Authorship Analysis and Identification Techniques: A Review

Mubin Shaukat Tamboli
Department of Computer Engg
Amrutvahini COE
Sangamner, India

Rajesh S. Prasad, Ph.D
Department of Computer Engg.
Zeal Education Society, Narhe
Pune, India

ABSTRACT

Trends in data mining are increasing over the time. Current world is of internet and everything is available over internet, which leads to criminal and malicious activity. So the identity of available content is now a need. Available content is always in the form of text data. Authorship analysis is the statistical study of linguistic and computational characteristics of the written documents of individuals. This paper describes review of various methods for authorship analysis and identification for a set of provided text. Surely research in authorship analysis and identification will continue and even increase over decades. In this article, we put our vision of future authorship analysis and identification with high performance and solution for behavioral feature extraction from set of text documents.

General Terms

Data Mining, Authorship Identification

Keywords

Features extraction, n-gram, lexical, structural, stylomatric features identification, Writeprint.

1. INTRODUCTION

Way to determine the authorship of handwritten document, and text document is a very old one. Now large volume of text is available in the form of digital content. Attribution for any text to a known ancient authority was essential to determining the text's veracity. This problem of author attribution is most important because of application in forensic analysis, humanities scholarship, electronic commerce, and the development of computational methods for addressing the problems.

In recent years, the use of data mining techniques are increased for several purposes. Data available might be in any format like text, images, binary, and multimedia. And several techniques of mining increased, modified, improved over the time. Here we focus on author identification techniques. Even before the world of computer, this technique was in its way shows in work of Mendenhall (1887). Today the availability of text document in electronic form increases the importance of using automatic methods to analyze the content of text documents. Initially identifying document was very time consuming, expensive and has its limit. That emerges text categorization in predefined categories called as classification. Categorization is based on certain properties called as features. There are various method for extraction of features. Writeprint is one analogue to finger print method. Another n-gram features, will give information about increasing word sequence according to its lengthf. Next is focus on stylomatric features. It includes a set of the style markers which are adapted for the automatic analysis of the text. A Source Code (programming code) Author Profiles (SCAP) represents a new, highly accurate approach to source code authorship identification. Another section, text categorization based on keywords that may appear uniquely, may dual sequences like computer+science, genetic+algorithm etc. In discriminative syntactic tree approach, there is direct mining from a given set of syntactic trees.

Later part of this paper is organized in research work done, challenges in feature extraction and classification.

2. RELATED WORK

2.1 n-gram Features

A document is represented by a feature vector that contains one Boolean attribute for each word that occurs in the training collection of documents. When this method generalized by using word sequences and form a sequence, termed as n-gram as a feature. For generation of n-gram features, consideration was made for small value of n, number n-gram features that can be discovered in document, which increases in such a way that for every n-gram, there is at list one n+1gram that has n-gram as a starting sequence. Thus features are growing linearly, the number of features with minimum frequency grows much slower. So, an efficient algorithm for generating these feature sets, should therefore avoid generating all n-grams. Algorithm utilizes three parameters as the collection of documents, MaxGramSize, MinFrequency. Algorithm is based on the APRIORI-algorithm for discovering frequent item subsets in databases. Outcomes from the project includes, learning algorithm which removes stop words, word sequence of length 2 or 3. Longer sequences reduces the performance. The results seem to indicate that the addition of n-grams to the set-of-words representation frequently used by text categorization systems improves performance. However, sequences of length n>3 are not useful and may decrease the performance as in [1].

The n-gram method for text categorization for Bangla in newspaper corpus as in [6]. Paper describes analyzing the efficiency of n-grams, and shows that tri-grams have much better performance for text categorization for Bangla [6]. The use n-gram to specific for two sequences as bigrams [4]. The research paper [4], mainly focuses to find bigrams in which at least one of the constituent words (hereafter unigrams) has a minimum document frequency in at least one of the categories.

Biagrams are selected in such a way that their occurrence and information gain is high.

