Journal of Computer Science 6 (3): 235-243, 2010
ISSN 1549-3636

Dropping down the Maximum Item Set: Improving the Stylometric Authorship Attribution Algorithm in the Text Mining for Authorship Investigation

Tareef Kamil Mustafa, Norwati Mustapha, Masrah Azrifah Azmi and Nasir B. Sulaiman
Faculty of Computer Science and Information Technology, University Putra Malaysia,
P.O. Box 43400, UPM Serdang, Selangor, Malaysia

Abstract: Problem statement: Stylometric authorship attribution is an approach concerned about analyzing texts in text mining, e.g., novels and plays that famous authors wrote, trying to measure the authors style, by choosing some attributes that shows the author style of writing, assuming that these writers have a special way of writing that no other writer has; thus, authorship attribution is the task of identifying the author of a given text. In this study, we propose an authorship attribution algorithm, improving the accuracy of Stylometric features of different professionals so it can be discriminated nearly as well as fingerprints of different persons using authorship attributes. **Approach:** The main target in this study is to build an algorithm supports a decision making systems enables users to predict and choose the right author for a specific anonymous author's novel under consideration, by using a learning procedure to teach the system the Stylometric map of the author and behave as an expert opinion. The Stylometric Authorship Attribution (AA) usually depends on the frequent word as the best attribute that could be used, many studies strived for other beneficiary attributes, still the frequent word is ahead of other attributes that gives better results in the researches and experiments and still the best parameter and technique that's been used till now is the counting of the bag-of-word with the maximum item set. **Results:** To improve the techniques of the AA, we need to use new pack of attributes with a new measurement tool, the first pack of attributes we are using in this study is the (frequent pair) which means a pair of words that always appear together, this attribute clearly is not a new one, but it wasn't a successive attribute compared with the frequent word, using the maximum item set counters. the words pair made some mistakes as we see in the experiment results, improving the winnow algorithm by combining it with the computational approach, achieved by using the CV statistical tool as a conditional threshold for attribute selecting; by doing so, the frequent pair result improved from 50% error to 0% in the improved frequent pair with a clear higher score result compared with the frequent word attribute. **Conclusion/Recommendations:** The new CV algorithm results improvement may lead to several new attributes usage that gave unsatisfying results before that might improve the direction for solving some hard cases couldn't be solved till now.

Key words: Text mining, Stylometric attribution, authorship attribution, winnow algorithm, computational stylistic

INTRODUCTION

Text mining is the "discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" (Argamon *et al*., 2003). A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation (Argamon *et al*., 2003; Hearst, 2003), while Style, concerns the way in which a document is written rather than its contents; stylistics is the study of style. Automated analysis of stylistics can be applied to a range of problems, from document attribution and authentication to matching document readability (Tareef, 2007). Stylometric Authorship Attribution (AA) can be considered as a typical clustering, classification and association rule problem, where a set of documents with known authorship are used for training and the aim is to automatically determine the corresponding author of an anonymous text (Mustafa *et al*., 2009). In contrast to other classification tasks, it is not clear which features of a text should be used to classify an author. Consequently,

Corresponding Author: Tareef Kamil Mustafa, Faculty of Computer Science and Information Technology, University Putra Malaysia, P.O. Box 43400, UPM Serdang, Selangor, Malaysia

the main concern of computer-assisted authorship attribution is to define an appropriate characterization of documents that captures the writing style of authors. Our proposal is to construct this characterization by pair-of-words sequences. It is important to mention that word pair sequences have been applied without much success compared with traditional frequent word classification (Argamon and Levitan, 2005) and our goal here is to improve the AA algorithm to have a new pair of words attributes works as successful as the frequent word. In this study, the keystone for improving is the accuracy of the Giving results for Authorship Attribution (AA) is by using new methodology, for a start, then by improving the computational stylistics algorithms with statistical methods and using a new computational function, gaining the ability to verify other attributes than the frequent words, by this combination, we have already built a new winnow algorithm for (AA), then use our improved winnow algorithm with a ratio scale to support decision making. Results show that the frequent pair experiment have no predicting mistakes, at the contrary, the words pair made some mistakes as we see in the experiment results; improving the winnow algorithm by combining it with the computational approach, achieved by using the CV statistical tool as a conditional threshold for attribute selecting; by doing so, the frequent pair result improved from 50% error to 0% in the improved frequent pair with a clear higher score result compared with the frequent word attribute

