Automatically Identifying Pseudepigraphic Texts

Moshe Koppel
Bar Ilan University
Ramat-Gan, 52900, Israel
moishk@gmail.com

Shachar Seidman
Bar Ilan University
Ramat-Gan, 52900, Israel
shachar9@gmail.com

Abstract

The identification of pseudepigraphic texts – texts not written by the authors to which they are attributed – has important historical, forensic and commercial applications. We introduce an unsupervised technique for identifying pseudepigrapha. The idea is to identify textual outliers in a corpus based on the pairwise similarities of all documents in the corpus. The crucial point is that document similarity not be measured in any of the standard ways but rather be based on the output of a recently introduced algorithm for authorship verification. The proposed method strongly outperforms existing techniques in systematic experiments on a blog corpus.

1 Introduction

The Shakespeare attribution problem is centuries old and shows no signs of abating. Some scholars argue that some, or even all, of Shakespeare's works were not actually written by him. The most mainstream theory – and the one that interests us here – is that most of the works were written by Shakespeare, but that several of them were not. Could modern methods of computational authorship attribution be used to detect which, if any, of the works attributed to Shakespeare were not written by him?

More generally, this paper deals with the unsupervised problem of detecting pseudepigrapha: documents in a supposedly single-author corpus that were not actually written by the corpus's presumed author. Studies as early as Mendenhall (1887), have observed that texts by a single author tend to be somewhat homogeneous in style. If this is indeed the case, we would expect that pseudepigrapha would be detectable as outliers.

Identifying such outlier texts is, of course, a special case of general outlier identification, one of the central tasks of statistics. We will thus consider the pseudepigrapha problem in the context of the more general outlier detection problem.

Typically, research on textual outliers assumes that we have a corpus of known authentic documents and are asked to decide if a specified other document is authentic or not (Juola and Stamatatos, 2013). One crucial aspect of our problem is that we do not assume that any specific text in a corpus is known a priori to be authentic or pseudepigraphic; we can assume only that most of the documents in the corpus are authentic.

The method we introduce in this paper builds on the approach of Koppel and Winter (2013) for determining if two documents are by the same author. We apply that method to every pair of documents in a corpus and use properties of the resulting adjacency graph to identify outliers. In the following section, we briefly outline previous work. In Section 3 we provide a framework for outlier detection and in Section 4 we describe our method. In Section 5 we describe the experimental setting and give results and in Section 6 we present results for the plays of Shakespeare.

2 Related Work

Identifying outlier texts consists of two main stages: first, representing each text as a numerical vector representing relevant linguistic features of the text and second, using generic methods to identify outlier vectors.

There is a vast literature on generic methods for outlier detection, summarized in Hodge & Austin (2004) and Chandola et al. (2009). Since our prob-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1449–1454,
Seattle, Washington, USA, 18-21 October 2013.

lem setup does not entail obtaining any labeled examples of authentic or outlier documents, supervised and semi-supervised methods are inapplicable. The available methods are unsupervised, principally probabilistic or proximity-based methods. A classical variant of such methods for univariate normally distributed data uses the the z-score (Grubbs, 1969). Such simple univariate outlier detectors are, however, inappropriate for identifying outliers in a high-dimensional textual corpus. Subsequent work, such as the *Stahel-Donoho Estimator* (Stahel, 1981; Donoho, 1982), *PCout* (Filzmoser et al., 2008), *LOF* (Breunig and Kriegel, 2000) and *ABOD* (Kriegel et al., 2008) have generalized univariate methods to high-dimensional data points.

In his comprehensive review of outlier detection methods in textual data, Guthrie (2008) compares a variety of vectorization methods along with a variety of generic outlier methods. The vectorization methods employ a variety of lexical and syntactic stylistic features, while the outlier detection methods use a variety of similarity/distance measures such as cosine and Euclidean distance. Similar methods have also been used in the field of intrinsic plagiarism detection, which involves segmenting a text and then identifying outlier segments (Stamatatos, 2009; Stein et al., 2010).