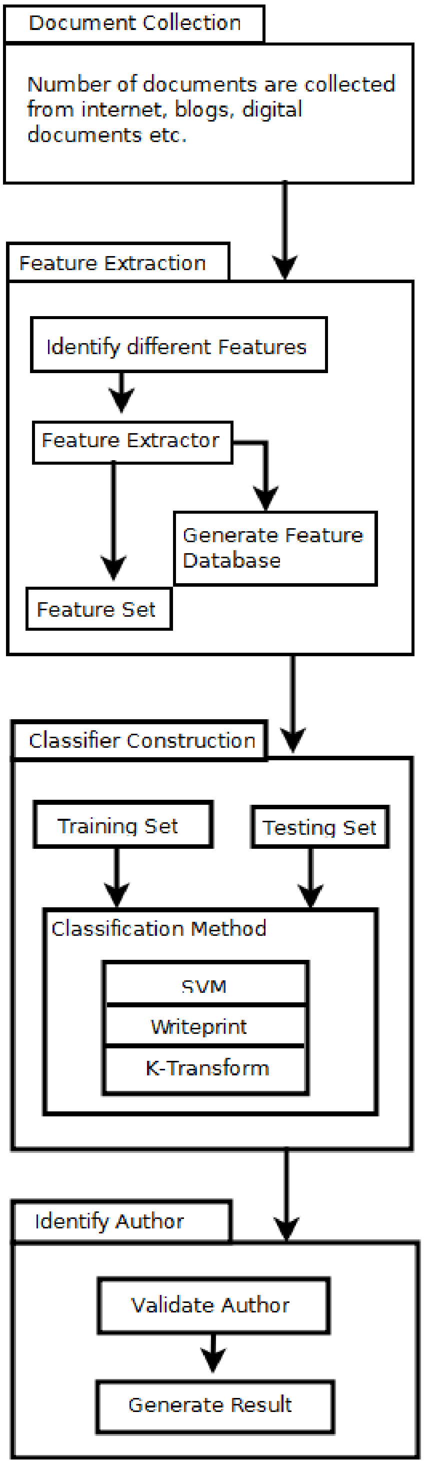

Fig 1: Data Mining Steps for Author Identification

2.2 Stylomatric Approach

A set of stylistic features, includes linguistic features, character based features, word based statistic features, syntactic features, structure based features, and function words. Classification process includes normalized features using max-min normalization method to put values between 0 and 1 as in [12].

Stylometric analysis techniques, which are categorized into supervised and unsupervised methods [15]. In character based features there are number of white characters, special characters etc. The word based features includes words with vocabulary richness. It also indicates writing trends emotional words, cognitive words, frequency of particular cues, appearance of related words. Syntactic features include punctuation marks such as comma, semicolon, question mark etc. These are very general and depend on facts like habits, mood, expression etc. Structural based features depends on layout, length of sentence, organization of writing skill. Function words shows vocabulary richness, lexical meaning, personal styles etc [15].

2.3 Writer dependent and independent Strategies

Along with above basic features, there is concern of originality, i.e. if original document is of one author and is copied by some other author. At this time to find real author of document is difficult. There are two different approaches [14], one writer dependent, and writer independent so as to build a robust method for identifying authorship. Again, it uses same stylometric features described in previous section based on conjunction and adverbs. Writer dependent model is based on individual author. In independent model, it is based on the forensic questioned document examination approach and classifies the writing in terms of authenticity, using the global model [14].

2.4 Structure Specific Architecture

The structure specific feature includes the structure of documents in terms of representation. Normally it includes length of sentence, separators used in sentences, length of paragraph, for specific fact represented by number of sentence, repetition of part of sentence, layout of whole document. Indirectly we can termed it as habit of author reflected by its writing style [21].

2.5 Other Strategies

Over internet huge amount of data / text is available in form of blogs, emails, digital contracts, books and many more. So correct identity of this available data is difficult. That might incur cybercrime. The problem of anonymity in online communication is addressed by applying authorship analysis techniques. In past times, a lot of research is going on for analysis and identification for owner of data content. There are many approaches for the authorship analysis. In research describes [18] frequent pattern based on writeprint, capturing stylistic variation, analysis based on different training sampling sizes, presentable evidence, removing burden from investigator, clustering by stylometric features for varying training samples and authorship characterization.