Stylistic analysis that has been done by Croft (1981) claimed that many references points that for a given author, the habits “of style” are not affected “by passage of time, change of subject matter or literary form. They are thus stable within an author's writing, but they have been found to vary from one author to another”.

Stylistics, which may be defined as the study of the language of literature, makes use of various tools of linguistic analysis. Corpus linguistics is opening up new vistas for the study of language and there are interesting similarities in the approaches of stylistics and corpus linguistics, using theories relating to phonetics, syntax and semantics. Theories and techniques of analysis from authorship attribution of documents has given some prior stylistic characteristics of the author’s writing extracted from a corpus of known works, e.g., authentication of disputed documents or literary works. Although the pioneering paper based on word length histograms appeared at the very end of the nineteenth century (Malyutov, 2006), the resolution power of this and other Stylometric approaches is yet to be studied both theoretically and on case studies such that additional information can assist in finding the correct attribution. The pioneering Stylometric study by Mendenhall (Malyutov, 2006) was based on histograms of word-length distribution of various authors (Malyutov, 2006). The study showed significant differences of these histograms for different languages and also for different authors “Dickens vs. Thackeray” using the same language. Other studies described the histograms for Shakespeare contemporaries commissioned and supported by Hemminway (Malyutov, 2006). After fitting appropriate parametric family of distributions (Poisson or negative binomial), they follow the Bayes rule for odds (posterior odds is the product of prior odds times the likelihood ratio).

In the history of authorship attribution, the analysis of The Federalist Papers (USA) plays an important role. The goal in this work was to perform a correct study by using a revised corpus of the Federalist papers based in large part on Rudman’s critique. They used machine-learning techniques for analyzing the use of lexical features for authorship attribution of the papers. Another goal of the study was to explore how different corruptions of the corpus may affect the accuracy of the classification results and the differences between them (Argamon and Shlomo, 2006).

The Federalist Papers were written during the years 1787 and 1788 by Alexander Hamilton, John Jay and James Madison (Argamon and Shlomo, 2006). These 85 propaganda tracts were intended to help to get the US Constitution ratified and were all published anonymously under the pseudonym “Publics”. According to Avalon project (Yale Law School) Hamilton wrote 51 of the papers, Madison wrote 15, Jay wrote five, while three papers were written jointly by Hamilton and Madison and 11 papers have disputed authorship-either Hamilton or Madison, although most evidence points to Madison as the author.

Moreover, the study provides additional support to the almost universally accepted allocation that Madison is the author of the disputed Federalist Papers (Argamon and Shlomo, 2006).

All the methods and algorithms that have been stated here, even if they are indexed according to their first testing appearance date, are still working together, that means, that no old method is replaced by a newer method, all the professionals and researchers choose a specific method and continued improving and testing it until now days (Tareef, 2007).

The methods and algorithms that are described here are the most frequently used, mentioned, developed and tested, not mentioning the methods that we can call “one time shot” that were more ad-hoc adventures and never been tested latter, also avoiding techniques that

gave unsatisfied results that would not assist our subject. These methods are:

- Content analysis (Krippendorf, 2003)
- Computational stylistic approach (Stamatatos *et al*., 1999)
- Exponentiated gradient learn algorithm (Argamon *et al*., 2003)
- Winnow regularized algorithm (Zhang *et al*., 2002)
- Long canons modeling as Markov chains (Malyutov, 2006)
- Burrows's delta method (Stien and Argamon, 2007)

There is no real need for entering into the details of each method listed above; researches interested in any of them can refer to the reference showing in front of each, preferring to go into the details of the new proposed method directly in the materials and methods.

