Authorship Disambiguation and Alias Resolution in Email Data

Freek Maes Johannes C. Scholtes

Department of Knowledge Engineering
Maastricht University, P.O. Box 616, 6200 MD Maastricht

Abstract

Given a data set of email messages we are interested in how to resolve aliases and disambiguate authors even if their names are misspelled, if they use completely different email addresses or if they deliberately use aliases. This is done by using a combination of string similarity metrics and techniques from authorship attribution and link analysis. These techniques are combined by using a voting algorithm that is based on a Support Vector Machine. The approach is tested on a cleaned subset of the ENRON email data set. The results show that a combination of Jaro-Winkler email address similarity, Support Vector Machine on writing style attributes and Jaccard similarity of the link network outperforms the use of each of these techniques separately.

1 Introduction

In this paper a description will be given of a new approach to the problem of disambiguating authorship and resolving aliases in email data. The techniques that are commonly used for authorship detection in literary texts cannot readily be applied to email data for a number of reasons. (1) The number of potential authors in an email data set can be very large, whereas traditional authorship attribution problems only deal with small author sets (2) email data is often sparse and can be very noisy because of the presence of forwards/replies, duplicates and system messages. (3) the written text contained in an email message can be very short, making it hard to distill style markers from it and (4) it is not known whether a particular person in the data set uses any aliases at all, so the candidate set is an open set.

A number of approaches exist to determining authorship of email data:

- Using *string similarity metrics*, such as Jaro-Winkler [11], on the email addresses it is possible to quickly generate a list of potential aliases of an author. These string metrics are able to capture superficial aliases that results from the use of different email domains/protocols (e.g. home or work email) and spelling errors. However, these metrics often give false positive aliases, such as ”John Barker” and ”John Baker” which might actually be two different persons. Moreover, they fail to find the more sophisticated aliases where the email addresses do not look alike, such as ”Bin Laden” and ”The Prince”.

- *Authorship attribution* techniques can be used to find the author of a given email solely by looking at the writing style that a particular author employs. By training a binary classifier on a combination of lexical, syntactic, content-specific and/or semantic features derived from training messages, it is possible to determine the author of a new anonymous message. Multiple classifiers can then be combined using a one-versus-all approach such that a multi-class problem can also be solved.

- The information that is captured in the *link network* of the author can also be utilized. For example: if two authors share a great number of direct contacts the likelihood that they might be the same person increases. Similarly, information from more distantly shared contacts can be used to provide additional information about the similarity between two persons.

	Features	**Description**
Lexical		
1	Total number of characters (C)	
2	Total number of alphabetic characters / C	
3	Total number of upper-case characters / C	
4	Total number of digit characters / C	
5	Total number of white-space characters / C	
6	Total number of tab spaces / C	
7-32	Frequency of letters	A-Z
33-53	Frequency of special characters	~ @ # $ % ^ & * - _ = + > < [ ] { } / \ |
54	Total number of words (M)	
55	Total number of short words / M	less than four characters
56	Total number of characters in words / C	
57	Average word length	
58	Average sentence length (in characters)	
59	Average sentence length (in words)	
60	Total different words / M	
61	Hapax legomena	Frequency of once-occurring words
62	Hapax dislegomena	Frequency of twice-occurring words
63-82	Word length frequency distribution / M	
83-333	TF*IDF of 250 most frequent 3-grams	
Syntactic		
334-341	Frequency of punctuation	, . ? ! : ; ’ ”
342-491	Frequency of function words	
Structural		
492	Total number of sentences	

Table 1: Feature set that has been used in the authorship SVM

Since the three approaches mentioned above use information from different domains, the hypothesis of this research is that combining them will yield better results than each technique on its own. In order to combine them, a separate binary classifier on the results of the three methods has been trained that can distinguish between good and bad combinations of results.

2 Approach

Authorship based on email address similarity: Christen [3] found that when dealing with surnames the Jaro similarity metric performed best out of 27 techniques. Cohen and Fienberg [5] evaluated different string metrics on different data sets and found that the Monge-Elkan distance performed best. However, they conclude that the Jaro-Winkler metric performed almost as well as the Monge-Elkan distance, but is an order of magnitude faster. Therefore, in the experiments to follow, a Jaro-Winkler similarity has been calculated between each author-candidate pair based on their email addresses. The Jaro distance calculates the similarity between two strings based on the number of matching characters, and the number of transpositions needed to transform one string into the other. The Winkler-enhancement increases the Jaro-score when the two strings share a common prefix. This approach will hereafter be referred to as ”Jaro-Winkler” or ”JW”.

