *Lang.* | *Genre* | *F-Bcubed* | *R-Bcubed* | *P-Bcubed* | *Av-Precision* |
|---|---|---|---|---|---|---|
| problem001 (50) | en | articles | 0.078401 | 1 | 0.0408 | 0.023508 |
| problem002 (50) | en | articles | 0.14815 | 1 | 0.08 | 0.04074 |
| problem003 (50) | en | articles | 0.088924 | 1 | 0.046531 | 0.011735 |
| problem004 (80) | en | reviews | 0.046398 | 1 | 0.02375 | 0.005601 |
| problem005 (80) | en | reviews | 0.03198 | 1 | 0.01625 | 0.003615 |
| problem006 (80) | en | reviews | 0.06354 | 1 | 0.032813 | 0.016949 |
| problem007 (57) | nl | articles | 0.82304 | 0.94737 | 0.72755 | 0 |
| problem008 (57) | nl | articles | 0.66797 | 0.81053 | 0.56806 | 0.040129 |
| problem009 (57) | nl | articles | 0.73889 | 0.78363 | 0.69898 | 0.011173 |
| problem010 (100) | nl | reviews | 0.28786 | 0.86 | 0.17286 | 0.02488 |
| problem011 (100) | nl | reviews | 0.33262 | 0.85167 | 0.20667 | 0 |
| problem012 (100) | nl | reviews | 0.37251 | 0.95667 | 0.23128 | 0.007143 |
| problem013 (55) | gr | articles | 0.083016 | 1 | 0.043306 | 0.01234 |
| problem014 (55) | gr | articles | 0.067093 | 1 | 0.034711 | 0.047392 |
| problem015 (55) | gr | articles | 0.045866 | 1 | 0.023471 | 0.004599 |
| problem016 (55) | gr | reviews | 0.043338 | 1 | 0.022149 | 0.011038 |
| problem017 (55) | gr | reviews | 0.10345 | 1 | 0.054545 | 0.028131 |
| problem018 (55) | gr | reviews | 0.059654 | 1 | 0.030744 | 0.030675 |

**Table 3:** Results from the Evaluator based on the Test Corpus and thresholds for clusters of en=0.5, nl=0.5, gr=0.5

| *Problem* | *Lang.* | *Genre* | *F-Bcubed* | *R-Bcubed* | *P-Bcubed* | *Av-Precision* |
|---|---|---|---|---|---|---|
| problem001 | en | articles | 0.054011 | 1 | 0.027755 | 0.002294 |
| problem002 | en | articles | 0.11393 | 1 | 0.060408 | 0.014387 |
| problem003 | en | articles | 0.033708 | 1 | 0.017143 | 0.010335 |
| problem004 | en | reviews | 0.042813 | 1 | 0.021875 | 0.014908 |
| problem005 | en | reviews | 0.030769 | 1 | 0.015625 | 0.002269 |
| problem006 | en | reviews | 0.065296 | 1 | 0.03375 | 0.038915 |
| problem007 | nl | articles | 0.76808 | 0.78947 | 0.74781 | 0.002083 |
| problem008 | nl | articles | 0.78436 | 0.91228 | 0.6879 | 0.0125 |
| problem009 | nl | articles | 0.65528 | 0.54887 | 0.81287 | 0 |
| problem010 | nl | reviews | 0.4605 | 0.96 | 0.3029 | 0.007047 |
| problem011 | nl | reviews | 0.41359 | 0.78667 | 0.28054 | 0.00927 |
| problem012 | nl | reviews | 0.48847 | 0.83667 | 0.34492 | 0.006292 |
| problem013 | gr | articles | 0.047031 | 1 | 0.024082 | 0.009168 |
| problem014 | gr | articles | 0.068585 | 1 | 0.03551 | 0.024812 |
| problem015 | gr | articles | 0.035285 | 1 | 0.017959 | 0.012965 |
| problem016 | gr | reviews | 0.047031 | 1 | 0.024082 | 0.014572 |
| problem017 | gr | reviews | 0.033708 | 1 | 0.017143 | 0.016435 |
| problem018 | gr | reviews | 0.062475 | 1 | 0.032245 | 0.017248 |

## 5 Conclusions and Future Work

In this paper, we discussed the adaptation of one of our approaches to a PAN 2016 problem of author clustering, and the contrast of this task to earlier tasks as might be conceived as author classification. Because of our participation in previous tasks, and approaches taken there, the changes we needed to make – largely around ingesting data and similarity score processing – were relatively minimal. However, the timing of a number of other priority efforts brought substantial limitations to the effort we were able to dedicate to this task, in contrast to that which we would have liked to dedicate, and because of this we do not expect that the results obtained to be particularly remarkable.

## Acknowledgements

The authors gratefully acknowledge both the efforts and patience of the PAN organizers in crafting and managing the task and the prior support from EPSRC/JISC (EP/I034408/1), the UK's Technology Strategy Board, now InnovateUK (TSB, 169201), HEFCE/Innovate UK through SETsquared, and the UK government.

## References

1. E. Stamatatos, W. Daelemans, B. Verhoeven, P. Juola, A. López-López, M. Potthast, and B. Stein. "Overview of the Author Identification Task at PAN 2015". In: Cappellato, L., Ferro, N., Jones, G., and San Juan, E., editors. CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2015.
2. A. Vartapetiance and L. Gillam, "Quite Simple Approaches for Authorship Attribution , Intrinsic Plagiarism Detection and Sexual Predator Identification - Notebook for PAN at CLEF 2012," in *Working Notes Papers of the CLEF 2012 Evaluation Labs*, 2012.
3. A. Vartapetiance and L. Gillam, "A Trinity of Trials : Surrey's 2014 Attempts at Author Verification - Notebook for PAN at CLEF 2014," *Work. Notes Pap. CLEF 2014 Eval. Labs*, 2014.
4. A. Vartapetiance and L. Gillam, "Adapting for Subject-Specific Term Length using Topic Cost in Author Verification - Notebook for PAN at CLEF 2015". In: Cappellato, L., Ferro, N., Jones, G., and San Juan, E., editors (2015).
5. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Clustering by Authorship Within and Across Documents. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016)
6. Church, K., Hanks, P.: Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, vol. 16(1), pp. 22, 1991.