Innovative Methods in Identifying Authors of Documents

Pandian A[a], Mohamed Abdul karim Sadiq[b]
[a]Dept. Of MCA, SRM University, Kattankulathur, 603203, India,
pandian.a@ktr.srmuniv.ac.in
[b]Dept. Of Information Technlogy, College of Applied Sciences, PO box 135, Sohar 311, Oman,
sadiq.ak@gmail.com

Abstract - With the advent of internet, we have loads of documents online. Many of these are anonymous or claimed by more than one person. Identifying the authors of such documents is beneficial for many reasons. The textual content is composed of linguistic domains. Each of these domains is governed by rules, yet within these rules and among the components, the grammar offers flexibility to the writers. In this paper we compare the various techniques used to identify the corresponding authors of documents.

Keywords: text processing; authorship attribution; feature extraction; machine learning

I - INTRODUCTION

Textual documents can be viewed as an outcome of particular choices made by its authors. This is the reason each document carries the specific characteristics of its creator. These can be referred to as fingerprints of text. While trying to determine authorship, the following assumptions arise.

- There is a single author
- There are choices the author decides
- The author is consistent in his/her preferred choices
- These choices are present and could be detected in all end products of that creator

Author Identification study is useful to identify the most plausible authors and to determine evidences to support the conclusion. Authorship analysis problem is categorized as follows,

- Authorship identification: Attribution determines the likelihood of a piece of writing to be produced by a particular author by examining other writings of that author.
- Authorship characterization: It summarizes the characteristics of an author and generates the author profile based on his/her writings like gender, educational, cultural background, and writing style.
- Similarity detection: It compares multiple pieces of writing and determines whether they were produced by a single author without actually identifying the author like plagiarism detection. To extract unique writing style from the number of online messages, various features such as lexical, syntactic, structural, content-free and content-specific need to be considered.

Although authorship attribution problem has been studied in the past, but in the last few decades, authorship attribution of online messages has become a forthcoming research area as it is a confluence of various research areas like Machine Learning, Information Retrieval and Natural Language Processing. Initially this problem started as the most basic problem of author identification of anonymous texts, now it has expanded for forensic analysis, electronic commerce etc. This extended version of author attribution problem has been defined as needle-in-a-haystack problem [12].

When people write an article on any topic, they use certain words unconsciously. Our objective is to find some underlying pattern of the author's style. The fundamental assumption of authorship attribution is that each author habitually uses specific words that make their writings unique. Extraction of features that distinguish one author from another includes use of statistical or machine learning techniques on large corpus of text.

In Section 2 below we review existing techniques used for Authorship Analysis along with their classification. Section 3 explains basic procedure for authorship analysis. Section 4 summarizes comparison of various techniques since the year 2004 till 2014. Section 5 reviews performance evaluation parameters required for Authorship Analysis Techniques. This is followed by section 6 to conclude the paper.

II - STATE OF THE ART TECHNIQUES

2.1 Brief History

The advent of non-traditional authorship attribution techniques can be traced back to 1887, when Mendenhall first created the idea of counting features such as word length. His work was followed by work from Yule (1938) and Morton (1965) with the use of sentence lengths to judge authorship.

2.2 Applications of Authorship Attribution

- To analyze anonymous or disputed documents and books such as the ancient articles and poems written by various authors.
- Plagiarism detection - to establish whether claimed authorship is valid.
- Criminal Investigation – to determine source of unauthorized or unsolicited Emails
- Forensic investigations - verifying the authorship of spam mails, newsgroups messages, or identifying the basis of a piece of intelligence.

2.3 Key Features

- When an author writes they use certain words unconsciously.
- Find some underlying 'fingerprint' for an author's style.
- The fundamental assumption of authorship attribution is that each author has habits in wording that make their writing unique.
- It is well known that certain writers can be quickly identified by their writing style.
- Extract features from the given text that differentiate an author from another
- Applying certain statistical or machine learning techniques on given training data
- Showing examples and counterexamples of an author's work

2.4 Issues involved in the process

Identification of authors needs expertise in linguistics, statistics, text authentication, literature, etc. Hence, this is an interdisciplinary area. Too many style measures have to be applied and style markers need to be determined. Although statistical methods may be complicated or simple, too many exist in the literature. The features are extracted only after parsing all the documents thoroughly. The results have to be combined in order to obtain certain characteristics about the authors. Apply each of the statistical or machine learning approaches to assign a given document to the most likely author.