MATERIALS AND METHODS

The methodology used in this work generally depends on the combination of the winnow algorithm (Zhang *et al*., 2002) and the computational stylistic approach for learning (clustering part), there is a reason to skip other Methods mentioned previously, such as the content analysis (Krippendorf, 2003), because it is the earliest type of the computational (Stamatatos *et al*., 1999), also for exponential and long canons, both methods are typical mathematical models, letter observing more than words or sentence, which is not quite the main interest of Stylometric and last method skipped is the burrow (Stien and Argamon, 2007), although it's a new improved technique, but it can be considered also a new diversion of winnow.

Testing techniques are used as a classification part and for the authentication attributes (considered as association rule part), three set types of attributes will be used in the experiment:

- Traditional frequent word
- Frequent word pair
- Improved frequent word pair

With weighting parameters given by Pearson correlation and by analyzing the proposed set of style markers, as in the Computational stylistic approach, which is based on the frequencies of the rewrite rules as they appear in a syntactically annotated corpus. Both high-frequent and medium-frequent rewrite rules give accuracy results comparable to lexically-based methods, avoiding the problems caused by the lexically based style markers that are highly language dependent, parameters that will be used with each style marker in winnow algorithm are the results of the linear regression measure represented by the Pearson correlation coefficient that has been proposed at the computational stylistics approach (Stamatatos *et al*., 1999):

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

X is the frequency of any attribute that is used by the Stylometric map for the author that we are investigating, while Y is the corresponding attribute from the frequent item set for the scripts under test or investigation, in other words, the x is gathered from the learning path, while y is from the testing path.

The proposed part in our algorithm will be by using a new threshold replacing the classic frequency threshold (maximal item set). The new threshold is the Coefficient of Variation (CV):

$$CV = \frac{S}{\bar{x}} \times 100$$

In probability theory and statistics, the Coefficient of Variation (CV) is a normalized measure of dispersion of a probability distribution. It is defined as the ratio of the standard deviation to the mean; this is only defined for non-zero mean and is most useful for variables that are always positive.

For each pair of words chosen in maximum item sets, to exclude the unstable pair of words frequency appearance, which means the pair of words that the author adopted heavily for some of his books, but not all books, so when getting the Mean for the learning task for 8 books to build a Stylometric map for the experiment author: Mark Twain, we find a big standard deviations also, showing the instability of that writing habit under consideration, i.e., (the king) pair was excluded because the high frequency in twains map came from just one book, "A Connecticut Yankee in King Arthur's Court" (Table 1).

That event was usually noticeable by excluding manually some of the frequent pair that appeared because of some certain futures used in the novel influenced by its environment and not by the writer's habits.

Table 1: The dataset

Author	Book title	Size (KB)	Task
Mark Twain	What is Man	532	Learn
Mark Twain	The Adventures of Huckleberry Finn	563	Learn
Mark Twain	The Prince and the Pauper	374	Learn
Mark Twain	Roughing It	922	Learn
Mark Twain	How to Tell a Story	40	Learn
Mark Twain	A Horse's Tale	107	Learn
Mark Twain	The Stolen White Elephant	60	Learn
Mark Twain	A Connecticut Yankee in King Arthur's Court	661	Learn
Shakespeare	The Tragedy of Antony and Cleopatra	167	Test
Jack London	The Mutiny of the Ellsinore	627	Test
Mark Twain	A Dogs Tale	641	Test
Mark Twain	The Adventure of Tom Sawyer	387	Test

It should be known that the standard deviation cannot be used here as threshold since the value is affected by the Mean value, so we skipped for CV which its main usage is passing this negative effect, even with this manual kind of help couldn't save the maximum item set for pair of words attribute at the end of the experiment as you will notice:

$$S = \sqrt{\frac{\sum_{i-1}^{n} X^2 - \frac{(\sum_{i-1}^{n} X_i)^2}{n}}{n-1}}$$

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots X_n}{n} = \frac{\sum_{j-1}^{n} x_j}{n} \frac{\sum X_j}{n}$$