3 Proximity Methods

Formally, the problem we wish to solve is defined as follows: Given a set of documents $\mathbf{D} = \{\mathbf{d_1},\ldots,\mathbf{d_n}\}$, all or most of which were written by author A, which, if any, documents in D were not written by A?

We begin by considering the kinds of proximity methods for textual outlier detection considered by Guthrie (2008) and in the work on intrinsic plagiarism detection; these will serve as baseline methods for our approach. The idea is simple: mark as an outlier any document that is too far from the rest of the documents in the corpus.

We briefly sketch the key steps:

1. *Represent a document as a numerical vector.*
   The kinds of measurable features that can be used to represent a document include frequencies of word unigrams, function words, parts-of-speech and character n-grams, as well as complexity measures such as type/token ratio, sentence and word length and so on.
2. *Measure the similarity of two document vectors.*
   We can use either inverses of distance measures such as Euclidean distance or Manhattan distance, or else direct similarity measures such as cosine or min-max.
3. *Use an aggregation method to measure the similarity of a document to a set of documents.*
   One approach is to simply measure the distance from a document to the centroid of all the other documents (*centroid* method). Another approach is to first measure the similarity of a document to each other document and then to aggregate the results by averaging all the obtained values (*mean* method):
   $$mean_{\boldsymbol{d_t}\in\boldsymbol{D}}\big(sim(\boldsymbol{d_i},\boldsymbol{d_t})\big)$$
   Alternatively, we can average the values only for the k nearest neighbors (*k-NN* method):
   $$mean_{\boldsymbol{d_t}\in\boldsymbol{D_k}}\big(sim(\boldsymbol{d_i},\boldsymbol{d_t})\big)$$
   (where $\mathbf{D_k}$ = k nearest neighbors of $\mathbf{d_i}$).
   Yet another method is to use median distance (*median* method).
   $$median_{\boldsymbol{d_t}\in\boldsymbol{D}}\big(sim(\boldsymbol{d_i},\boldsymbol{d_t})\big)$$
   We note that the centroid method and mean method suffer from the *masking effect* (Bendre and Kale, 1987; Rousseeuw and Leroy, 2003): the presence of some outliers in the data can greatly distort the estimator's results regarding the presence of other outliers. The k-NN method and the median method are both much more robust.
4. *Choose some threshold beyond which a document is marked as an outlier.*
   Choosing the threshold is one of the central issues in statistical approaches. For our purposes, however, the choice of threshold is simply a parameter trading off recall and precision.

4 Second-Order Similarity

Our approach is to use an entirely different kind of similarity measure in Step 2. Rather than use a first-order similarity measure, as is customary, we employ a second-order similarity measure that is the output of an algorithm used for the *authorship verification problem* (Koppel et al. 2011), in which we need to determine if two, possibly short, documents were written by the same author.

That algorithm, known as the "impostors method" (*IM*), works as follows. Given two docu-

ments, $\mathbf{d_1}$ and $\mathbf{d_2}$, generate an appropriate set of impostor documents, $\mathbf{p_1},\ldots,\mathbf{p_m}$ and represent each of the documents in terms of some large feature set (for example, the frequencies of various words or character n-grams in the document). For some random subset of the feature set, measure the similarity of $\mathbf{d_1}$ to $\mathbf{d_2}$ as well as to each of the documents $\mathbf{p_1},\ldots,\mathbf{p_m}$ and note if $\mathbf{d_1}$ is closer to $\mathbf{d_2}$ than to any of the impostors. Repeat this k times, choosing a different random subset of the features in each iteration. If $\mathbf{d_1}$ is closer to $\mathbf{d_2}$ than to any of the impostors (and likewise switching the roles of $\mathbf{d_1}$ and $\mathbf{d_2}$) for at least θ% of iterations, then output that $\mathbf{d_2}$ and $\mathbf{d_1}$ are the same author. (The parameter θ is used to trade-off recall and precision.)