2.5 Current Techniques

Computerized analysis of documents was developed in 1980's, from the previous statistical analysis of literary style. This is termed "Stylometry". In order to quantify some of the features of an author's style, the following measures are explored.

Word or Sentence Length: This is a method developed in the origin of Stylometry. Due to the naïve quantification, it is not a reliable method.

Function Words: This method relies on word usage and context-free words. Using this method, we can analyze words' frequency, position, and immediate context of words. This is a criticized method, and cannot reliably distinguish between certain literature types.

Vocabulary Distributions: In this method, we measure the richness or diversity of an author's vocabulary. It analyzes the frequency profile of word usage to glimpse the author's extent of vocabulary.

Content Analysis: This method tabulates the frequency of types of words in a text. It aims to reach the denotative or connotative meaning of the text.

III CLASSIFICATION OF METHODS

The methods for authorship attribution are broadly classified based on the statistical or machine learning approach adopted for the purpose. These are summarized in Figure 1 below. The statistical univariate methods include Naïve Bayes Classifier, Cusum Statistics procedure and Cluster Analysis.

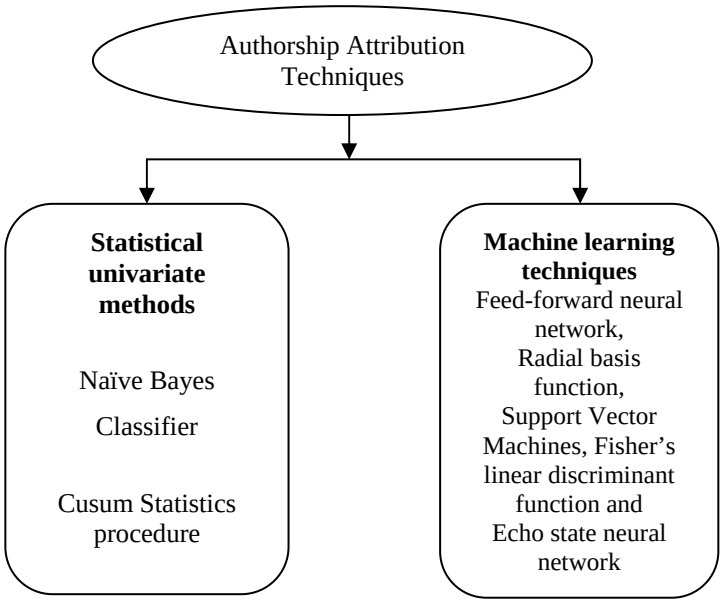

Fig. 1.Main Techniques in Authorship Attribution

The machine learning techniques are Feed-forward neural network, Radial basis function, Support Vector Machines, Fisher's linear discriminant function and Echo state neural network.

IV TYPICAL PROCEDURE

The procedure followed in identifying authors typically consists of four stages as shown in Figure 2. The first step is data collection. During this phase, we collect materials written by potential authors from various sources and store them in digitized form.

