ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Applicability of Relevant Authorship Attribution Technique in Malayalam

Shabeeb PK

P.G. Student, Department of IT, School of Engineering, CUSAT, Kochi, Kerala, India

ABSTRACT:Authorship attribution has been around arguably for a long time, from the federalist papers to the 16th book of Oz it has been successfully used to find out authors of disputed texts in question. With the advent of computers authorship attribution techniques got faster and smarter but it still has its limitations. Only countable courts accept the results of these techniques accountable before law. But the major challenge of authorship attribution is the language barrier, there has not been any serious attempts made in to implement authorship attribution in to languages other than English and a few others. This paper tries to identify and adopt a widely used authorship attribution technique implemented in English to Malayalam, comparatively a complex language and infer its accuracy and challenges when implementing the same. With this work it is expected that the field of authorship attribution finds a whole new research groundto go forward.

KEYWORDS: Authorship Analysis, Attribution, Stylometry, Statistical text processing, Natural Language Processing

I. INTRODUCTION

Authorship attribution is the method/technique of ascertaining the author of a particular transcript in question by analyzing the previous works of the same author.The research in authorship attribution is very extensive innature, but it still lacks implementation of the technique incomplex language such as agglutinative languages. Malayalamis an agglutinative language which is very different frominflectional language such as English/European languages,hence the technique such as authorship attribution is notimplemented in Malayalam. The major challenge of authorshipattribution is the language barrier, there has not been anyserious attempts made in to implement authorship attribution into languages other than English and a few others. Thispaper tries to identify and adopt a widely used authorshipattribution technique implemented in English to Malayalam,comparatively a complex language and infer its accuracy andchallenges when implementing the same. With this work it isexpected that the field of authorship attribution finds a wholenew research ground to go forward.

II. PROBLEM DEFINITION

Although authorship analysis is available in many languages, there is still a lack of authorship analysis implementation in the languages of Indian subcontinent, in this paper we are addressing the issue of adopting the extensively employed authorship analysis technique used in European languages such as English, French, Dutch etc.: in to South-Indian language Malayalam an agglutinative language. Agglutinative tends to express concepts words consisting of many elements rather than by inflection or by using isolated elements (Some examples are: Hungarian, Turkish, Korean, Malayalam). Along with adopting the technique we should also be able to identify the feasibility and challenges of the technique with respect to the language.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

III. SCOPE

The scope of authorship attribution is large. Today authorship attribution techniques are used in an array of applications and the prospects of its use future is also high. Some of the notable

1) Cyber forensics: The use of authorship attribution is immense; authorship attribution can be used to identify the author of anonymous scriptures circulating in the internet. With the immense utilization of internet as an arena for preaching and recruiting terrorists by major terror groups, authorship attribution helps the officials to zero in or get a picture of the people behind the scriptures.

2) In source code fingerprinting: Authorship analysis can also be used in identifying the author of different source codes (codes of computer programs) that are usually anonymous programmers. It will be very much helpful in aiding investigations involving computer hacking by using computer programs. It will also be helpful to find a resolution in source code disputes between different corporations.

3) Finding authorship of disputed literature: One of the classical example of authorship attribution is that of finding the authorship of disputed literature. Recently famed author JK Rowling was discovered to be the author of a crime novel named “Cuckoos Calling¨ which was written in a pseudo-name by the same.

4) Analyzing Historical artifacts: The research on authorship attribution was largely influenced by analyzing and attributing the authorship of historical artifacts that are generally questioned or unknown. The works of Mostfeller and Wallace on Federalist Papers and determining the author of 16th book of Oz are the prominent examples.

5) Assisting hate-crime investigations: Hate Crimes are increasing on a daily basis with the aid of social media and internet. In most of the hate-crimes there will be written transcripts written by radicals to preach the issues related to the crime to a much common group of people, most often these transcripts arise from few of the brain-centers in the group and thus the technique of authorship attribution allows us to pin-point these people or get an idea of their number thus assisting in these kind of investigations.

6) Analysis of notes like Suicide notes: The infiltration of electronic media has changed many aspects of our life, nowadays we write most of our notes in electronic way. Sometimes they include personal notes such as diary and suicide notes, authenticity of electronic suicide notes can be analyzed through the technique of authorship attribution.