Adapting that method for our purposes, we use the proportion of iterations for which $\mathbf{d_1}$ is closer to $\mathbf{d_2}$ than to any of the impostors as our similarity measure (adding a small twist to make the measure symmetric over $\mathbf{d_1}$ and $\mathbf{d_2}$, as can be seen in line 2.2.2 of the algorithm). More precisely, we do the following:

Given: Corpus $\mathbf{D=\{d_1,\ldots,d_n\}}$

1. Choose a feature set **FS** for representing documents, a first-order similarity measure *sim*, and an impostor set $\{\mathbf{p_1},\ldots,\mathbf{p_m}\}$.
2. For each pair of documents $<\mathbf{d_i}, \mathbf{d_j}>$ in set **D**:
   - 2.1. Let $sim2(\mathbf{d_i}, \mathbf{d_j}) := 0$
   - 2.2. Iterate *K* times:
     - 2.2.1. Randomly choose 40% of features in **FS**
     - 2.2.2. If $sim(\mathbf{d_i}, \mathbf{d_j})^2 > max_{u\in\{1,..,m\}} sim(\mathbf{d_i}, \mathbf{p_u}) * max_{u\in\{1,..,m\}} sim(\mathbf{d_j}, \mathbf{p_u})$, then $sim2(\mathbf{d_i}, \mathbf{d_j}) := sim2(\mathbf{d_i}, \mathbf{d_j}) + 1/K$
3. For each document $\mathbf{d_i}$ in set **D**:
   - 3.1. Compute $sim2(\mathbf{d_i}, \mathbf{D}) = agg_{w\in\{1,..,n\}}[sim2(\mathbf{d_i}, \mathbf{d_w})]$ where *agg* is some aggregation function
   - 3.2. If $sim2(\mathbf{d_i}, \mathbf{D}) < \theta$ (where θ is a parameter), then mark $\mathbf{d_i}$ as *outlier*.

The method for choosing the impostor set is corpus-dependent, but quite straightforward: we simply choose random impostors from the same genre and language as the documents in question. The choice of feature set **FS,** first-order similarity measure *sim*, and aggregation function *agg* can be varied. For **FS**, we simply use bag-of-words (BOW). As for *sim* and *agg*, we show below results of experiments comparing the effectiveness of various choices for these parameters.

Using second-order similarity has several surface advantages over standard first-order measures. First, it is decisive: for most pairs, second-order similarity will be close to 0 or close to 1. Second, it is self-normalizing: scaling doesn't depend on the size of the underlying feature sets or the lengths of the documents. As we will see, it is also simply much more effective for identifying outliers.

5 Experiments

We begin by assembling a corpus consisting of 3540 blog posts written by 156 different bloggers. The blogs are taken from the blog corpus assembled by Schler et al. (2006) for use in authorship attribution tasks. Each of the blogs was written in English by a single author in 2004 and each post consists of 1000 words (excess is truncated).

For our initial experiments, each trial consists of 10 blog posts, all but *p* of which are by a single blogger. The number of pseudepigraphic documents, *p*, is chosen from a uniform distribution over the set {0,1,2,3}. Our task is to identify which, if any, documents in the set are not by the main author of the set. The pseudepigraphic documents might be written by a single author or by multiple authors.

To measure the performance of a given similarity measure *sim*, we do the following in each trial:

1. Represent each document in the trial set **D** in terms of BOW.
2. Measure the similarity of each pair of documents in the trial set using the similarity measure *sim*.
3. Using some aggregation function *agg*, compute for each document $\mathbf{d_i}$: $sim(\mathbf{d_i}, \mathbf{D}) = agg_{w\in\{1,..,n\}}[sim(\mathbf{d_i}, \mathbf{d_w})]$.
4. If $sim(\mathbf{d_i}, \mathbf{D}) < \theta$, mark $\mathbf{d_i}$ as an outlier (where θ is a parameter).

