Homotopy Based Classification for Author Verification Task

Notebook for PAN at CLEF 2015

Josue Gutierrez[1], Jose Casillas[2], Paola Ledesma[3], Gibran Fuentes[1], and Ivan Meza[1]

[1]Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas (IIMAS)
[2]Facultad de Ciencias (FC)
Universidad Nacional Autonoma de Mexico (UNAM)
[3]Escuela Nacional de Antropologia e Historia (ENAH)
http://www.enah.edu.mx

Abstract This paper presents our experience implementing a homotopy-based classification (HBC) system for the 'PAN 2015 Author Identification' [20]. Known documents from a specific author and randomly selected impostor documents are used as a dictionary to generate a contested document. Given the contribution of the known documents to the contested document we can verify the authorship of the document. This classification is embedded into the General Impostor Method resulting in an ensemble of the SBC model.

1 Introduction

Author verification has multiple applications in several areas including information retrieval and computational linguistics, and has an impact in fields such as law and journalism [8,10,18]. In this edition of the *PAN 2015 Author Identification*, the task was formally defined as follows[1]:

> *Given a small set (no more than 5, possibly as few as one) of "known" documents by a single person and a "questioned" document, the task is to determine whether the questioned document was written by the same person who wrote the known document set. The genre and/or topic may differ significantly between the known and unknown documents.*

This edition had documents in English, Spanish, Dutch and Greek.

In this work we present our approach for author verification based on sparse-based classification. Homotopy-based Classification (HBC) was first proposed for face recognition in this setting the goal is to measure the contribution of known faces in the generation of an unknown face. The amount of contribution determines the identity of the person with the unknown face [21]. This work is a continuation from the previous version of our system [15]. In this version we have normalized the extraction of document representation; additionally, we have added character-level features.

[1] As described in the official website of the competition http://pan.webis.de/ (2015).

2 Previous work

Author verification is considered a corner stone of the authorship analysis together with authorship attribution, author profiling and plagiarism detection tasks [19]. Current work on the field depends on similarity metrics among texts such as: Jaccard, *cosine*, Euclidean and *min-max* similarities. As aforementioned the general impostor method has been successful at using similarity measures relative to documents in the domain [17,9]. On the other hand, clustering approaches highly depends on similarity measures [7].

Alternative methods for combining distances have been also proposed [2]. Even the Common N-gram (CNG) method which was originally proposed for author profiling had been adapted to author verification and it can be interpreted as a particular similarity metric [4,13]. Metric distances have been important on the field since they facilitate an unsupervised framework for the task. In order to surpass some of the limitations of similarity metrics supervised approaches had been explore [7,16]. Hybrid approaches on which model is built on a feature space based on similarity metrics had also been proposed with mixed results [5,14].

3 Document representation

We use the vector space model to represent the documents. In this edition we use the following features:

1. **Bag of words** Frequencies of words in the document.
2. **Bigram of words** Frequencies of two consecutive words.
3. **Punctuation** Frequencies of punctuations.
4. **Trigram of words** Frequencies up to three consecutive letters.

Table 3 presents the final configurations of feature per language

Table 1. Features used per language.

Feat	Dutch	English	Greek	Spanish
1	*	*	*	*
2	*	*	*	*
3	*		*	
4	*	*	*	

4 Methods

In order to present our proposal first we review the GI method, and the homotopy-based classification, to follow with our proposal.

4.1 The general impostor method

The GI method is a second order binary similarity metric for collections of documents [11,12]. It uses two functions: the similarity metric sim that compares pairwise documents, and the aggregate function agg to allow for comparing collections of documents. The aggregate function does not work directly with the similarity function, it rather aggregates the score calculated by the original impostor method which is also a pairwise metric. This method has been described as an ensemble of random models since several comparisons are performed with randomly selected impostors [9].

The procedure to generate the impostor collection is the following: First, randomly select n terms of a document and made a query to a search engine. Second, from the results keep the first m results. Third for each result only take the k first words. Finally, repeat this m times.

4.2 Homotopy-based Classification

At the core of the proposal is performing variable selection over the equation system represented in Figure 1 by optimizing the following objective $x' = argmin||x||_1$. $A \in \Re^{MxN}$ is a matrix composed of N columns of document examples for which we know their associated authorship: known author (K) or impostor (I). $x \in \Re^N$ is a vector that when linearly combined with A generates the $y \in \Re^M$ vector of the questioned document. In particular a desirable x will be sparse so that variables are zeroed and ignored in the reconstruction of y. To reach this type of solution we chose the $L1$-homotopy algorithm which was used to find a sparse solution on an undetermined system of equations [3,1].

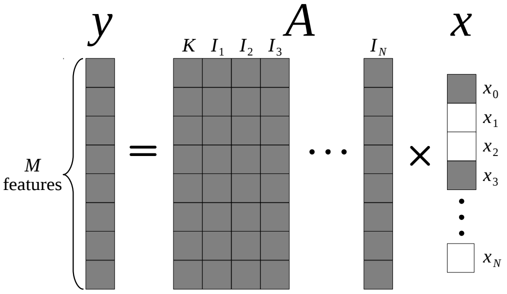

Figure 1. Equation system for the reconstruction of y using a matrix A of example documents.