IV. SUCCESSFUL SYSTEMS

Authorship attribution has been successfully implemented in a few ways, but they have been primarily focused on transcripts that are available in the language of English. Some of the notable implementation of authorship attribution are given below.

1) JGAAP: JGAAP (Java Graphical Authorship Attribution Program) is a Java-based, modular, program for textual analysis, text categorization, and authorship attribution i.e. stylometry / textometry. JGAAP is intended to tackle two different problems, firstly to allow people unfamiliar with machine learning and quantitative analysis the ability to use cutting edge techniques on their text based stylometry / textometry problems, and secondly to act as a framework for testing and comparing the effectiveness of different analytic techniques’ performance on text analysis quickly and easily. JGAAP is developed by the Evaluating Variation in Language Laboratory (EVL Lab) [6]

2) Signature: Signature is a program designed to facilitate “stylometric” analysis and comparison of texts, with a particular emphasis on author identification [4]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

Signature has many distinct features which includes:

- More powerful file-handling and filtering tools
- Ability to specify relevant alphabets and punctuation etc. for different languages/genres
- Wordlist facilities extended to accommodate phrases
- of specified length(s)
- Similar facilities for bigrams/trigrams etc. C
- Choice of Keynes measures for key word/phrase identification
- Fully automatic creation of frequent word/phrase lists
- Automated monitoring of previously specified words
- Powerful concordance, enabling also punctuation and proximity searches etc.
- Principal Component Analysis, applicable to all data types
- Burrows' Delta analysis, applicable to all data types
- Multiple chi-square analysis, applicable to all data types
- Main parameters of all facilities easily configurable
- Comprehensive help and theoretical documentation

3) JSAN: JSAN is a writing style analysis and anonymization framework. It consists of two parts: JStylo – authorship attribution framework and Anonymouth - authorship evasion (anonymization) framework JStylo is used as an underlying feature extraction and authorship attribution engine for Anonymouth, which uses the extracted stylometric features and classification results obtained through JStylo and suggests users changes to anonymize their writing style [5].

V. THEORETICAL BACKGROUND

Theoretical background of authorship attribution is completelythat of statistics present inside the textual content ofthe document. Statistical study of disputed authorship shouldinvolve comparisons of the disputed text with work by eachof the possible candidate authors using appropriate statisticaltest on quantifiable features of the texts, features which reflectthe "style" of the writing [3].
Typically, given an authorship attribution or related task,researchers will choose a set of function words, and construct datasets consisting of frequency vectors of function wordusage, for each of several sections of texts with known authorship. A statistical and/or a machine learning methodis then applied to these data, yielding a classifier. The classifier is then applied to test data, which, in a real case ofdisputed authorship, will be function word frequency vectors associated with the disputed text(s). The familiar range ofclassification methods have been attempted, including neural networks, support vector machines, and various statisticaland probabilistic approaches, including linear discriminant analysis and evolutionary search [2].

VI. REVIEW OF LITERATURE

Authorship attribution studies began in 1887, when Mendenhall reported using word-length distributions to study certain of the works of John Stuart Mill, comparing them to work by others on the same topic. Mendenhall followed this up in 1901, by applying his method to certain works of Shakespeare and of Bacon. Though seminal, Mendenhall's work can be criticized for mistakenly revealing differences in word-length distributions between poetry and prose, rather than between different author's styles. The first to examine sentence-length distribution (rather than word-length) was Yule, who attempted to characterize authorship in terms of, for example, mean and standard deviation of number of words per sentence. This method showed some success, leading to more sentence-length based studies, however such studies were supplanted in the 1960s by characterizations in terms of function words, and more generally on vocabulary distribution, which measures the diversity of an author's vocabulary. Typically, vocabulary distribution is

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

modeled by frequency distributions of the number of words appearing exactly r times (for example) for various r. The function words approach, in contrast, deals only with specific words (such as pronouns, conjunctions, prepositions, and so forth) that have no significant meaning, but are grammatically and syntactically important. Research in function word based authorship attribution flourished following Mosteller and Wallaces demonstration of their use for the case of the disputed Federalist Papers. Work in this area still tends overwhelmingly to concern English texts, and focuses on the comparison of different learning methods and/or augmentations to a function word approach.