Last step in the algorithm is implicating the computational results in the winnow algorithm, for the suspected authors as a comparing measure, to compare it with the winnow result of the tested text under investigation, to give the automated decision in the three types of attribute sets; below we show the classic steps for winnow before any improving steps (Zhang *et al*., 2002):

The winnow algorithm (simple version)

1. Initialize the weights $w_1 \ldots, w_n$ of the variables to 1.
2. Given an example $x = \{x_1, \ldots, x_n\}$,
   Output 1 if $w_1x_1 + w_2x_2 + \ldots + w_nx_n \geq n$
   And output 0 otherwise
3. If the algorithm makes a mistake:
   (a) Predict negative on a positive example, then for each x_i equal to 1, double the values of w_i
   (b) Predict positive on a negative example, then for each x_i equal to 1, cut the values of w_i in half
4. Go to 2.

The suggested steps for improved winnow algorithm will be:

1. Initialize the weights $w_1, \ldots, w_n$ of the variables to the Pearson correlation coefficient r extracted from the computational approach, for r = [-1,1]
2. For all x_i=:
   (a) If x_i is negative example then $x_i = -1$, negative example is extracted for an author we already know that he's not a candidate for the investigated corpus in the learning process
   (b) Else xi is positive, then $x_i = 1$, for positive example, is r extracted for an author we know already that he is the right author for the investigated corpus
3. Winnow result W will be [-n, n] for n = number of attributions used in the set
4. End.

- The main deference in the proposed algorithm as you will see is replacing the 'Go To' step (step 4 above in the simple winnow version) with the Pearson correlation
- There is no loops for the proposed algorithm because we weight our attributes with a fuzzy rate [-1,1] using the Pearson correlation coefficient instead of losing time by testing in each loop and then multiplying by 2 or by ½ depending on negative or positive result

The proposed methodology cannot be clarified without getting into details, As noted in Fig. 1, our methodology starts with the Data set that is been used, the methodology steps will be described clearly for each part, step by step.

Data set: The data set is taken from the web site www.Gutenberg.org dataset, its the same dataset used in "Searching with Style: Authorship Attribution in Classic Literature" by Zhao and Zobel (2007),

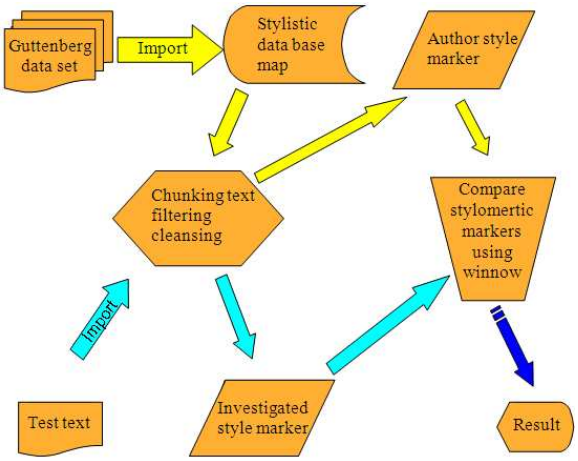

Fig. 1: Proposed methodology design

to further explore the properties of Attribution identification (AA) methods, we apply them to a corpus of novels extracted from the Gutenberg project. While not a large corpus by text collection standards, it is more substantial than the collections used in most previous work for AA and contains a substantial cross-section of 19th-century English literature as well as other work. Using this collection with no poetry volumes collected neither dictionaries, nor languages other than English, Individual short stories, However we did keep both plays and novels.

The exact data set has been used in the "Computational Stylomertic Approach Based on Frequent Word and Frequent Pair in the Text Mining Authorship Attribution" (Mustafa *et al*., 2009) which will give us a perfect comparison platform between the methods adopted.

