Computational Stylometic Approach Based on Frequent Word and Frequent Pair in the Text Mining Authorship Attribution

Tareef Kamil Mustafa[1], Norwati Mustapha[2], Masrah Azrifah Azmi Murad[2] and Md. Nasir Sulaiman[2]

Faculty of Computer Science and Information Technology
University Putra Malaysia
Serdang, Selangor, Malaysia

Summary
Stylometric Authorship attribution is one of the new approaches in the text mining field that has been showing recently because of its delicateness. This approach is concerned about analyzing texts, e.g. Novels and plays that famous authors wrote, trying to measure the author 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. To achieve that, this paper discusses several algorithms which are used frequently and skipping the one time, ad-hoc adventures in this field. This paper is also opens the way for future works to merge and improve these techniques by showing experimentally the accuracy level of using both frequent words and frequent word pair depending on the computational approach.

Key words:
Authorship attribution, computational stylometric, text mining.

1. Introduction

Text mining is a diverted subject from the well-known field “Data mining”. As for authorship investigation that using the writing style of the author is a sub field of text mining called “Authorship attribution” or “Stylometric Text mining”. All these subjects need to be defined to get the picture well clarified.

A. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what we're familiar with in web search. In searching, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently isn't relevant to your needs in order to find the relevant information. Unlike what’s In text mining, the goal is to discover heretofore unknown information, some thing that no one yet knows and so could not have yet written down [3][5].

B. Authorship attribution (AA***)*** is the process of attempting to identify the likely authorship of a given document, given a collection of documents whose authorship is known. Most of the methods described in the research literature consist of two components, an indexing mechanism and a comparison mechanism. The indexer converts each document to a set of tokens or markers whose properties are assumed to be characteristic in some way of a particular author. The comparator uses these markers to assign an author to un-attributed documents [12].

C. 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 to the abilities of the user.

D. Frequent words is the most common words and most frequently used by authors in their texts. For example, the word "the" was represented in a numerical vector as the number of times it occurs in the text divided by the number of words in the text [1].

E. Word collocation [2] is defined as a certain pair of words occurring within a given threshold distance of each other (such as "is" and "certain" appearing within 5 words of each other in this sentence).

F. In the literature of stylistic analysis, we find many references claiming that for a given author there are habits (of style) and they are not affected by (1) passage of time, (2) change of subject matter (3) literary form. They are thus stable within an author's writing, but they have been found to vary from one author to another [4].

The importance of Dactyloscopy (fingerprint) and DNA profiling in forensic and security applications is universally recognized after successful testing of their resolution power and standardization of analyzing tools. Much less popular so far is a similar approach to the attribution of disputed texts based on statistical study of patterns appearing in texts written by professional writers. The best tests and their power are yet to be estimated both theoretically and by intensive statistical examination of Stylometric differences [6].

Manuscript received March 5, 2009
Manuscript revised March 20, 2009

Most of the stylistic analysis has forgotten that “Style” means how the author combines and arranges statements consciously and unconsciously using words to create statements. Therefore, statements become our concerns instead of words and for the starting towards that, we do the analysis in this paper based on the computational stylistic method. Our intention on doing the analysis is to prove that style feature of professionals can be discriminated as well as fingerprints of different persons using authorship attributes. Improving the accuracy of the discrimination is the goal of each method in the authorship attribution problem.

In section 2, we will describe the present works with the different algorithms and techniques used to solve the AA problems. Then, the methodology used and the details implementation of the computational approach in this field is discussed in Section 3 and Section 4 respectively. Experimental results by using frequent word authorship attributes as well as word pair attributes are reported in Section 5 and Section 6 gives the conclusion about the analysis that has been done together with our views on future works.

2. Preview

Even all the methods and algorithms that have been stated in this work are indexed according to their first testing appearance date but they are still working together. This is means that no method took the place of a previous one, or replaced by new ones. All the scientist and researchers choose a specific method, and continued improving and testing until now days [10].

The methods that are described in this section are the most frequently used and developed and tested. We are avoiding the methods that we call “one time shot” means that the methods which are more ad hoc adventures and never been tested. We are also avoiding techniques that gave unsatisfied results that would not assist on achieving our goal.