A glance at the literature review

Serial No:	Year/Author	Features	Techniques	Corpus
1	(2006) Rong Zheng, Jiexun Li, Hsinchun Chen, Zan Huang	Lexical, syntactic, structural, content Specific	SVM	English Internet newsgroup messages & Chinese Bulletin Board System (BBS) messages
2	(2006) Ahmed Abbasi and Hsinchun Chen	Lexical, syntactic, structural, content Specific	Classification	USENET forum, Yahoo group forum , website forum for the White Knights
3	(2007) Daniel Pavelec, Edson Justino, and Luiz S. Oliveira	Linguistic Features	N-Gram	2 Brazilian newspapers, Gazeta do Povo
4	(2008) EFSTATHIOS STAMATATOS	Stylistic Fearures	N-Gram	CorpusVolume1(RCV1) Arabic Corpus
5	(2008) Kim Luyckx and Walter Daelemans	Syntactic Features	Memory based	Personal corpus
6	(2008) Chun Wei	Email features	Clustering	Email dataset
7	(2008) Farkhund Iqbal, Rachid Hadjidj, Benjamin C.M. Fung, Mourad Debbabi	Stylometric Features	Frequent Pattern	Enron Dataset

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

8	(2008)M.Connor	Syntactic	Decision Trees/KNN.	Emails collected from users
9	(2009) Rachid Hadjidj, Mourad Debbabi, Hakim Lounis, Farkhund Iqbal,Adam Szporer, Djamel Benredjem	Stylometery Features	Stastical Analysis, ML	Enron Dataset
10	(2011) George K. Mikros and Kostas Perifanos	Linguistic features	N-Gram	Greek Tweets

VII. PROPOSED SYSTEM

- Step1: Data collection: -Collect relevant corpus written by potential authors.
- Step2: Feature Extraction: -After extraction, each unstructured text is represented as a vector of writing-style features.
- Step3: Model Generation:-Dataset should be divided into training and testing set.
- Step4: Author Identification:-Developed model can be used to predict the authorship of unknown transcript.

Authorship attribution system for Malayalam transcripts based on the n-gram Model is proposed. This model uses n-grams for representing the text. N Gram Model was first used in text categorization based on the statistical information gathered from the usage of sequence of characters [4]. N grams are consecutive overlapping characters formed from an input stream. N-gram means that token of n words is used extracting the words from the passages and these n-grams matched. Then the resemblance measures are computed for text categorization. In this approach, a document collection D = D1, D2, D3, … is used as reference corpus, and a suspicious(plagiarized) document collection P = P1, P2, P3, …. are used. Each suspicious document Pi will be compared to the documents in D.

This model uses n-grams for representing the text. N Gram Model was first used in text categorization based on the statistical information gathered from the usage of sequence of characters [4]. Character and word n-grams have been used successfully previously in AAI tasks with character bigrams to appear as early as 1976 in the relative literature (Bennett 1976). They exhibit significant advantages over other stylometric features since their identification can be achieved easily and they are language-independent [7].

N-gram profiles, i.e. sequences of different n-gram sizes have also been used as feature group in authorship attribution research but they are limited in one level each time, either using characters or words. word and character level n-grams function complementary and capture different text regularities and combining these two levels could result in more robust and accurate results [Peng 2004]. Taking into consideration the complementary nature of character and word level information, we propose a combined vector of both character and word n-grams of different size.

an n-gram is a contiguous sequence of n items from a given sequence oftext.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

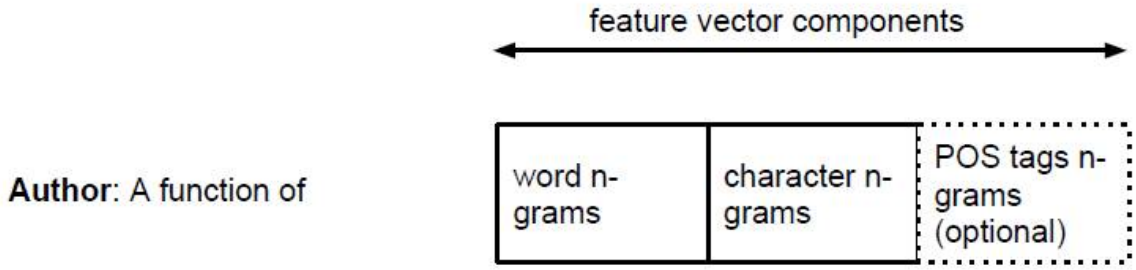
(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