Stylistics database map and test text: Represented in the experiment by The stylistic database that is designed and the database tables and relations prepared to import data into our system, then we can deal with the data as structured, able to mine and analyze and prepare to the chunking and filtering and cleansing steps that are familiar in data mining, reminding that the attributes measures in the first and second experiments depend on maximum item set for the frequency, while our last experiment depends on the CV totally dropping out the maximum item.

Chunking text, filtering and cleansing: We start to analyze the data set that we collected preparing for learning algorithm procedures, the procedures are for teaching the proposed system to act like an "expert" in checking the text styles of authors, cleansing and filtering are common preprocessing procedures to get the proper data the can be clearly analyzed without any distortion or noise, these terms here are represented by, multi spaces between words (since statement is a collection of sequential words, each word is distinguished by a single space before and after) multi punctuating similar signs, titles of sections and parts. After cleansing and filtering comes chunking to tokens, that shreds the text into table of author stylistic mark or classifier and their frequencies, the marks or classifier that we are interested in is the frequent word and frequent pair that the author is addicted to use frequently in all off his texts.

The main contribution here is dropping down the usage of the frequent (word or pair) and going towards selecting attributes by there CV result, meaning that we wont select any more attributes depending on its frequency in usage, but the selection will be made depending on the CV result that will show the stability of there usage no matter how frequent the attributes are, this contribution will involve deferent kinds of AA to be selected, high, medium, or low frequent.

The only frequency assumption will be assumed is that the AA selected should be more than 30 frequency (30 frequencies out of 8 books is less than 4 time appearances in each book as an average which means that the low frequent AA will be selected also) to support the principle of the effective sample size results in normal distribution, selecting attributes less than this size gives less satisfactory results (James and Muth, 2007).

Author style marker and investigated style marker: Both learning and test data that were cleansed and chunked and analyzed are now stated as AA classifiers (clustering), one as an expert opinion and the other as the tested under investigation text, for further winnow algorithm comparison.

Compare Stylometric markers using winnow: Winnow algorithm here is used as a classification step, after getting the comparison done by:

- Classical attribution selection for 1st and 2nd experiment by putting the frequent pair for the map facing the corpuses under test, with threshold = 15%
- The CV statistical measure for 3rd part of experiment done by putting the highest CV degrees for each pair in the map facing the corpuses under test, with threshold range (50-70)

The algorithm used here for computational approach is:

- Post processing filtering the tokens for the 15% threshold
- Sorting the tokens descending depending on their frequency
- Searching for each map token in the four test books and putting the corresponding token frequency for each test author
- Generate the comparison table

Results will show in two ways:

- Histogram comparison as it shows next
- Pearson correlation coefficient

By using the computational Stylometric measure r (Pearson correlation) that is used to find each classifiers weight (association role step) to give the final automated result as described previously by using the Pearson correlation coefficient, giving two variables involved, x presents the Stylometric twain learning map and y is the corresponding test map for each three authors, that will give us a reasonable decision support tool for the judgment.

Decision making using the results: For a straight ahead accurate decision making, we compute the winnow result from the computational Pearson parameters counted previously and the highest winnow result belongs to the best attribute set investigated. Given (+1) for each x that represent a positive example and (-1) for each negative example, positive example is the value that we multiply by the Pearson correlation calculated for an author that is appointed previously as the right author for the script, while negative parameter represents the false author:

$$\text{Winnow result} = w_1x_1+w_2x_2+\ldots+w_nx_n$$

RESULTS

Programs and codes needed to distinguish our methodology and algorithm were written in visual FoxPro 7.0 language that is very effective and flexible as a database engine and text mining tool.