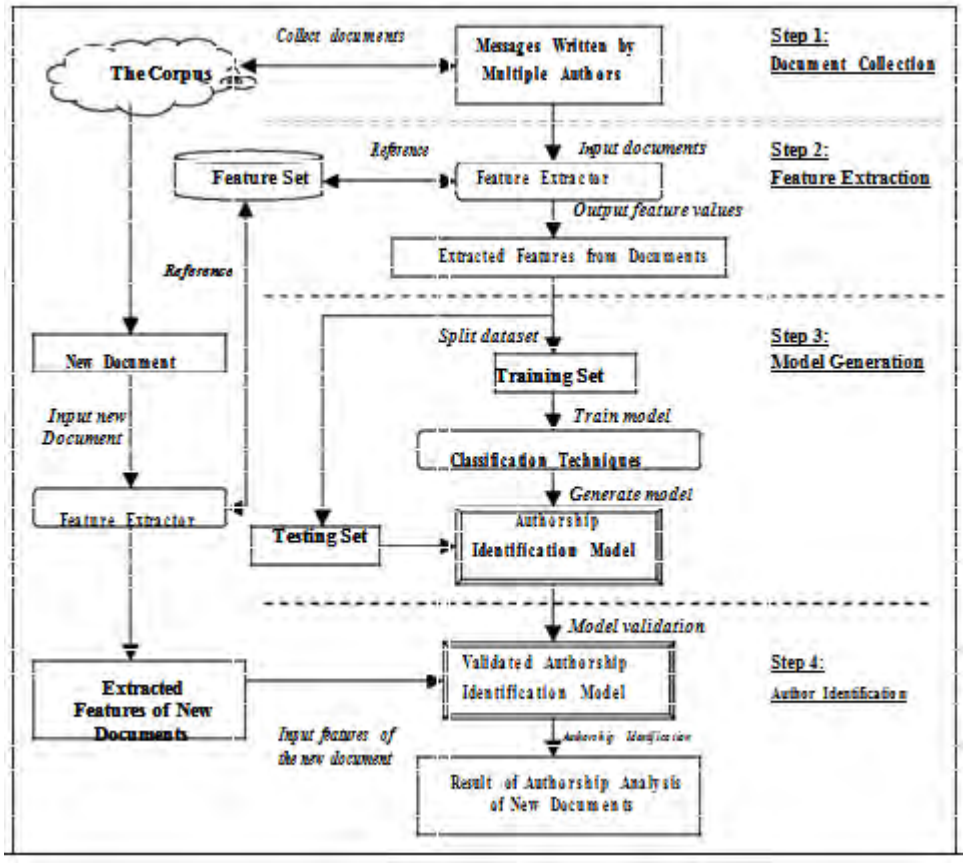

Fig.2.Stages in Authorship Attribution

Feature extraction is the second stage. After extraction, each unstructured text is represented as a vector of writing-style features. The next step is model generation. The dataset is a large collection of textual documents. This should be divided into training and testing sets. Classification techniques are applied, while an iterative training and testing process is undertaken. Finally, in the fourth stage, author identification is done. The developed model is used to predict the authors.

V COMPARISON OF TECHNIQUES

This section summarizes the various techniques used for authorship identification reported in research forum since 2004 till 2014. History of studies on authorship attribution problems is presented in tabular format year wise.

For each method, we identify the corpus on which the method was tested, the feature types used and the categorization method used, along with the size of training set. Table 1 lists the comparative study of all authorship techniques.

TABLE I

Comparison of various Authorship Attribution methods in English language (NB=Naïve Bayes; NN=neural nets; k-NN=k nearest neighbors; MVA=multivariate analysis; PCA=principle component analysis; LDA=linear discriminant analysis

Author (s)	Year	Corpus	Features	Techniques Used
Mendenhall	1887	Bacon/Marlowe/Shakespeare	Sentence length, word length	Distance
Mascol	1888 (a,b)	Pauline Epistles	FW(10s), punctuation	Distance
Yule	1938	De Gerson	Sentence length	Distance
Yule	1944	De Gersen	Vocabulary richness (K-measure)	Distance
Fucks	1952	English and german authors	Word length	Distance
Brinegar	1963	QCS letters	Word length	Distance
Mosteller & wallace	1964	Federalist papers	FW(10s)	NB
Morton	1956	Ancient Greek Prose	Sentence length	Distance
Burrows	1987	Austen/S.Fielding/H.Fielding	FW(10s)	MVA++PCA
Burrows	1992(a)	Brontes	FW(10s)	MVA+PCA
Matthews & Merriem	1993	Shakespeare/Fletcher	FW(1s)	NN
Kjell	1994 (a,b)	Federalist Papers	Character n-grams	NN,NB
`Merriam & Mathews	1994	Shakespeare/Marlowe	FW(1s)	NN
Ledger & Merriam	1994	Shakespeare/Fletcher	Character n-grams	MVa
Holmes & Forsyth	1995	Federalist Papers	FW(10s), vocabulary richness	MVA, genetic algorithm
Kjell et al	1995	WSJ	Character n-grams	NN, k-NN
Lowe & Mathews	1995	Fletcher/Shakespeare	Fw(1s)	RBF-NN
Martinedale & McKenzie	1995	Federalist Papers	Words	MVA+LDA,NN
Mealand	1995	Book of Luke	FW(10s),POS	MVA
Baayen et al	1996	Federalist Papers	Syntax	NN
Merriam	1996	Shakespeare	FW(1s)	MVA+PCA
Tweedie et al	1996	Federalist Papers	FW(1s)	NN