So, for the sentence above:

word 2-grams (or bigrams): [(an, n-gram), (n-gram, is), (is, a), (a, *contiguous),* …]
char 2-grams: ['an', 'n ', ' n', 'n-', '-g', …]



Then the resemblance measures are computed for text categorization.
As stated above in this approach, a document collection D= {D1, D2, D3, …}is used as reference corpus, and a suspicious document collection P= {P1, P2, P3…} are used.
Each suspicious document Pi will be compared to the documents in *D*

The resemblance measures are computed as follows:

$$R\_Score_n = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|}$$

Where S(A) is the set of n-gram from passage A,
where A S(B) is the set of n-gram from passage B where B ε P.

The Matched n-grams are calculated as:

$$M = S(A) \cap S(B)$$

And the total number of n-gram is computed as:

$$N = S(A) \cup S(B)$$

$R_score_n$is the resemblance metric of two documents when segmented by n,

$$0 \leq R\_score_n \leq 1 .$$

When R_score *n* =0, document A and B have no identical n-gram,
And if R_score *n* =1, document A and B are identical.

The smaller n is the more candidates will be detected. The value of R ranges between 0 and 1. We have set a threshold of 75% resemblance as the threshold for classifying texts.

VIII. CONCLUSION

A large number of authorship attribution studies have previously implied that authorship traits span across a wide spectrum of linguistic levels. N-grams still represent one of the best suited features for representing stylometric profiles of small size texts. Employed method proved highly efficient compared to single n-gram feature groups in all text sizes.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