A. Content analysis is “one of the most important research techniques in the social sciences. It seeks to understand data not as a collection of physical events but as symbolic phenomena and to approach their analysis unobtrusively [6]. Methods in the natural sciences do not need to be concerned with meanings, references, consequences, and intentions. Methods in social research that derive from these "hard" disciplines manage to ignore these phenomena for convenience. Yet, nobody doubts the significance of symbols in society. This method is one of the pioneer ones, but it is still the technique that was more on listing some descriptive statistical measures and making the decisions by pure self opinions.

B. Computational stylistic approach method is based on the computational analysis of the input text using a text-processing tool. Besides the style markers relevant to the output of this tool it also use analysis-dependent style markers, that is, measures that represent the way in which the text has been processed [8]. Effort is required regarding the selection of the most appropriate set of words that best distinguish a given set of authors, Moreover, the statistical methodology of multivariate linear multiple regressions was applied to the training corpus. Multiple regressions provide predicting values of a group of response (dependent) variables from a collection of predictor (independent) variable values. The response is expressed as a linear combination of the predictor variables, namely:

$$y_i = b_o + z_l b_{li} + z_2 b_{2i} + ... + z_r b_{ri} + e_i$$

where y, is the response for the *i*-th author, z1,z2,..and zr are the predictor variables (i.e., in our case r=22), bo, bl,, b2,...,br are the unknown coefficients, and e, is the random error. During the training procedure the unknown coefficients for each author are determined using binary values for the response variable (i.e., 1 for the texts written by the author, 0 for the others). Thus, as much greater the response variable of a certain author, the more likely to be the author of the text. Some statistics measuring the degree to which the regression functions fit the training data. Note that R is the coefficient of determination that defined as follows:

$$R^2 = \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2}$$

C. Exponentiated Gradient learning algorithm (EG) [3] considers the use of computational stylistics for performing authorship attribution of electronic messages, addressing categorization problems with as many as 20 different classes (authors), effective stylistic characterization of text is potentially useful for a variety of tasks, as language style contains clues regarding the authorship, purpose, and mood of the text. All of these would be useful adjuncts to information retrieval or knowledge-management tasks. The problem focus in this work is to determine the author of an anonymous message, based on the message text. They compare results using several multi-class generalizations of the (EG) to outperform other popular learning methods for stylistic classification [3]. EG is a very successful algorithm, but with electronic letters. In addition, it is used on small corpus texts and it is not tested on large amount of texts such as novels.

D. Winnow regularized algorithm is the algorithm to enhanced linguistic features. In this paper, a description of text chunking system is using regularized Winnow. Since regularized Winnow is robust to irrelevant features, we can construct a very high dimensional feature space and let the algorithm pick up the important ones. The article shows that state of the art performance can be achieved by using this approach [11]. Furthermore, the method proposed is more computationally efficient than all other systems reported in the literature.

E. Modeling of long canons as Markov chains with some order composed of English letters and auxiliary symbols is a new approach suggested in [7]. Given a non-attributed text T and a collection of firmly attributed (to author k) canons T(k) of approximately the same length for training the Markov model of, say, order 1, with transition probabilities P(k, i, j) between symbols i and j, k=1,..., M, the log likelihood of T being written by the k-th author is :

$$\sum \log(p(k,i,j))N(i,j) + \log \pi_k(x(1))$$

where the sum is over all i and j, N(i, j) is the frequency of i followed by j, _k denotes the stationary probability of the k-th Markov chain, and x(1) is the first symbol in T. Second order Markov chain modeling admits similar expressions for the likelihood. The author with maximal likelihood is chosen, which is practically equivalent to minimizing the cross entropy of empirical and fitted Markov distributions and to minimizing the prediction error probability of a next symbol given the preceding text.

F. Burrows's Delta Method [9] is a simple and effective method. Its goal is to automatically determine, based on a set of known training documents labeled by their authors, who is the most likely author for an unlabeled test document. The Delta method uses the most frequent words in the training corpus as the features to make these judgments. The Delta is defined as mean of the absolute differences between the z-scores for a set of word variables in a given text-group and the z-scores for the same set of word-variables in a target text. The Delta of the test document is computed with respect to each of the training documents, and that author whose training document has minimal Delta with the test document is chosen for attribution. The Delta between these documents can be reformulated as below:

$$\sum_{i=1}^{n} |z(X_i) - z(Y_i)| = \sum_{i=1}^{n} \left| \frac{X_i - \mu_i}{\sigma_i} - \frac{Y_i - \mu_i}{\sigma_i} \right|$$

$$= \sum_{i=1}^{n} \left| \frac{(X_i - \mu_i) - (Y_i - \mu_i)}{\sigma_i} \right|$$

$$= \sum_{i=1}^{n} \left| \frac{X_i - Y_i}{\sigma_i} \right|$$

3. Methodology

Our work is based on computational stylistics approach for learning and testing techniques. There are two types of test used for the authentication attributes (AA) :

- Frequent words with deferent thresholds.
- Frequent word pair with deferent thresholds.