To get results, we implement empirically the proposed methodology in three levels to show the accuracy improvement for predicting the right author for the scripts under investigation, as we described previously, these levels are:

- Frequent word results for large item set 250 (most common used which equals 10% from the maximal item set in most cases) (Fig. 2):

Pearson (Twain map, London) = 0.9669896
Pearson (Twain map, shakes) = 0.8785840
Pearson (Twain map, twain 1) = 0.9801125
Pearson (Twain map, twain 2) = 0.9919130

By implementing the Eq. 4 as we showed previously:

Winnow1 = (-1) (0.9669896)+(-1) (0.8785840)+
(1) (0.9801125)+(1) (0.9919130) Winnow1
= 0.1264519

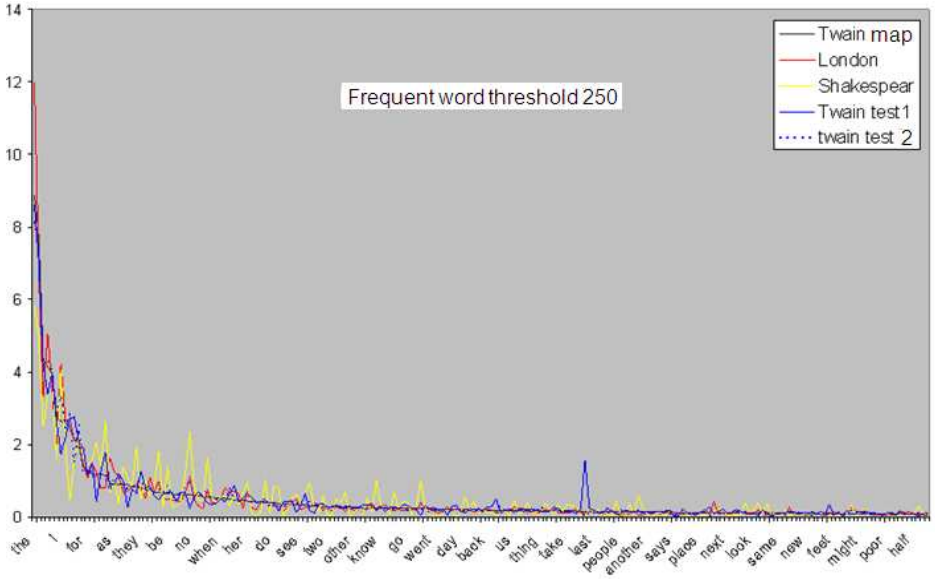

Fig. 2: Frequent word Stylometric map

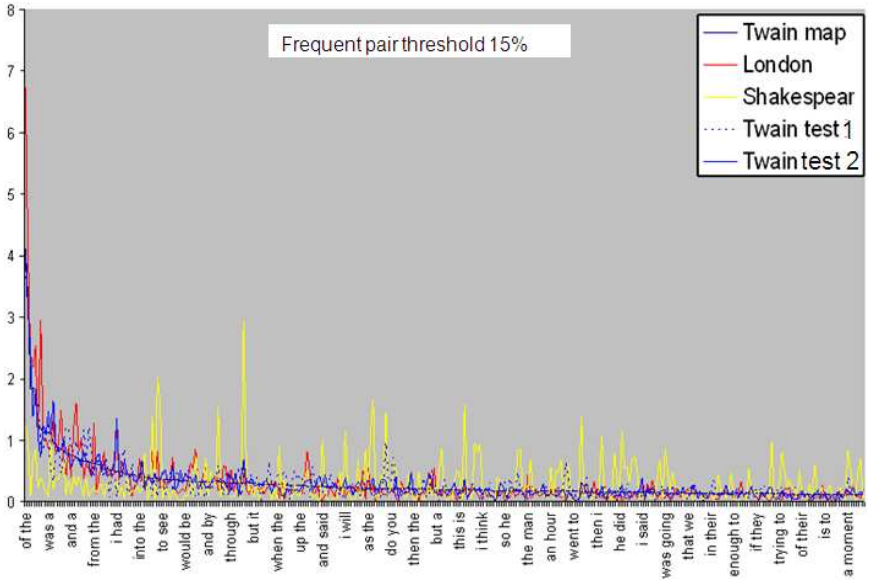

Fig. 3: Frequent pair Stylometric map

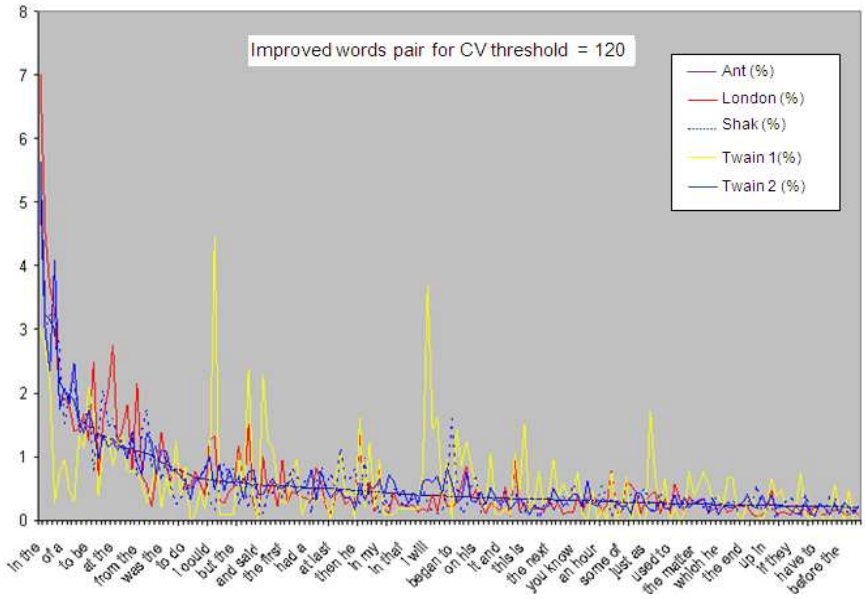

Fig. 4: Improved frequent pair Stylometric map

- Frequent pair results for large item set (Fig. 3):

Pearson (Twain map, London) = 0.9291140 (negative mistake)
Pearson (Twain map, shakes) = 0.1773040
Pearson (Twain map, twain 1) = 0.9068599 (positive mistake)
Pearson (Twain map, twain 2) = 0.9496353

By implementing the Eq. 4 as we showed previously:

Winnow2 = (-1)(0.9291140)+(-1)(0.1773040)+ (1)(0.9068599)+(1)(0.9496353)
Winnow2 = 0.7500772

- Improved CV pair result for large item set with Threshold = 60% (174 attributes) (Fig. 4).

Pearson (Twain map, London) = 0.9447206
Pearson (Twain map, shakes) = 0.6314261
Pearson (Twain map, twain 1) = 0.9612207
Pearson (Twain map, twain 2) = 0.9607192

By implementing the Eq. 4 as we showed previously:

Winnow3 = (-1)0.9447206+(-1)0.6314261+0.9612207
+0.9607192 Winnow3
= 0.3457932

DISCUSSION

There are two conditions to compare between the winnow results we got from the implementation:

- A sharp condition is that, there should be no mistake or the least positive and negative mistakes appearance in the AA detecting
- Highest winnow degree is chosen among the winnows we got from the experiment representing the highest expectation given for predicting the author

Depending on the previous criteria, we note that the frequent pair experiment (winnow1) has no predicting mistakes, at the contrary; the words pair (winnow2) has two mistakes as we see in the experiment results.

Combining the winnow algorithm with the computational approach, achieved by using the CV statistical tool as a new measure with new threshold for attribute selecting, the frequent pair (winnow2) result improved from 50% error to be 0% in the improved frequent pair (winnow3) with a clear higher score result compared with the frequent word attribute.