Another specific category in which information grouped is topic modeling. In one description [17] author describes idea for topic modeling. Topic is identified by distribution over the words and their frequency in corresponding corpus. This distribution are found from the text document with the help of statistical techniques such as Dirichlet allocation or Gibbs sampling. Another way is Hierarchical topic model.

In research work [16], author uses hirachical generative model. In this model each word associated with two variables one is author and other is topic. A set of N dimensional vector is used indicating defined variables , topics and author assigned for N words.

In research study [3], a fully automated approach to the identification of authorship on unrestricted text that exclude lexical measures. Described method excludes distributional lexical measures. Instead of using sentence length, punctuation marks and syntax based (noun phrase count, verb phrase count. It provides the way of analyzing text and way of capturing information. This method lacks linguistic theory as it is based on statistical measures. This paper also describes the method to capture diversity of an author's vocabulary, one is type-token ratio V/N where V is the size of the vocabulary of the sample text, and N is the number of tokens which form the sample text. Another way of measuring the diversity of the vocabulary is to count how many words occur once (i.e., hapaxlegomena), how many words occur twice (i.e., dislegomena) etc. These measures are strongly dependent on text-length.

Research information [4] introduces the detail about biagrams to improve text categorization. Presented algorithm used the information gain metric combined with various frequency thresholds and bigram can substantially raise the quality of feature sets by increasing breakeven point and measures.

The solution for disputed authorship is present as in [5]. Due to causal basis likelihood judgment and conditional dependencies, scholar makes critical errors. Study provides Bayesian inference in distributed authorship. Two hypotheses (H and ~H) are to examine the passages of each document and judge the extent to which each passage supports or refuses each hypothesis.

A method for source code authorship identification is uses a SCAP (source code author profile) method [8]. Wide ranges of features are considered for java and common lisp and depending on programs, comments, layout features and packages selected naming influences, classification accuracy other like user defined names, program related features not appeared to influences accuracy. Feature considered for the same are programming layout metrics style, metrics structure, and linguistic metrics. Where SCALP approach makes use of n-gram contiguous sequences defined at lower level attribute of a program. Program content categorizes into features like comment layout, identifier programming structure etc [8].

Documents are observed in hierarchical fashion, stylistic characteristic of author, group of author specific rules are used to build classifier and recursive data mining approach is performed as in [9]. Approach used to perform identification was RDM (Recursive Data Mining). Using token and patterns as a feature, performs as well as Navie Bayes, SVM (Support Vector Machine), and RDM using statistical information as a feature. Result on experiments shows the capturing stylistic pattern in SEA and Enron dataset and also used for organizational role of authors. The method divides the semantic knowledge for semantically related pattern.

The paper [11], describes a method as a navie bayes algorithm for feature and word selection, for the purpose of text classification. This algorithm is for multidimensionality classification. For that, it uses feature clustering to reduce dimensionality of feature vector for classification and for that put a fuzzy similarity based self constructing algorithm for feature clustering. The described algorithm improves performance of algorithm with elite strategy.

In the paper [12], which is for gender identification, it is based on human psychology. In this paper, total of 545 psycho-linguistic and gender-preferential cues along with stylometric features are used to build the feature space for this identification problem. Three machine learning algorithm are designed for gender identification based on the proposed feature. In the described technique, all features are collected and normalized using maxmin normalization method as described in equation (i) to insure all feature values from 0 to 1.

Normalized-xij = (xij-min(xj)) / (max(xj)- min(xj)) ……(I)

where xij is the jth feature in the ith example,

min xj and max xj are the minimum and maximum feature values of the jth feature, separately.

For classification techniques used are Bayesian-based logistic regression, Ada- Boost decision tree and SVM classifiers separately, using Reuters and Enron corpora.

Paper [14] defines two approaches one is writer dependent, and writer independent. Because of this strategy, it becomes robust. This method used features as forensic stylistic which is a subfield of forensic linguistic which aims at applying stylistics to the context of author identification, where is based on two writers do not write in the same way, and writer does not write in the same way all the time. Proposed work use conjunctions and adverbs of the Portuguese language to find author. In this work the authors extract features using a compression algorithm and achieve success rate of 78%.