After building the style map for a specific proposed author (Mark Twain), we compare and analyze the rewrite rules as they appear. Both of high-frequent and medium-frequent rewrite rules give accurate results. The parameters that will be used for comparing the style proposed map with each test text is the linear regression measure represented by the Pearson correlation coefficient that has been proposed at the computational stylistics approach.

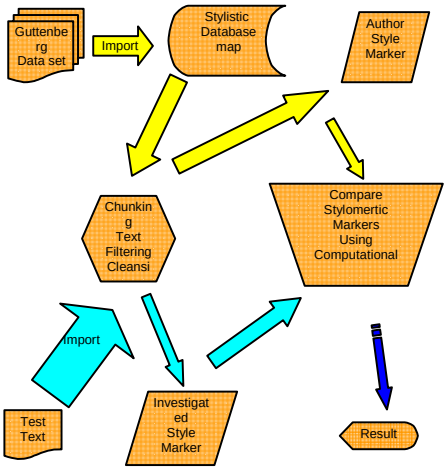

Fig 1 The Proposed Methodology

In other words, take the result of learning approach for the suspected authors as a comparing measure to compare it with the result of the tested text under investigation to give the automated decision. The results will be shown in two ways; first is visualization using histograms and second is the correlation coefficient ratio scale with three thresholds for experimental reasons.

We are using Visual Fox Pro 7.0 language to implement our methodology. It is very effective and flexible as a database programming language and text mining tool because of the SQL engine that's been embedded with.

4. Implementation of Computational Stylometric Approach

The Details of each step in our methodology is shown in Fig 1 which is based on computational Stylometric approach, as described below:

A. Data set from the web site www.Gutenberg.org dataset: It is the same dataset that has been used in "Searching with Style: Authorship Attribution in Classic Literature" by Ying Zhao Justin Zobel [12]. To further explore the properties of Attribution Authorship (AA) identification 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, we gathered 8 books from a famous author called Mark Twain and two other books from Shakespeare

and Jack London for testing with two more from Mark Twain wasn't used in the learning path, so we can compare the results. In selecting the books, we avoided choices that we felt were inconsistent with the aims of our experiments. We did not collect volumes of poetry, dictionaries, or text in languages other than English. Individual short stories were avoided as well (see Table 1).

Table 1: The Dataset

No.	Author	Book Title	Size	Task
1	Mark Twain	What Is Man	532KB	Learn
2	Mark Twain	The Adventures of Huckleberry Finn	563KB	Learn
3	Mark Twain	The Prince and The Pauper	374KB	Learn
4	Mark Twain	Roughing It	922KB	Learn
5	Mark Twain	HOW TO TELL A STORY	40KB	Learn
6	Mark Twain	A Horse's Tale	107KB	Learn
7	Mark Twain	The Stolen White Elephant	60Kb	Learn
8	Mark Twain	A Connecticut Yankee in King Arthur's Court	661KB	Test
9	Shakespeare	THE TRAGEDY OF ANTONY AND CLEOPATRA	167KB	Test
10	Jack London	The Mutiny of the Ellsinore	627KB	Test

B. Stylistics Database Map and test text: The database that is designed and files and relations that have been prepared to import data into our system. We can deal with the data as structured, able to mine and analyze by preparing it for chunking and filtering and cleansing steps that are familiar in data mining. Fig 2 is the overlook on the Stylistic Database map.