Our objective is to show that results using second-order similarity are stronger than those using first-order similarity. Before we do this, we need to determine the best aggregation function to use in our experiments. In Figure 1, we show recall-precision breakeven values (for the *outlier* class) over 250 independent trials, for each of our four first-order similarity measures (inverse Euclidean, inverse Manhattan, cosine, min-max) used in conjunction with each of four aggregation functions (centroid, mean, k-NN mean, median). As is evident, k-NN is the best aggregation function in each case. We will give these baseline methods an advantage by using k-NN as our aggregation function in all our subsequent experiments.

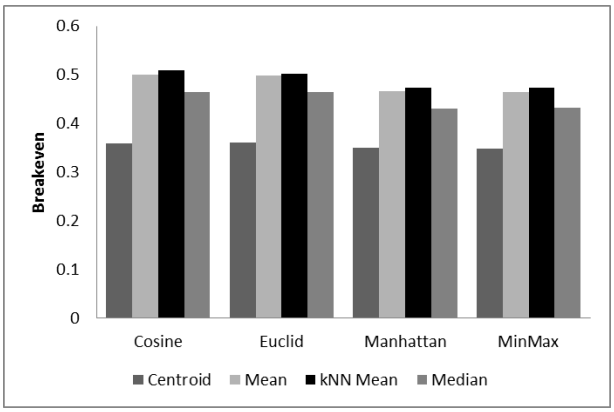

Figure 1. Breakeven values on first-order similarity measures with various aggregation functions.

We are now ready to perform our main experiment. We use BOW as our feature set and k-NN as our aggregation function. We use 500 random blog posts as our impostor set. In Figure 2, we show recall-precision curves for outlier documents over 250 independent trials, as just described, using four first-order similarity measures as well our second-order similarity measure using each of the four as a base measure. As can be seen, even the worst second-order similarity measure significantly outperforms all the standard first-order measures. In Figure 3, we show the breakeven values for each measure, pairing each first-order measure with the second-order measure that uses it as a base. Clearly, the mere use of a second-order method improves results, regardless of the base measure.

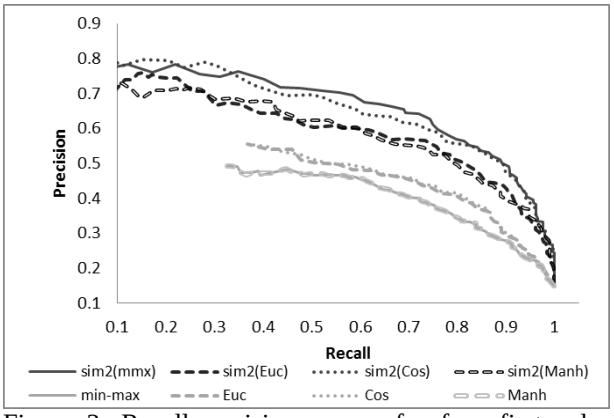

Figure 2. Recall-precision curves for four first-order similarity measures and four second-order similarity measures, based on 250 trials of 10 documents each.

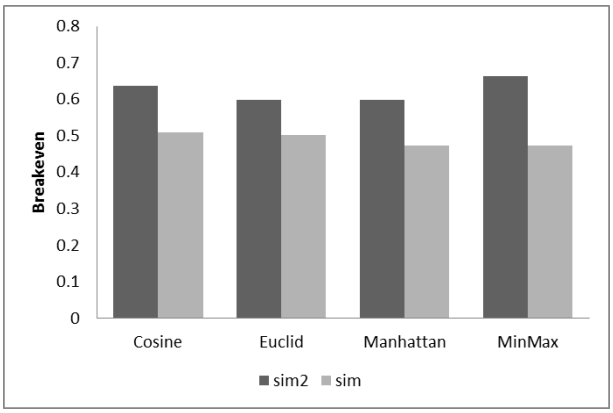

Figure 3. Breakeven values for first-order measures and corresponding second-order measures.