Paper [7], [23] describes the feature extraction methods includes word similarity among sentence and their frequency occurred in statement. Similar sentence repeated over document, paragraphs, and provide a solution for classification using evolutionary programming with the help of fuzzy logic and artificial neural network. The description in paper provide a view for hybrid classification method which provide a direction for smart feature extraction.

In [15], the research study uses stylistic features, including lexical, syntactic, structural, content-specific, and idiosyncratic attributes. Writeprint method is also described in this paper. This study describes that, existing methods have focused on author identification task, but there are limitation for similarity detection, and provide summary of some features. Stylistic features represents lexical, syntactic, structural, content-specific, and idiosyncratic style markers. Lexical features include words, characters and their variance, length distributions. Syntactic feature includes function words, punctuation, n-grams. Structural features are as file extension, font, colors etc. Content specific features which are keywords, phrases, and topic name like word n-grams. Idiosyncratic features have misspellings, grammatical mistakes, and other usage anomalies. Author introduces extended feature set along with baseline feature. Extended features includes static and dynamic features. For its classifier constructions Writeprints technique used. This technique has creation and pattern disruption. For finding writing style variation Karhunen-Loeve transforms are applied with a sliding window in order to capture stylistic variation with a finer level of granularity.

The research desecribed in [17], a learning method to derive information about author and topic from text collection. Article gives simple probabilistic model for defining relationship between authors, documents, topics and words.

Research study of [8] introduces the method for classification of author based on high level programming features. Paper describes author identification using high-level features that contribute to source code authorship identification using a tool the SCAP method. Source code author profile (SCAP) is based on byte level feature which is used to assess high level programming features. Author describes, previous method of classification based on features, programming layout, style, structure and linguistic metrics.In a set of experiments, author uses feature in identifier, symbol name identifier, package name identifier, comments, layout metrics.

GA feature selection model is represented by [22], which is used to identify the writeprint features. In his model, features are represented by bits. Number of bits depends on candidate features. This defines accuracy. These features are generated successively and finally this GA model generates different combinations and utilized for classification and termed as key writeprint features to discriminate writing style of several authors.

3. CONCLUSION

Feature extraction and classification to identify the authorship of document is active research area. Above discussion provides the direction to which we could move. In review of various research papers we found unigram feature [1],[6], which focused on word sequences but other features like two and more sequences [4]. Other view after the n-grams, focus was made on syntax representation using lexical measures [3],[21]. Very common features are writing style of author in various ways [3],[9],[12],[15],[20],[21]. They are utilized in different ways to identify the author. All these are very general static methods of identifying and analysis of authorship.

In the presented research regarding error made by human due to stress, tension, facts are taken into consideration. Due to this authorship analysis may not be correct at every time. To overcome this, researcher has made thought on the cognitive error arise for regular author [5]. There is challenge to identify such cognitive error and differ as mentality of author may not be same for certain task. And for same person cognitive error changes as the age of human increases. So need to build a model which can identify this feature which supports on time factor. Similar the case for author style, so need to add a decision support system which builds from smart learning algorithm for collecting features in a group.

Apart from the pattern of individual author and analyses them in by representing as hierarchical fashion [9], specific word, phrase used in document, one can guess, who the owner of that document is. These features are used along with above features. This was described in research paper [4],[11]. The method for writer dependent and independent documents authorship identification and linguistic model uses features regarding writing methodology followed by other features. Selected features are reduced by clustering using genetic algorithm. At this stage challenge is to build a feature set which can be applied on both situations. So feature selection algorithm should have ability for selecting such features which not depends on who is writing rather than the owner of document.

4. REFERENCES

[1] Johannes Furnkranz, “A Study using n-gram Feature for Text Categorization”, Technical report OEFAI-TR-98-30, 1998

[2] Maria Fernanda Caropreso, “Statistical Prases in Automated Text Categorization,” IEI-B4-07-2000. Pisa, IT, (2000).