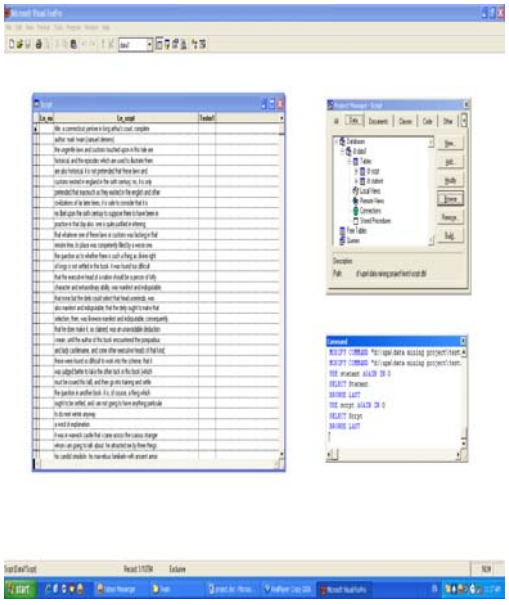

Fig 2 Overlook on the Stylistic Database map

C. Cleansing and Chunking Text: We start to analyze the data set that we collected to prepare it 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 procedures to get the proper data that can be clearly analyzed nearly without any distortion or noise, these terms are represented by multi spaces between words. Since frequent pair 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 part etc. using the SQL commands.

D. Chunking: After cleansing and filtering then chunking is needed to shreds the text into table of author stylistic mark or classifier and find their frequencies. The marks or classifier that we are interested in is the "frequent pair", that is the double words statements that this author is addicted to use frequently in all of his texts. This procedure is enrolled together for the learning and the testing data for comparison purposes. Instead of simply chunking the corpus without counting the frequency, we added 'Group' SQL command at the end to perform the grouping and counting which giving faster.

E. 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. Learning as an expert opinion, and the other one as the tested under investigation text. For algorithm comparison purposes, we are choosing three different thresholds once at a time to see the reflections on the final results; 1000,500 and 250 for the frequent word attribute from the Stylometric map of Mark Twain that contains 472546 tuple and keeping about 42, 117 and 202 attributes from each threshold respectively.

As mentioned before, results will be shown in two ways (see Fig 3).

a. Visual histograms: This is clearer, because it shows the result as a chart.
b. Pearson correlation: This is more specific because it is a ratio result.

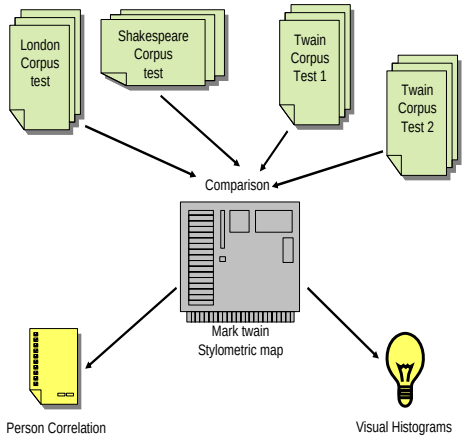

Fig 3 Pearson Correlation and Visual Histogram to display the results

F. Compare Stylometric Markers via computational approach. This research presets the empirical experiment into two parts:

1. Frequent pair attributes experiment.
2. Frequent word attributes experiment.

Using different threshold for each experiment to compare and choose the best result obtained.

5. Experimental Results

A. Frequent pair attributes experiment.
The comparison is done by putting the frequent word table for the map facing the corpuses under test, with threshold = 1000.

The algorithm used here for computational approach is:

- Filtering the tokens for the 1000 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.
- At the end we generate the comparison table, showing the results in two ways.

1) Histogram comparison as it shows next.
2) Pearson correlation for each test.

The histogram in Fig 4(a), 4(b), 4(c) show the main Twain Stylometric map curve (black) for frequent pair tokens, the chart shows that Shakespeare curve (yellow) is far from the collation, while more of London curve (red) is more near and the twin test corpuses are the nearest collation for the learning twain map, which means that the test result is correct, however it is still hard to recognize.

B. Frequent pair attributes experiment

Histograms in Fig 5(a), 5(b), 5(c) are for frequent word results. They show better performance in recognizing the author compared with word pair. Showing that the threshold used here is numerical and not percentage, the reason is that in frequent word, unlike frequent pair, the threshold where concentrated on 2 words taking more than 10% of the whole frequent words, so even taking for 10% or 15%, it won't make a reasonable amount of attributes that can help to make a decision about the author. Due to this reason, it leads for taking a ratio threshold between 250 and 1000. It's preferred to take the result from three thresholds, min, max and mid to make some judgment by comparing these results.