Thus far we have considered authorial corpora consisting of only ten documents. In Figures 4 and 5, we repeat the experiment described in Figures 2 and 3 above, but with each trial consisting of 50 documents including any number of pseudepigraphic documents in the range 0 to 15. The same phenomenon is apparent: second-order similarity strongly improves results over the corresponding first-order base similarity measure.

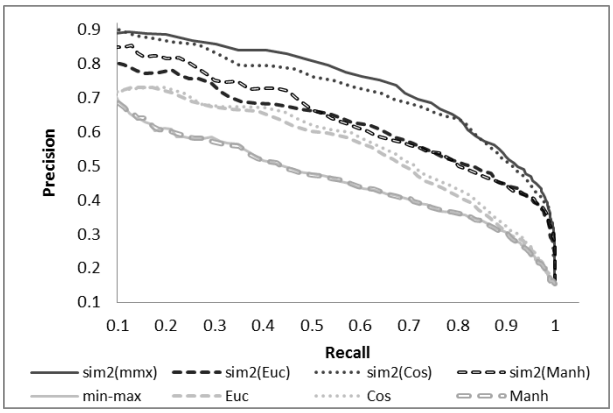

Figure 4. Recall-precision curves for four first-order similarity measures and four second-order similarity measures, based on 250 trials of 50 documents each.

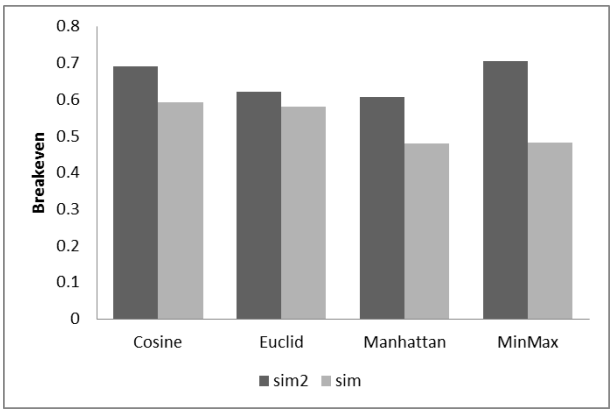

Figure 5. Breakeven values for first-order measures and corresponding second-order measures

6 Results on Shakespeare

We applied our methods to the texts of 42 plays by Shakespeare (taken from Project Gutenberg). We included two plays by Thomas Kyd as sanity checks. In addition, we included three plays occasionally attributed to Shakespeare, but generally regarded by authorities as pseudepigrapha (*A Yorkshire Tragedy*, *The Life of Sir John Oldcastle* and *Pericles Prince of Tyre*). We also included *King Edward III* and *King Henry VI* (Part 1), both of which are subjects of dispute among Shakespeare scholars. As impostors we used 39 works by contemporaries of Shakespeare, including Christopher Marlowe, Ben Jonson and John Fletcher.

We found that the two plays by Thomas Kyd and the three pseudepigraphic plays were all among the seven furthest outliers, as one would expect. In addition, *King Edward III* was 9th furthest. *King Henry VI* (Part 1) was not found to be an outlier at all. Curiously, however, three undisputed plays by Shakespeare were found to be greater outliers than *King Edward III*. These are *The Merry Wives of Windsor*, *The Comedy of Errors* and *The Tragedy of Julius Caesar*. *The Merry Wives of Windsor* is a particularly distant outlier, even further out than *Oldcastle* and *Pericles*. We leave it to Shakespeare scholars to explain the reasons for these anomalies.

7 Conclusion

In this paper we defined the problem of unsupervised outlier detection in the authorship verification domain. Our method combines standard outlier detection methods with a novel inter-document similarity measure. This similarity measure is the output of the impostors method recently developed for solving the authorship verification problem. We have found that use of the kNN method for outlier detection in conjunction with this second-order similarity measure strongly outperforms methods based on any outlier detection method used in conjunction with any standard first-order similarity measures. This improvement proves to be robust, holding for various corpus sizes and various underlying base similarity measures used in the second-order similarity measure.