Wright et. al propose to calculate the residuals for each i identity represented in the matrix A using the following formula:

$$r_i(y) = |y - Ax'_i| \tag{1}$$

A vector x'_i is created for which we zeroed the values of x' that do not correspond to the i identity. Thus the r_i represents the difference between the unknown document and the reconstructed document using only elements corresponding to the same identity. After calculating the residual per identity, we look for the lower residual for figure out the identity.

Following this procedure we are able to assign one of the identities to the questioned document: $True$ if the residual i corresponds to the author of documents D, if not $False$.

4.3 GI with Homotopy-based Classification

Algorithm 1 shows the adaptation of the GI method to be used on the homotopy based classification. The aggregate function agg iterates over a pairs of documents in the collections to compare. It aggregates similarities based on the Homotopy-based Classification procedure (HC) described above. At the end, the GI_{hc} is a voting system over randomly selected impostors.

Algorithm 1 The General Impostor within sparse approximation

```
procedure HC($D$,$y$,$I$)
    $I_r \leftarrow (random(I, \%, N * |D|))$
    $D_r \leftarrow (random(D, \%))$
    $A \leftarrow I_r + D_r$
    $x' \leftarrow$ homotopy$(A, y)$
    for $i \leftarrow I$ do
        $r_i \leftarrow |y - Ax'_i|$
    end for
    if $\mathrm{argmin}_i\ r_i = D_i$ then
        return$True$
    else
        return$False$
    end if
end procedure
procedure $GI_{HC}$($D_1$,$D_2$,$I$)
    for $d_j \leftarrow D_2$ do
        for $k \leftarrow K$ do
            agg$[HC(D_1, d_j, I)]$
        end for
    end for
    return$agg$
end procedure
```

5 Results

The performance of our approach is presented in Table 5 calculated using the TIRA [6]. Our approach performed the best for English (3rd position) while the worst performance was for Dutch (9th position). From our development experiments we hypothesize that this was related to the texts being shorter for this language than the rest.

Table 2. Detailed final scores for language.

Approach	AUC	C@1	Score
Dutch	0.59	0.56	0.33
English	0.74	0.69	0.51
Greek	0.80	0.72	0.58
Spanish	0.76	0.67	0.51

6 Discussion

In this year's submission we have explored the use of Homotopy-based Classification (HBC) for the verification of authorship. This is a continuation from our previous work. In particular in this edition we embedded our approach into the general impostor method. The performance of our system was stable for three languages: English, Greek and Spanish; but it was severely affected by the size of text in the Dutch case.

References

1. Asif, M.S., Romberg, J.: Sparse recovery of streaming signals using l1-homotopy. arXiv preprint arXiv:1306.3331 (2013)
2. Castillo, E., Cervantes, O., Vilariño, D., Pinto, D., León, S.: Unsupervised method for the authorship identification task. In: Working Notes for CLEF 2014 Conference. pp. 1035–1041 (2014)
3. Donoho, D.L., Tsaig, Y.: Fast solution of-norm minimization problems when the solution may be sparse. Information Theory, IEEE Transactions on 54(11), 4789–4812 (2008)
4. Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C.E., Howald, B.S.: Identifying authorship by byte-level n-grams: The source code author profile (scap) method. International Journal of Digital Evidence 6(1), 1–18 (2007)
5. Fréry, J., Largeron, C., Juganaru-Mathieu, M.: Ujm at clef in author identification. In: Working Notes for CLEF 2014 Conference. pp. 1042–1048 (2014)
6. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). pp. 1125–1126. ACM (Aug 2012)
7. Halvani, O., Steinebach, M., Zimmermann, R.: Authorship verification via k-nearest neighbor estimation notebook for pan at clef 2013. In: Working Notes for CLEF 2013 Conference (2013)
8. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3), 233–334 (Dec 2006)
9. Khonji, M., Iraqi, Y.: A slightly-modified gi-based author-verifier with lots of features (asgalf). In: Working Notes for CLEF 2014 Conference. pp. 977–983 (2014)
10. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60(1), 9–26 (2009)
11. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60(1) (2009)
12. Koppel, M., Seidman, S.: Automatically identifying pseudepigraphic texts. In: EMNLP. pp. 1449–1454 (2013)

13. Layton, R., Watters, P., Dazeley, R.: Local n-grams for author identification–notebook for pan at clef 2013. In: Working Notes for CLEF 2013 Conference (2013)
14. Ledesma, P., Fuentes, G., Jasso, G., Toledo, A., Meza, I.: Distance learning for author verification. In: CLEF 2013 Evaluation Labs and Workshop - Online Working Notes (2013)
15. Mayor, C., Gutierrez, J., Toledo, A., Martinez, R., Ledesma, P., Fuentes, G., Meza, I.: A single author style representation for the author verification task. In: Working Notes for CLEF 2014 Conference. pp. 1079–1083 (2014)
16. Moreau, E., Jayapal, A., Vogel, C.: Author verification: Exploring a large set of parameters using a genetic algorithm. In: Working Notes for CLEF 2014 Conference. pp. 1092–1103 (2014)
17. Seidman, S.: Authorship verification using the impostors method–notebook for pan at clef 2013. In: Working Notes for CLEF 2013 Conference (2013)
18. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
19. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2009)
20. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., Lopez Lopez, A., Potthast, M., Stein, B.: Overview of the Author Identification Task at PAN 2015. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2015), http://www.clef-initiative.eu/publication/working-notes
21. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(2), 210–227 (2009)