The visualized results are always uncertain but it still gives a fast opinion if the result was clear enough, As much as there are distortion and diversions and sharp angles between the frequent map curve and the compared curve, it means that the system is suggesting these authors are not the same. As much as the curves are collaborating and smoothly going side by side, it means that the system decision is positive which is the authors are same or predicted as the most expected

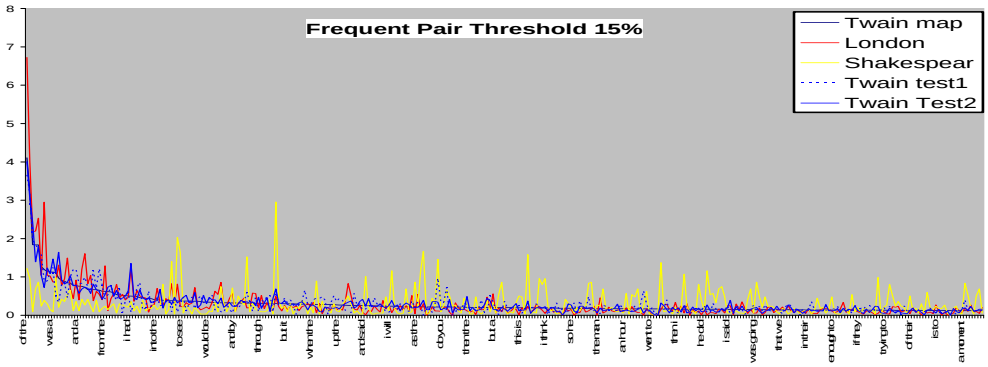

Fig 4(a) Result of Frequent Pair with Threshold 15%

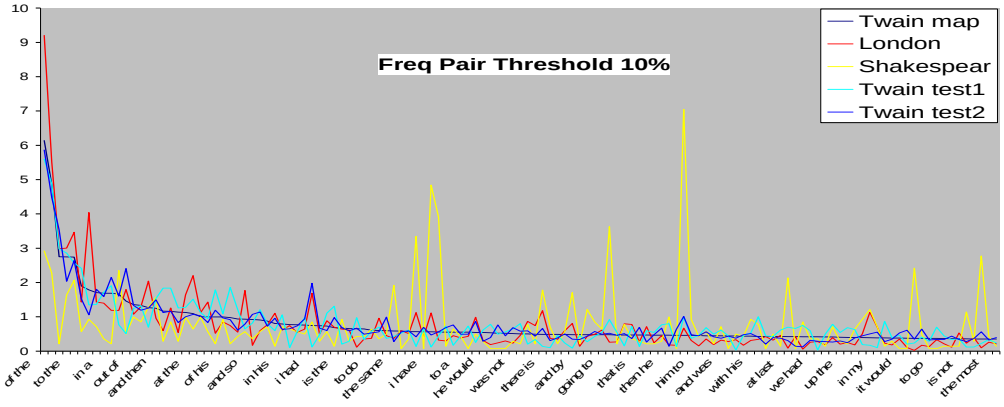

Fig 4(b) Result of Frequent Pair with Threshold 10%

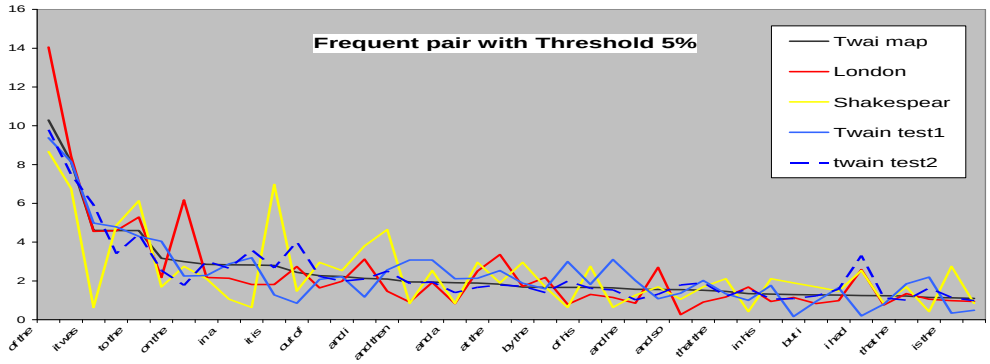

Fig 4(c) Result of Frequent Pair with Threshold 5%

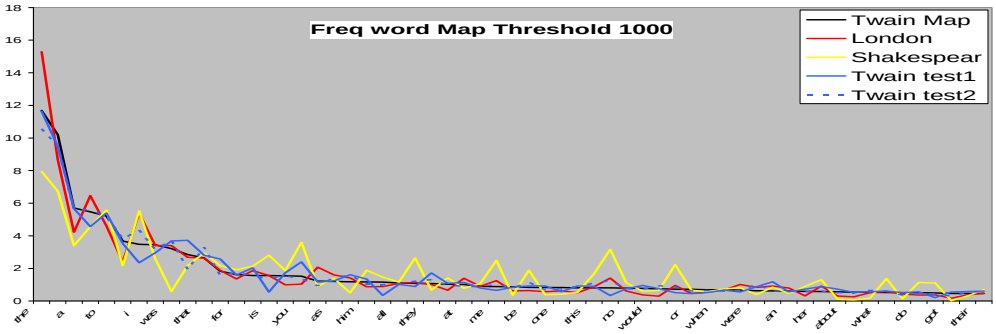

Fig 5(a) Result of Frequent Word Pair with Threshold 1000

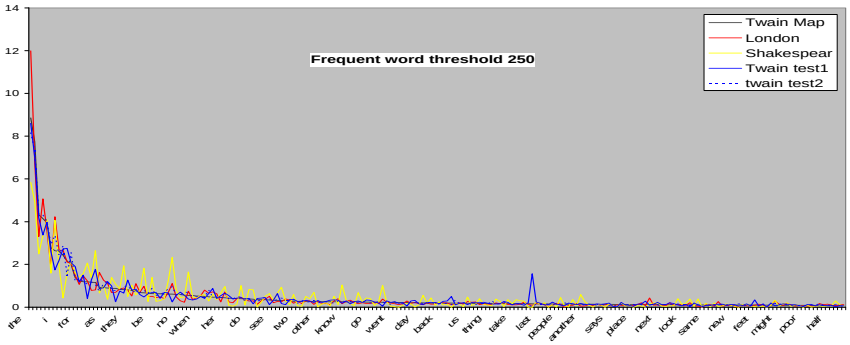

Fig 5(b) Result of Frequent Word Pair with Threshold 250

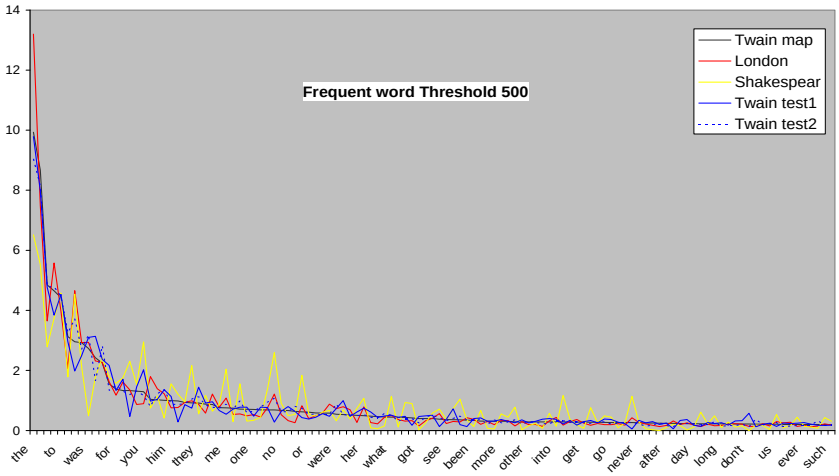

Fig 5(c) Result of Frequent Word Pair with Threshold 500

The second way of showing the result is Pearson correlation (measure r), It is used to find each classifiers weight to give the final automated result as described in [7]. By using the Pearson correlation coefficient below and by giving two variables which are x presents the Stylometric twain learning map and y is the corresponding test map for each four test corpuses, the result as shown in Table 2:

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

Table 2: Pearson Correlation Results

	Frequent Word Attributes			Frequent Pair Attributes		
	Ratio Threshold			Percentage Threshold		
Authors name	>1000	>500	>250	5%	10%	15%
London	0.9593833	0.9648779	0.9669896	0.9281808	0.9293098	0.9291140
Shakespeare	0.8552423	0.8690814	0.8785840	0.7220018	0.2298419	0.1773040
Twain test 1	0.9828431	0.9843868	0.9801125	0.9186169	0.9173343	0.9068599
Twain test 2	0.9905909	0.9915572	0.9919130	0.9429911	0.9502439	0.9496353
# of Attributes	42	117	202	42	129	331

Decision making here is depend on choosing the best column or more from the two experiments computed; frequent word and frequent pair attributes. The best column here is the one that have less diversion in deciding which one of the test corpuses really belong to the Stylometric map built in the learning path. In other words, if the result in any cell of the previous table is higher than other cell in the same column, that means the test corpus in the cell does belong to the author under system learning, and here its Mark Twain.