Authorship based on content: For every email message a combination of lexical, syntactic and structural features has been extracted. Examples of features that have been used are word and character frequencies, frequency of punctuation, vocabulary richness measures, 3-grams, sentence length and frequency of function words. The complete list of features, partially adapted from [12] and extended with a number of additional feature to create a larger overall feature variance, can be found in table 1. For each author a Support Vector Machine (SVM) has been created using the author’s email as positive, and a random selection of other emails as negative training examples. Both classes have been balanced in the number of training instances. This approach will be referred to as ”authorship SVM”.

Authorship based on link-analysis: Two link analysis methods have been employed in order to detect

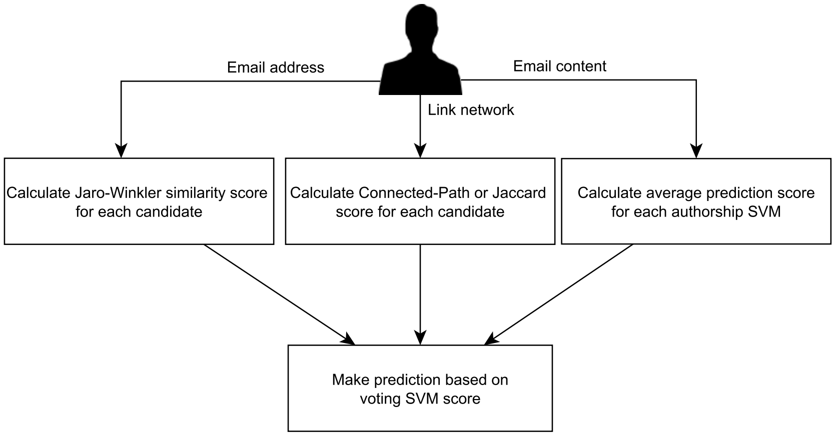
Email address
Email content
Link network
Calculate Jaro-Winkler similarity score for each candidate
Calculate Connected-Path or Jaccard score for each candidate
Calculate average prediction score for each authorship SVM
Make prediction based on voting SVM score

Figure 1: The structure of the framework.

aliases in the link network. The first one is the well-known Jaccard similarity [8], which will be referred to as ”Jaccard”. Let v, w be two authors in the data set and $N(v), N(w)$ the direct neighbors of v and w respectively. The similarity between v and w is calculated as follows:

$$Jaccard(v, w) = \frac{|N(v) \cap N(w)|}{|N(v) \cup N(w)}$$

The second link analysis method is a more sophisticated method referred to as ”Connected Path” or ”CP”. Connected Path [1] has been shown to outperform a range of well-known algorithms and metrics that can be used for alias detection in link networks, such as Jaccard similarity, Connected Triples, Pagerank and PageSim. The Connected Path-algorithm values shorter paths between authors higher than longer ones. Moreover, the more connections an author has to other authors, the lower each connection is valued. By aggregating in a smart way over all possible paths between two authors the algorithm derives a similarity metric that indicates how similar the two authors' link networks are.

Combining the results using an SVM voting algorithm: The results of these different techniques were then used to train a separate SVM. This SVM will be referred to as the ”Voting SVM”. Since the SVM performs feature ranking internally, it automatically assigns weights to different combinations of results and can distinguish between successful and unsuccessful combinations of results. If the results of one technique are ambiguous, another technique can possibly aid in making the classification decision. The general structure of the framework is summarized in figure 1. Two combinations of techniques have been tested, namely Jaro-Winkler, Connected Path similarity and authorship SVM (”JW-SVM-CP”), and Jaro-Winkler, Jaccard similarity and authorship SVM (”JW-Jaccard-SVM”). In order to avoid over-fitting, the voting SVM is trained on instances that did not occur in the test set.