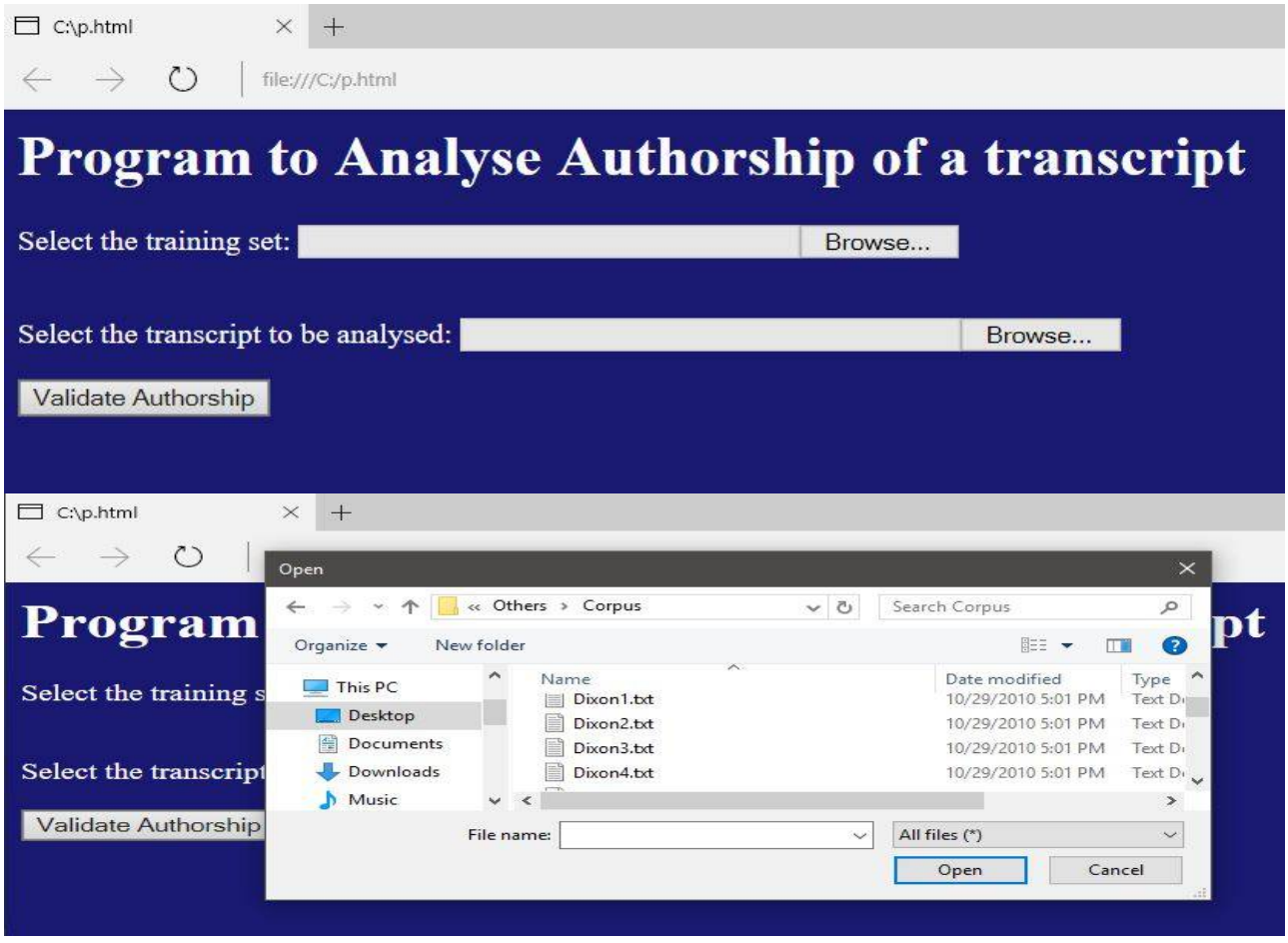
Vol. 6, Issue 2, February 2017

(a) The obtained results indicated that optimal results are achieved when both training and testing sets for authorship attribution contained merged transcripts.
(b) The text-size effect in the authorship attribution accuracy correlates directly with the test set. When the test set consists of text units containing merged tweets, we found that bigger texts imply higher attribution accuracies.

Although Malayalam as a language poses many difficulties in many authorship attribution techniques, unlike the function word technique the method employed in this paper proved to be more effective in case of Malayalam authorship attribution.

In the near future we plan to extend the present study in a number of ways. We will continue to develop the Malayalam corpus in order to enlarge its users and record their full metadata. Furthermore, we will expand n-gram feature representation with bigger n-grams and compare different machine learning algorithms in order to find the best blend. Finally, we can try to include speech to text functionality in to the program and use the collected data to increase the efficiency of person identification in biometric systems.