There are two types of error appears in this situation.

- Positive error: the system gives low result in testing a corpus does belong to the author.

- Negative error: the system gives high result for a corpus that does not belong to the author under consideration.

We are using different levels of threshold to compare and chose the most optimal one for such test. Table 2 shows that the result of frequent word attributes gives better decision than the frequent word pair because of one negative error which is the column for (London, frequent pair) compared with (Twain test 1, frequent pair) in Table 2.

For all threshold levels showing that the system supports and favorites “London” against “Twain” by giving higher result for the wrong decision. The unsatisfying result here shows that the computational approach is confused because of the noise affecting more on the frequent pair attribute that makes both authors have the same extensively high result.

The best result we got in frequent word is for the >1000 threshold (minimum attributes) with nearly no error in the decision supporting results.

It should be mentioned here that in selecting the attributes of the Twain Stylometric map, manually we dropped three attributes from the frequent pair list. The pairs are (the king, the king's, and says:) and obviously these pairs were belonging all to one novel “A Connecticut Yankee in King Arthur's Court” but they where so frequent which they can changed the whole Stylometric map. By dropping them out, the results became much more convenient.

6. Conclusion

Based on the experimental results given by computational Stylometric approach, we can conclude that Stylometric features of different professional authors can be discriminated nearly as well as fingerprints of different persons using authorship attributes.

The results of frequent word experiment were more satisfying than the frequent pair attributes. However, the computational stylistic method depends more on statements than words to differentiate between authors’ styles. In addition, noisy results appear while trying to get a minimized threshold which shows that higher threshold with large dataset gives better results even the attributes were less frequent.

Therefore, it needs some improving techniques to overcome its noisy effect are needed.

References

[1] Argamon L., Shlomo L. (2006), “Fixing the Federalist: Correcting Results and Evaluating Editions for Automated Attribution”, Abstract Submitted to Association Computer Humanities, Sorbonne, Paris.

[2] Argamon, S. & Levitan, S. (2005). Measuring the Usefulness of Function Words for Authorship Attribution, Association for Literary and Linguistic Computing/ Association Computer Humanities, joint conference ACH/ALLC Conference 2005 June 15 - June 18, 2005

[3] Argamon, Saric, Stien (2003), “Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results”, Dept. of Computer Science Illinois Institute of Technology Chicago, Proceedings of ACM Conference on Knowledge Discovery and Data Mining.

[4] Croft D.J (1981), “Book of Mormon "Word prints" Reexamined”, Sun stone publishing’s.

[5] Hearst M. A (2003), “What is Text Mining”, ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland School of Information Management & Systems.

[6] Krippendorf K. (2003), “CONTENT ANALYSIS – An introduction to its Methodology”, Sage Publications, Inc; 2nd edition.

[7] Malyutov M.B. (2006), “Authorship attribution of texts: a Review”, Springer Berlin / Heidelberg, LNSC series.

[8] Stamatatos E., Fakotakis N. and Kokkinakis G. (1999), “Automatic Authorship Attribution”, Ninth Conference of the European Chapter of the Association for Computational Linguistics, university of Bergen, Norway.

[9] Stien S, Argamon S. (2006) “A Mathematical Explanation of Burrows’s Delta”, Linguistic Cognition Laboratory Department, website http://lingcog.iit.edu/

[10] Tareef kamil, Norwati Mustapha (2007) , , The Ninth International Conference on Intelligent Technologies proceedings, Samui Thailand, October 7-9, 2008

[11] Zhang T, Damerau F. and Johnson D. (2002), “Text Chunking using Regularized Winnow”, CiteSeer website, digital library NEC Research Institute, http://citeseerx.ist.psu.edu.

[12] Zaho Y., Zobel J. (2006), “Search with Style: Authorship Attribution in Classic Literature”, The Thirtieth Australasian Computer Science Conference, Ballarat, Jan. 2007.

Mr. Tareef Kamil Mustafa completed his Master degree at Technology university- IRAQ – Baghdad in 2005. At present he is a PhD candidate at University Putra Malaysia, Serdang, Selangor, Malaysia. He is a Lecturer in Baghdad university – Science College – Computer department and was involved in several database Projects as a Database manager and System Analyses.