The ENRON Data set: The new approach has been tested on the ENRON-data set that was made available by the Federal Energy Regulatory Commission during its investigation into fraudulent activities at ENRON [7]. A well-known version of the data set, containing roughly 500,000 email messages from 151 Enron employees, was first made available by William Cohen [4]. Later, Shetty & Adibi [10] applied preprocessing such as removing empty messages and duplicates to the data set. The Shetty & Abidi-version of the data set has been used in this research. Many records in this data set consisted of system messages, emails with little or no original text (e.g. forwards or empty messages) and duplicates. These messages have been removed in order to reduce noise. Messages where the number of words (after removing forwarded information) was smaller than or equal to 10 were also removed, since they contained too little useful information.

According to Burrows [2] 10,000 words per author is a reliable minimum for authorship attribution, whereas Sanderson and Guenter [9] mention a minimum of 5,000 words per author. Since Hirst and Feiguina

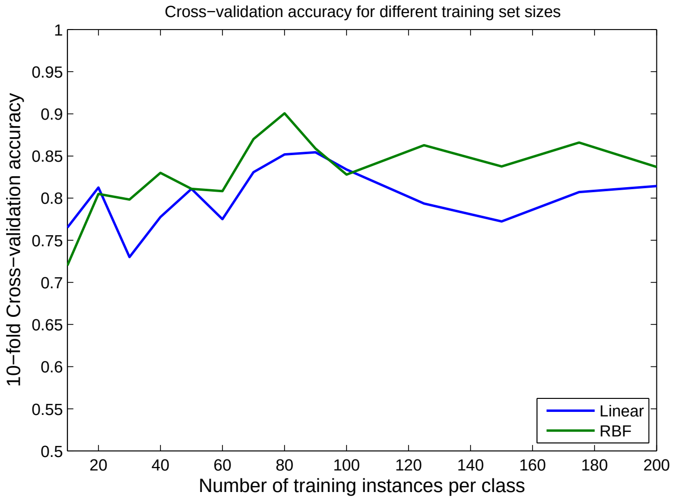

Figure 2: Averages of 10 times 10-fold cross-validation using different training set sizes and kernels for the authorship SVM.

[6] conclude that using multiple short texts for authorship attribution overcomes the problem of not having sufficiently long training texts available, there is no need to concatenate the e-mails from a single author into one long e-mail. Empirical findings on the ENRON data set, displayed in figure 2, show that using 80 training instances per class in combination with a Radial Basis Function-kernel (RBF) achieves the highest accuracy. Therefore, it was decided that a RBF-kernel should be used and that authors with a total number of emails less than 80 should be discarded. Additionally, in order to preserve balance in the number of training instances per author, authors that had sent more than 600 messages were also removed from the data set. In the final data set the average number of words per email equals 209, and with at least 80 emails per author it is ensured that each author has a reliable number of words to train on. After preprocessing the data set consisted of 44,912 emails by 246 different authors.

Training and Evaluation: Since there was no data to verify whether the ENRON-data set actually contained any real aliases, authors that had a total of more than 200 messages were split up into aliases of 100-200 messages each. For each author with more than 200 messages there were two possibilities:

- The author is split up into 1 or more artificial aliases yielding high Jaro-Winkler similarity. These are easy-to-recognize aliases, for example: *john.doe@enron.com* is split up into the aliases *john.doe@enron.comA* and *john.doe@enron.comB*.
- The author is split up into 1 or more artificial aliases yielding low Jaro-Winkler similarity. These are hard-to-recognize aliases, for example: *jane.doe@enron.com* is split up into the aliases *bin_laden* and *abu_abdallah*.

In total, 41 authors were split up into aliases with high Jaro-Winkler similarity, and 12 authors were split up into aliases with low Jaro-Winkler similarity. Emails from and to the original authors were randomly assigned to one of the author's artificial aliases, and separate authorship SVM's were trained for each alias. Splitting up authors may results in aliases with the same e-mail signatures (name, position, telephone number, etc.). However, there are not many emails in the dataset that contain such an extensive signature. In addition, the content-based approach is not affected by this since it does not take into account the most frequent n-grams per author, but the most frequently occurring n-grams in the complete data set. In order to evaluate the results of the different techniques two different test sets have been created, both of which can be seen in table 2. The first test set, called the *mixed* test set, has a fairly equal division of alias types. The second test test, called the *hard* test set, is substantially more difficult since the majority of the aliases are not easy to recognize by their email addresses. The authors in each test set were chosen at random from their respective alias categories.