[3] E. Stamatatos, N. Fakotakis and G. Kokkinakis, “Computer-Based Authorship Attribution Without Lexical Measures” , Kluwer Academic Publishers, Computers and the Humanities 35, 2001, pp 193-214.

[4] Chade-Meng Tan, Yuan-Fang Wang, Chan-Do Lee, “The use of Bigrams to enhance Caegorization,” Inf. Process. Manage. 38(4): 529-546 (2002.).

[5] Kevin Burns, “Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support” Information Sciences 176, 2006 pp1570–1589.

[6] Munirul Mansur, Naushad UzZaman and Mumit Khan, “Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus,” School of Engineering and Computer Science (SECS), BRAC University, 2006.

[7] Prasad, R.S., U.V. Kulkarni and J.R. Prasad, “A novel Evolutionary Connectionist Text Summarizer (ECTS)”, Proceedings of IEEE International Conference on Anti-Counterfeiting, Security and Identification, Aug. 20-22, IEEE Xplore Press, Hong Kong, pp: 606-610. DOI: 10.1109/ICASID.2009.5277003.

[8] Georgia Frantzeskou, Stephen MacDonell, EfstathiosStamatatos, StefanosGritzalis, “Examining the significance of high-level programming features in source code author classification” The Journal of Systems and Software 81, 2008 pp. 447–460.

[9] Vineet Chaoji, Apirak Hoonlor and Boleslaw K. Szymanski, “Recursive Data Mining for Author and Role Identification” Proc. 3rd Annual Information Assurance Workshop ASIA’08, 2008, pp. 53-62.

[10] Moshe Koppel, Jonathan Schler, Shlomo Argamon, “Computational Methods in Authorship Attribution”.

[11] B. Rama Krishna, J. Ramesh, “An Efficient Self Constructing Algorithm for Text Categorization” International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 7, 2012, ISSN: 2278-0181.

[12] Na Cheng, R. Chandramouli, K.P. Subbalakshm, “Author gender identification from text” Eslevier Digital Investigation 8 (2011), pp 78-88.

[13] Abdur Rahman, Haroon A. Babri, Mehreen Saeed, “Feature Extraction Algorithms for Classification of Text Documents”, ICCIT 2012, pp. 231-236.

[14] Daniel Pavelec, Edson Justino, Leonardo V. Batista, and Luiz S. Oliveira, “Author Identification using Writer-Dependent and Writer-Independent Strategies” SAC’08 March 16-20, 2008, ACM 978-1-59593-753-7/08/0003, pp. 414-418.

[15] Abbasi, A. and Chen, H. “Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace” ACM Trans. Inf. Syst. 26, 2, Article 7 (March 2008), pp. 1-29.

[16] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M. “Learning author topic models from text corpora” ACM Trans. Inform. Syst. 28(1), Article 4 January 2010, pp. 1-38.

[17] Giacomo Inches, Fabio Crestani, “Online Conversation Mining for Author Characterization and Topic

Identification” PIKM’11, October 2011, ACM 978-1-4503-0953-0/11/10.

[18] Farkhund Iqbal, HamadBinsalleeh, Benjamin C.M. Fung, MouradDebbabi, “A unified data mining solution for authorship analysis in anonymous textual communications” Elseveir Pub., Information Sciences 231 (2013) pp. 98–112.

[19] Jacques Savoy, “Authorship attribution based on a probabilistic topic model,” Information Processing and Management 49 (2013) Elsevier Pub. pp. 341–354.

[20] ShlomoArgamon, Marin Sari, Sterling S. Stein, “Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results,” SIGKDD’03, August 2003 pp. 24-27, Washington, DC, USA, ACM 1-58113-737-0/03/0008.

[21] Rong Zheng, “A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques,” Wiley Periodicals, Inc., Published online 21 December 2005 (www. interscience. wiley.com).

[22] Jiexun Li, RongZheng, and Hisinchun Chen, “From Fingerprint to Writeprint,” Communication of ACM, April 2006 Vol. 49 No. 4 pp. 76-82.

[23] Prasad, R.S., U.V. Kulkarni, “Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization,” Journal of Computer Science 6 (11) 2010, pp. 1366-1376, ISSN 1549-3636.