IX. SIMULATION AND RESULTS



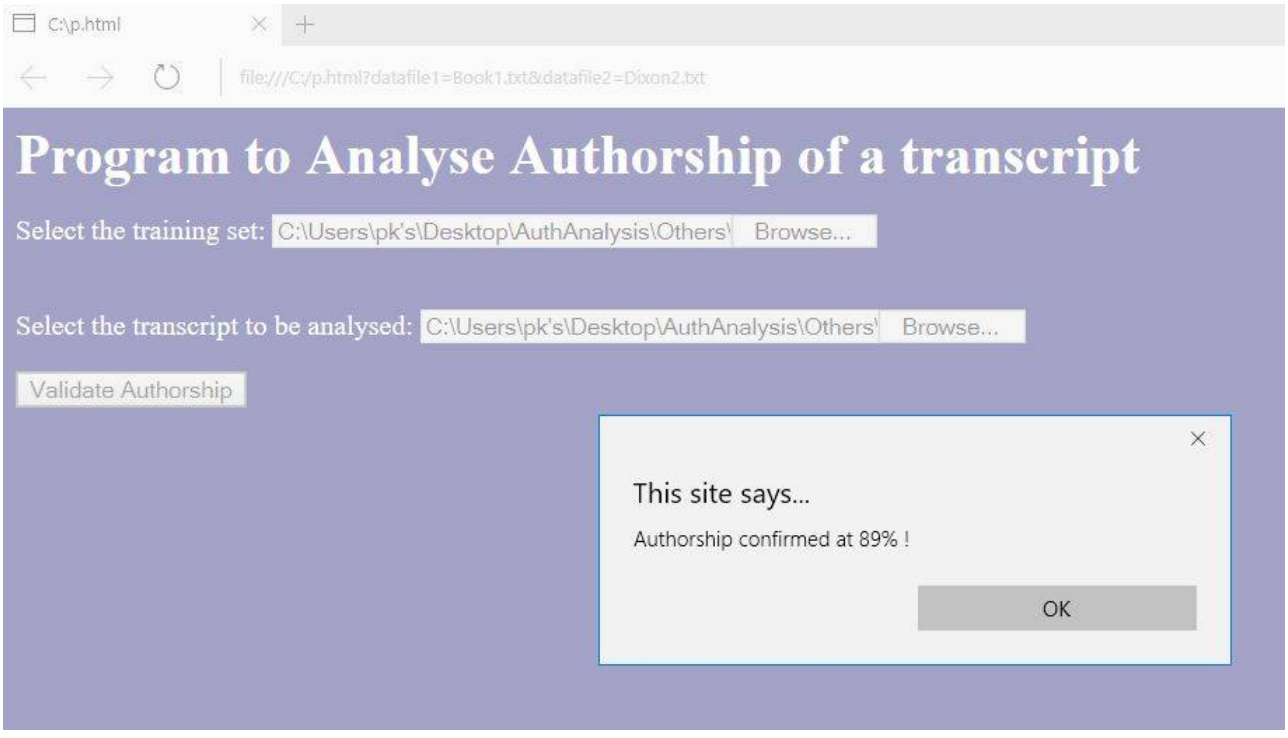
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017



REFERENCES

[1] Abbasi, Ahmed and Chen, Hsinchun "Applying authorship analysis to extremist-group Web forum messages", IEEE Intelligent Systems Volume: 20, Issue: 5, Sept.-Oct. 2005

[2] Shaker, Kareem and Corne, David, "Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis", 2010 UK Workshop on Computational Intelligence (UKCI), 2010

[3] Holmes, David I, "Authorship Attribution",Vol 28, Issue 2, Page 87-106, 2016

[4] Website://philoComp.net

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 2, February 2017

[5] Andrew McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman and Rachel Greenstadt. "Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization". Privacy Enhancing Technologies Symposium (PETS 2012)Xuelian Long and James Joshi," A HITS-based POI Recommendation Algorithm for Location-Based Social Networks" IEEE/ACM 2013

[6] Juola, P., Sofko, J., Brennan, P. (2006). A Prototype for Authorship Attribution Studies. Literary and Linguistic Computing 21:169-178Pavlos Kosmides, Chara Remoundou, Konstantinos Demestichas, Ioannis Loumiotis, Evgenia Adamopoulou, MichaelTheologou," A Location Recommender System for Location Based Social Networks" IEEE 2014

[7] Keim, Daniel A. and Oelke, Daniela "Literature fingerprinting: A new method for visual literary analysis" VAST IEEE Symposium on Visual Analytics Science and Technology 2007, Proceedings 2007

[8] Li, J., Chen, H., & Huang, Z. "A Framework for Authorship Identification of Online Messages: Writing-Style Features and classification Technique", Journal of the American Society for Information Science, 57(3), 378–393. doi:10.1002/asi,2006.

[9] Pavelec, D. and Oliveira, L. S. and Justino, E. and Neto, F. D Nobre and Batista, L. V. "Compression and stylometry for author identification", Proceedings of the International Joint Conference on Neural Networks, 2009

[10] Stuart, Lauren M. and Tazhibayeva, Saltanat and Wagoner, Amy R. and Taylor, Julia M. "Style features for authors in two languages", Proceedings - 2013 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2013

[11] Sallis, Philip and Shanmuganathan, Subana, "A blended text mining method for authorship authentication analysis", Proceedings - 2nd Asia International Conference on Modelling and Simulation, AMS 2008

[12] Stuart, Lauren M. and Tazhibayeva, Saltanat and Wagoner, Amy R. and Taylor, Julia M." On identifying authors with style", Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013