Test set:	**Mixed**	**Hard**
High Jaro-Winkler	6	2
Low Jaro-Winkler	8	16
No alias	6	2

Table 2: Distribution of alias-types for two different test sets.

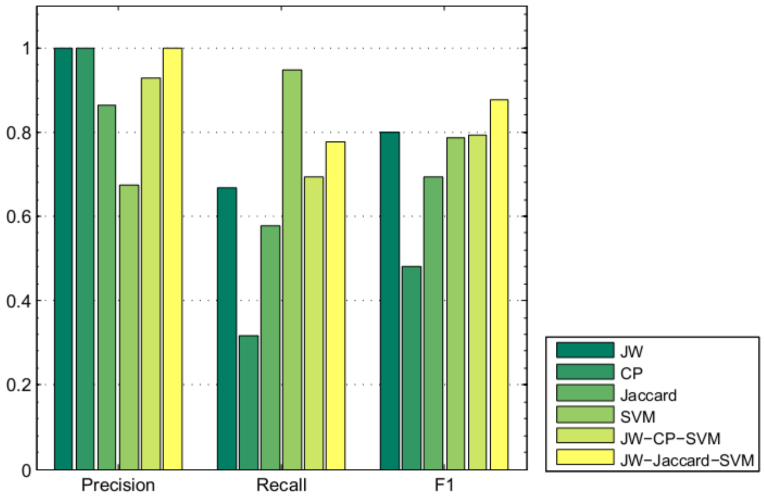

Figure 3: Precision, Recall and F1-scores for different techniques, evaluated on the mixed test set.

3 Results

Figure 3 gives an overview of the precision, recall and F1-values that correspond to the best F1-score for each technique on the mixed test set. Figure 4 gives on overview of these values for the hard test set.

Jaro Winkler: The Jaro-Winkler approach gave good results on the mixed test set, but failed on the hard test set. The high F1-score of 0.80 on the mixed test set can be explained by the fact that many of the artificial aliases had a high Jaro-Winkler similarity. The hard test set more closely mimics a real-world scenario where aliases do not look as much alike. Therefore, the best F1-score achieved by Jaro-Winkler on this test set is only 0.28. However, the results still shows that using a simple string metric can detect many aliases resulting from spelling errors or the use of different email addresses for work, home, etc.

Connected Path: The Connected Path method achieves an F1-score of 0.48 on the mixed test set, and a score of 0.53 on the hard set. It can be concluded that the Connected-Path algorithm failed to achieve good results because of three reasons. First, since authors have been split up into aliases and some have been removed all together, the link network's structure might have been corrupted. This especially affects link analysis that goes beyond the analysis of direct neighbors, since it takes into account more complicated link connections. Second, the link network search has been performed to depth 3, which means that only the information contained in paths of length 2 and 3 have been used in the calculation of the similarity score. Third, the Connected Path method can only return similarity scores for authors within close proximity of the original author. If there was no Connected Path score returned for a particular author-alias pair the alias had to be counted as a false negative.

Jaccard: Using Jaccard similarity yielded better results than the Connected Path algorithm: an F1-score of 0.69 and 0.67 for the mixed test set and hard test set respectively. Since the Jaccard similarity only takes into account direct neighbors, it is less affected by changes in the link network. Moreover, the Jaccard similarity can be calculated between any two authors in the data set, which is why it scored significantly better than the Connected Path method.

Authorship SVM: The use of authorship SVM's gave good results overall, with an F1-score of 0.79 on the mixed test set and 0.76 on the hard test set. The results are especially good considering the fact that there are 314 candidate aliases for each author and that the training texts are short.