CONCLUSION

- The frequent word experiment (winnow1) has no predicting mistakes, on the contrary; the words pair (winnow2) has two mistakes as we see in experiment results showing that the classical frequent word performs better in the classical algorithm
- The frequent pair (winnow2) result improved from 50% error (classical measure) to 0% in the improved CV pair (winnow3) with a clear higher score result compared with the frequent word attribute showing the accuracy improvement in the new CV proposed algorithm
- The new CV algorithm results improvement may lead to several new attributes usage that gave unsatisfying results before that might improve the direction for solving some hard cases couldn't be solved till now
- Selecting the right threshold for the new algorithm needs to be tested empirically
- Using a specific threshold for each author in his own Stylometric map as a new AA is something worthwhile by researchers
- Expanding the data set to certify the accuracy improvement in this study should be under consideration

REFERENCES

Argamon, S., M. Saric and S.S. Stien, 2003. Style mining of electronic messages for multiple authorship discrimination: First results. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 24-27, ACM Press, Washington DC., USA., pp: 475-480. DOI: 10.1145/956750.956805

Argamon, S. and S. Levitan, 2005. Measuring the usefulness of function words for authorship attribution. Proceeding of the Joint Conference on Association for Literary and Linguistic Computing/Association Computer Humanities, June 15-18, University of Victoria, Canada, pp: 1-3. http://lingcog.iit.edu/doc/paper_162_argamon.pdf

Argamon, L. and L. Shlomo, 2006. fixing the federalist: Correcting results and evaluating editions for automated attribution. Proceedings of Association Computer Humanities in Digital Humanities, July 5-9, Sorbonne. Paris, pp: 323-328. DOI: 10.1.1.73.2352

Croft, D.J., 1981. Book of Mormon word prints reexamined. Sun Stone Publish., 6: 15-22. http://lds-mormon.com/wordprin.shtml

Hearst, M.A., 2003. Untangling text data mining. Proceeding of the 37th Annual Meeting on Association for Computational Linguistics, June 20-26, University of Maryland, College Park, Maryland, pp: 3-10. DOI: 10.3115/1034678.1034679

James, E. and D. Muth, 2007. Basic statistics and pharmaceutical statistical applications. J. Int. Biometr. Soc., 63: 630. DOI: 10.1111/J.1541-0420.2006.00787_9.X

Krippendorf, K.H., 2003. Content Analysis-An Introduction to its Methodology. 2nd Edn., Sage Publications Inc., USA., ISBN: 13: 978-0761915454, pp: 440.

Malyutov, M.B., 2006. Authorship attribution of texts: A review. Lecture Notes Comput. Sci., 4123: 362-380. DOI: 10.1007/11889342_20

Mustafa, T.K., M. Norwati, A. Masrah and S. Nasir, 2009. Computational stylomertic approach based on frequent word and frequent pair in the text mining authorship attribution. Int. J. Comput. Sci. Net. Secur., 9: 262-269. http://paper.ijcsns.org/07_book/200903/20090335.pdf

Stamatatos, E., N. Fakotakis and G. Kokkinakis, 1999. Automatic authorship attribution. Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics, June 8-12, Association for Computational Linguistics, Morristown, New Jersey, USA., pp: 158. DOI: 10.3115/977035.977057

Stien, S. and S. Argamon, 2007. Interpreting Burrows's delta: Geometric and probabilistic foundations. Literary Linguist. Comput., 23: 2-131. DOI: 10.1093/LLC/FQN003

Tareef, K.M., 2007. Measuring the stylometric authorship attributes in text mining. Proceeding of the 9th International Conference on Intelligent Technologies, Oct. 7-9, Samui Thailand, pp: 159-163. http://www.intech.scitech.au.edu/register/Intech08/schedule.asp

Zhang, T., F. Damerau and D. Johnson, 2002. Text chunking using regularized winnow. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, July 6-11, Association for Computational Linguistics, Morristown, New Jersey, USA., pp: 539-546. DOI: 10.3115/1073012.1073081

Zhao, Y. and J. Zobel, 2007. Search with style: Authorship attribution in classic literature. Proceedings of the 13th Australasian Conference on Computer Science Volume 62, Jan. 30-Feb. 2, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, pp: 59-68. http://portal.acm.org/citation.cfm?id=1273757