The method can be used to resolve historical conundrums regarding the authenticity of works in questioned corpora, such as the Shakespeare corpus briefly considered here. This is currently the subject of our ongoing research.

References

S. M. Bendre and B. K. Kale. 1987. *Masking effect on tests for outliers in normal samples*, Biometrika, 74(4):891-896.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander. 2000. *LOF: Identifying Density-Based Local Outliers*, ACM SIGMOD Conference Proceedings.

Varun Chandola, Arindam Banerjee and Vipin Kumar. 2009. *Anomaly detection: a survey*. ACM Computing Surveys 41, 3, Article 15.

David L. Donoho. 1982. *Breakdown properties of multivariate location estimators*. Ph.D. qualifying paper, Harvard University.

Peter Filzmoser, Ricardo Maronna and Mark Werner. 2008. *Outlier identification in high dimensions*. Computational Statistics and Data Analysis, 52:1694-1711.

David Guthrie. 2008. *Unsupervised Detection of Anomalous Text*. PhD Thesis, University of Sheffield.

Frank E. Grubbs. 1969. *Procedures for detecting outlying observations in samples*, Technometrics.

V.J. Hodge and J. Austin. 2004. A survey of outlier detection methodologies. Artificial. Intelligence Review, 22 (2). pp. 85-126.

Patrick Juola and Efstathios Stamatatos. 2013. *Overview of the Author Identification Task at PAN 2013*. P. Forner, R. Navigli, and D. Tufis (eds) CLEF 2013 Evaluation Labs and Workshop –Working Notes Papers.

Moshe Koppel and Jonathan Schler 2004. *Authorship verification as a one-class classification problem*. In ICML '04: Twenty-first International Conference on Machine Learning, New York, NY, USA.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. *Authorship attribution in the wild*. Language Resources and Evaluation, 45(1): 83–94.

Moshe Koppel M. and Yaron Winter. 2013. *Determining If Two Documents Are by the Same Author*. J. Am. Soc. Inf. Sci. Technol.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Reading, Mass. Addison Wesley.

Hans-Peter Kriegel, Matthias S. Schubert and Arthur Zimek. 2008. *Angle-based outlier detection in high dimensional data*. Proc. KDD.

Thomas C. Mendenhall. 1887. *The characteristic curves of composition*, Science 9, 237-259.

Sridhar Ramaswamy, Rajeev Rastogi and Kyuseok Shim. 2000. *Efficient Algorithms for Mining Outliers from Large Data Sets*. Proc. ACM SIDMOD Int. Conf. on Management of Data.

Peter J. Rousseeuw. 1984. *Least median of squares regression*. Journal of the American Statistical Association, 79(388):87-880.

Peter J. Rousseeuw and Annick M. Leroy. 2003. *Robust Regression and Outlier Detection*. John Wiley & Sons.

J. Schler, M. Koppel, S. Argamon and J. Pennebaker. 2006. *Effects of Age and Gender on Blogging*. in Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Werner A. Stahel. 1981. *Breakdown of covariance estimators*. Research Report 31, Fachgruppe f¨ur Statistik, Swiss Federal Institute of Technology (ETH), Zurich.

Efstathios Stamatatos. 2009. *Intrinsic plagiarism detection using character n-gram profiles*. Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse. pp. 38–46.

Benno Stein B, Nedim Lipka and Peter Prettenhofer. 2010. *Intrinsic Plagiarism Analysis*. Language Resources and Evaluation, 1–20. 2010.

Benno Stein B, Nedim Lipka and Peter Prettenhofer. 2010. *Intrinsic Plagiarism Analysis*. Language Resources and Evaluation, 1–20. 2010.