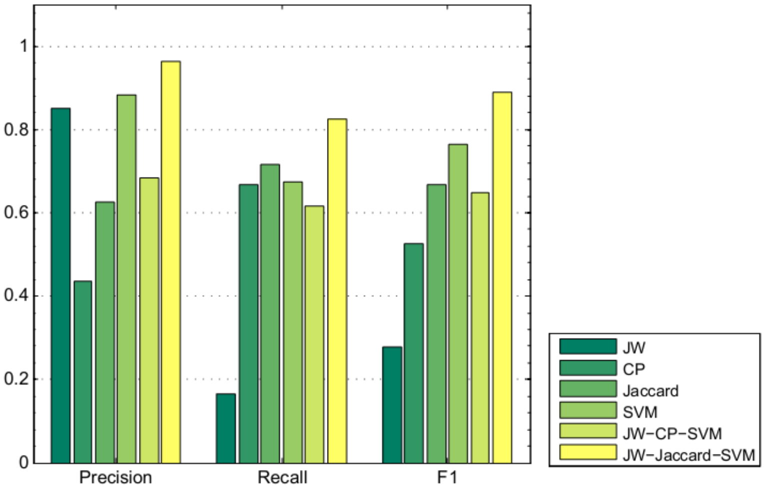

Figure 4: Precision, Recall and F1-scores for different techniques, evaluated on the hard test set.

Combined techniques: It can be concluded that the highest F1-score for both test sets is achieved by JW-Jaccard-SVM. For the mixed test set an F1-score of 0.88 was achieved, whereas on the hard test set an F1-score of 0.89 was achieved. These results confirm our hypothesis that a combination of techniques can yield better results than using these techniques individually. However, the combination of JW-CP-SVM on the mixed test set performed as good as authorship SVM or even Jaro-Winkler alone, with an F1-score of 0.80. For the hard test set it performed even worse, achieving an F1-score of 0.65. Because of aforementioned reasons, the Connected Path method failed to achieve good results in general. In combination with the low Jaro-Winkler performance on the hard data set this resulted in the combination JW-CP-SVM failing to achieve reasonable results.

4 Conclusion

The combination of Jaro-Winkler similarity, authorship SVM and Jaccard similarity outperforms individual and other combinations of techniques, achieving an F1-score of 0.89. It is important to note that the relative improvement in F1-score of the combined techniques over the individual techniques is dependent on the number of low Jaro-Winkler aliases in the test set. This indicates that the different techniques are indeed complementary and can work together to achieve better results. It can therefore be concluded from these results that it is beneficial to combine techniques from different domains using a voting SVM.

5 Future research

This paper showed that combinations of techniques can outperform the use of a single technique when applied to a real-life data set. It will be interesting to see how well these techniques perform on a full data set with real aliases, which could not be found to use in this research. Should such a collection not exist, it is worthwhile to create one.

The link analysis techniques that have been used in this paper only use information from the direct neighborhood of the authors. Boongoen et al.[1] have already shown that searching to a greater depth yields better results, so it would be useful to look at how the algorithm can be optimized to be less computationally intensive in order to search to greater depths.

Finally, the assumption has been made that the results from various techniques are independent of each other. These assumptions have not been tested, and it is not clear if and in what way various techniques affect each other. Therefore, it is important that more research will be done to examine the best choice of feature sets, techniques and aggregation methods.

References

[1] Tossapon Boongoen, Qiang Shen, and Chris Price. Disclosing false identity through hybrid link analysis. *Artificial Intelligence and Law*, 18(1):77–102, February 2010.

[2] John Burrows. All the way through: Testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22(1):27–47, January 2007.

[3] P. Christen. A comparison of personal name matching: Techniques and practical issues. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, pages 290–294. IEEE, 2006.

[4] William W. Cohen. Enron Email Dataset. Retrieved from: `http://www.cs.cmu.edu/~enron/`, 2009.

[5] William W. Cohen, P. Ravikumar, and S.E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, pages 73–78, 2003.

[6] O. Feiguina and G. Hirst. Authorship attribution for small texts: Literary and forensic experiments. In *Proceedings of the 30th International Conference of the Special Interest Group on Information Retrieval: Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (SIGIR)*, 2007.

[7] FERC. Information Released in Enron Investigation. Retrieved from: `http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp`.

[8] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[9] C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 482–491. Association for Computational Linguistics, 2006.

[10] Jitesh Shetty and Jafar Adibi. The enron email dataset: Database schema and brief statistical report. Technical report, Information Sciences Institute, 2004.

[11] William E Winkler. The state of record linkage and current research problems. *Statistical Research Division US Census Bureau*, pages 1–15, 1999.

[12] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.