S.Argamon et al	1998	Newspapers & magazines	FW(100s), POS n-grams	ID3,Ripper
Tweedie & Baayen	1998	English prose	FW(10s), vocabulary richness	Distance,MVA+PCA
Binongo & Smith	1999	Shakespeare	FW(10s)	MVA+PCA
Craig	1999	Middleton	Words	Distance
Hoom et al	1999	Dutch poets	Character n-grams	NN,NB,k-NN
Stamatatos et al	2000	Greek newspapers	Syntactic chunks	Distance
Waugh et al	2000	Renaissance plays, Federalist paper	Words	NN
Kukushkina et al	2001	Russian texts	Character n-grams,POs n-grams	Distance(Markov)
Chaski	2001	Four Women	Syntax,punctuation,various	Distance
De Vel et asl	2001	Emails	FW(10s),complexity,various	SVM
Holmes et al	2001(a)	Pickett letters	FW(10S)	MVA+PCA
Holmes et al	2001(b)	Crane articles(purported)	FW(10s)	NVA+PCA
Stamatatos	2001	Greek newspapers	Syntactic chunks	Distance(LDA)
Baayen et al	2002	Dutch texts	FW(10s),syntax	MVA+PCA
Benedetto et al	2002	Italian Texts	Character n-grams	Distance(compression)
Burrows	2002(a, b)	Restoration-era poets	FW(10s)	MVA+PCA
Hoover	2002	Novel and articles	Words, word n-grams	MVA
Khmelev & Tweedie	2002	Federalist papers, various	Character n-grams	Distance(Markov)
Binongo	2003	Oz books	FW(10s)	MVA+PCA
Clement & Sharp	2003	Movie reviews	Character n-grams	NB
Diederich et al	2003	German newspapers	Words	SVM
Hoover	2003(a)	Novels and articles	Words, word n-grams	MVA
Hoover	2003(b)	Orwell/Golding/Wilde	Words, word n-grams	MVA
Hoover	2003(c)	Novels	Vocabulary richness	MVA
Keselj et al.	2003	English noverls, Greek newspapers	Character n-grams	MVA
Khmelev & Teahan	2003	Russian texts	Character n-grams	Distance (Markov)
Koppel & Schler	2003	Emails	FW(110s), POS n-grams, idiosyncrasies	SVM, J4.8
Argamon et al.	2003	BNC	FW(100s), POs n-grams	Winnow
Hoover	2004(a)	American novels	Words	MVA+PCA
Hoover	2004(b)	Novels and articles	Words	MVA+PCA
Peng et al.	2004	Greek newspapers	Character n-grams, word n-grams	NB
Van Halteren	2004	Dutch texts	Word n-grams, syntax	MVA
Abbasi	2005	Arabic forum posts	Characters, words,	SVM, J4.8

& Chen			vocabulary, richness, various	
Chaski	2005	10 anonymous authors	Character n-grams, word n-grams, POS n-grams, various	Distance (LDA)
Juola & Baayen	2005	Dutch texts	FW(10s)	Distance (cross-entropy)
Zhao & Zobel	2005	newswire stories	FW(100s)	NB, J4.8, k-NN
Koppel et al.	2005	Learner English	FW(100s), POS n-grams, idiosyncrasies	SVM
Kopper et al.	2006a	Brontes, BNC	FW(100s), POS n-grams	Balanced Window
Zhao et al.	2006	AP stories, English novels	FW(100s), POS, punctuation	SVM, distance
Madigan et al.	2006	Federalist papers	Characters, FW(100s), words, various	Bayesian regression
Zheng et al. Li et al.	2006	English and Chinese newsgroups	Characters, FW(100s), syntax, vocabulary richness, various	NN, J4.8, SVM
Argamon et al.	2007	novels and articles	FW(100s), syntax, SFl	SVM
Burrows	2007	Restoration poets	Words	MVA+zeta
Hirst & Feiguina	2007	Brontes	Syntax	SVM
Pavelec et al.	2007	Portuguese newspapers	Conjunction types	SVm
Zhao & Zobel	2007	Shakespeare, Marlowe, various	FW(100s), POS, POS n-grams	distance (infogain)
Abbasi & Chen	2008	Emails, online comments, chats	Characters, FW(100s), syntax, vocabulary richness, various	SVM, PCA, other
Argamon et al.	2008	Blogs, student essays, learner English	Words, SFl	Bayesian regression
Stamatatos	2008	English and Arabic news	Character n-grams	SVM
Farkhund Iqbal et al.	2010	Enron E-mail Dataset which contains 200,399 e-mails	lexical, syntactic, structural, and content-specific features	EM, k-means, and bisecting k-means
Sarwat Nizamani & Nasrullah Memon	2013	Enron E-mail Dataset which contains 200,399 e-mails	lexical, structural, syntactic features and content specific	Cluster-based Classification (CCM) technique

TABLE II
Details of Authorship Identification Techniques used in other languages corpus

Authors / Year & Languages	Features	Techniques Used	Corpus	No. of Authors	Used Training Set
E. Stamatatos, n. Fakotakis & G. Kokkinakis *2001* Greek	Text length, frequency etc.,	Sentence and chunk boundaries Detector	*TO BHMA* Greek weekly newspaper	10 authors	300 texts
Paulo Varela, Edson Justino & Luiz S. Oliveira *2010* (Brazilian Portuguese)	Verbs and Pronouns	SVM	Brazilian newspapers, Gazeta do Povo www.gazetadopovo.com.br & Tribuna do Paraná www.paranaonline.com.br	20 Authors	Collection of Short Articles
Tanmoy Chakraborty & Sivaji Bandyopadhyay *Feb - 2011* (Bengali)	Detection of Stylometry	cosine-similarity, chi-square measure, Euclidean distance	30 Stories written by Indian Nobel laureate Rabindranath Tagore	One	20 Stories
Bei Yu *June – 2012* (Chinese)	Function words	EM Clustering Algorithm	Federalist Papers Dataset	Many Authors	Novels, Essays and Blogs
Sreeraj.M & Sumam Mary Idicula *2012* (Malayalam)	scale, space and orientation from images	Scale Invariant Features Transform	Collection of Handwriting Samples	280 Writers	Handwriting samples of all writers
Jayashree R1, Srikantamurthy K1 and Basavaraj S Anami *Sep – 2013* (Kannada)	Word Occurrence and No of Unique words	Naïve Bayesian Method, dimensionality reduction Techniques	Comprehensive Kannada Text Resource – TDIL	Many Authors	1791 paragraphs
Hemlata Pande & H. S. Dhami *Oct – 2013* (Hindi)	Mean Word Length, Average Deviation, Frequency of words of length etc.,	Discriminant analysis	Navbharat Times & ELRA-W0037	Many Authors	337 Texts
Vishnu Murthy.G, Dr. B. Vishnu Vardhan, K. Sarangam & P. Vijay pal Reddy *Nov - 2013* (Telugu)	100 features	Naive Bayes (NB), Support Vector Machine (SVM) and k Nearest Neighbor (kNN) classifiers	Telugu News Papers - Eenadu, Andhra Prabha & Sakshi	Many Authors	800 News Articles
A.Pandian and Md. Abdul Karim Sadiq *December 2013* (Tamil)	322 features of Emails	Fisher's linear discriminant function, Radial basis function & Echo state neural network	Emails of 50 Authors	50 Authors	500 Emails
R. Lakshmi Priya and G. Manimannan *January 2014* (Tamil)	morphological and function words	Principal Component Analysis (PCA) and Multivariate Discriminant Analysis (MDA)	Tamil Language Magazine "India" 1906	3	92 Articles

TABLE III
Details of Authorship Identification Techniques used in Multilingual Corpuses:

Authors / Year & Languages	Features	Techniques Used	Corpus	No. of Authors	Used Training Set
Rong Zheng, Jiexun Li, Hsinchun Chen, & Zan Huang *Dec – 2005* (English & Chinese)	lexical, syntactic, structural, and content-specific features	Decision trees, back propagation neural networks, and support vector machines	English News group messages & Chinese Bulletin Board System Messages	Many Authors	Online Messages
Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford & Ben Hutchinson *2007* (Arabic & English)	Demographics and psychometrics features of the authors	Text Attribution Tool	Emails of Arabic and English Writers and Speakers	1,030 Arabic & 1,033 English Authors	8,028 Arabic Emails & 9,836 English Emails
Yohei Seki, Noriko Kando, & Masaki Aono *2009* (English & Japanese)	grammatical subjects and predicates, nouns and adjectives/verbs	SVM & Lexicon Based Heuristics	NTCIR-6 Opinion Corpus & MPQA Corpus	155 Authors from Japanese & 565 Authors from English	Sample Topics
Maciej Eder *2011* (English, Polish, German, Latin)	Frequencies of frequent words	Delta Method	Collection of Prose Texts	20	Around 70 Texts
Jacques Savoy *2012* (English, French, German)	Word types and Lemmas	Principal Component Analysis and Delta Approach	19th and 20th Century Novels	78 Authors	52 Excerpts from English, 44 Segments from French and 59 excerpts German

VI CONCLUSION AND FUTURE WORK

To determine prediction accuracy, the number of authors and the size of training data set both play vital role. This comparative study concluded that if number of author's increases and size of training sets decreases then performance degrades. So far, there were no studies examining their impact on the authorship-identification performance in a systematic way. The problem of authorship attribution is explored well in the area of English language, but limited work has been done for the authorship identification in other languages and multilingual. Thus, by considering all these further research direction is to do work in various languages other than English and also concentrate on authorship identification of a multilingual text having more than one language or a single method used to identify more than one language.

REFERENCES

[1] E. Stamatatos, et al.,(2001), Computer-Based Authorship Attribution Without Lexical Measures, Computers and the Humanities 35: 193–214.

[2] Rong Zheng, et al., (2005), 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.

[3] Dominique Estival, et al., (2007), TAT: an author profiling tool with application to Arabic emails, Proceedings of the Australasian Language Technology Workshop 2007, PP. 21-30.

[4] Yohei Seki, et al., (2009), Multilingual opinion holder identification using author and authority viewpoints, Information Processing and Management 45 PP. 189–199

[5] Paulo Varela, et al., (2010), Verbs and Pronouns for Authorship Attribution, 17th International Conference on Systems, Signals and Image Processing.

[6] Tanmoy Chakraborty & Sivaji Bandyopadhyay, (2011), Inference of Fine-grained Attributes of Bengali Corpus for Stylometry Detection, Polibits (44) , PP. 79 – 83.

[7] Maciej Eder, (2011), Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint, Studies in Polish Linguistics, PP. 99 – 114.

[8] Jacques Savoy, (2012), Authorship Attribution: A Comparative Study of ThreeText Corpora and Three Languages, Journal of Quantitative Linguistics 19, issue 2, PP. 132-161.

[9] Bei Yu, (2012), Function Words for Chinese Authorship Attribution, Workshop on Computational Linguistics for Literature, PP. 45–53
[10] Sreeraj.M & Sumam Mary Idicula, (2012), The Effect of SIFT Features as Content Descriptors inthe Context of Automatic Writer Identification in Malayalam Language, Optical Networking Technologies and Data Security, PP. 632 – 636.
[11] Jayashree R, et al., (2013), Suitability of Naïve Bayesian Methods for Paragraph Level Text Classification in the Kannada Language using Dimensionality Reduction Technique, International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 5, PP 121 – 131.
[12] Hemlata Pande & H. S. Dhami, (2013), Analysis for the significance of statistical word-length features in genre discrimination of Hindi texts, IOSR Journal of Mathematics, Volume 8, Issue 1, PP 05-10.
[13] Vishnu Murthy.G, et al., (2013), A Comparative Study on Term Weighting Methods for Automated Telugu Text Categorization with Effective Classifiers, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.6, PP. 95 – 105.
[14] A.Pandian, Md. Abdul Karim Sadiq,(2013), Authorship Attribution in Tamil Language Email For Forensic Analysis, International Review on Computers and Software, Vol. 8 No. 12.
[15] R. Lakshmi Priya and G. Manimannan (2014), A Study of Ambiguous Authorship in Tamil Articles using Multivariate Statistical Analysis, International Journal of Computer Applications, Volume 86 – No,1.

AUTHOR PROFILE

A. PANDIAN received his B.Sc.,and MCA degree from Bharathidasan University, Tiruchi. He received his M.Tech degree from Punjabi University, Patiala, Punjab and M.Phil. degree from Periyar University, Salem. He is doing Ph.D.(Computer Science & Engineering) in SRM University, Chennai. He has over Eighteen years of experience in teaching. He is working as Assistant Professor (Sr.G) in the Department of MCA, SRM University, Chennai. His areas of interest are text processing, information retrieval and machine learning. He is a Life member of ISTE and ISC.

Dr. M. Abdul Karim Sadiq holds Ph.D. in Computer Science & Engineering from Indian Institute of Technology, Madras. He has over Sixteen years of experience in software development, research, management and teaching. His areas of interest are text processing, information retrieval and machine learning. Having published papers in many international conferences and refereed journals of repute, he filed a patent in the United States Patent and Trademark Office. He is an associate editor of Soft Computing Applications in Business, Springer. Moreover, he organized certain international conferences and is on the program committees. He has been awarded a star performer in the software industry.