ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 6, August 2012

Review and Study of Different Methods for Author Identification

Akhil Sanjeev Gokhale
Department of Computer Engineering
VIIT,Pune-48.

Rajendra Krishnat Dalbhanjan
Department of Computer Engineering
VIIT,Pune-48.

Dr.Rajesh.S.Prasad
Department of Computer Engineering
DCOER,Pune.

Abstract**—**As assessment of an individual has become an important aspect in the field of distance learning, it is difficult for these institutions to verify whether the individual participated is the person enrolled. Author identification has played a vital role to prevent plagiarism as well as authenticating an individual's identity. This paper gives an overview of the methods which includes machine learning techniques such as "support vector machines", that are useful for author identification to prevent plagiarism. It also includes various methods such as "QSUM technique", "Type-Token ratio", "Readability Measures", "Content Analysis" and "Artificial Neural Networks". These methods are found to be useful in text classification and a variety of practical applications. This paper concludes that above all mentioned methods Support Vector Machine (SVM) holds an upper hand compared to these methods.

Index Terms—author identification, machine learning, support vector machine, stylometry, Artificial Neural Networks.

I. INTRODUCTION

In this era of E-learning, the difficulty of the institutions is to safely deliver the material related to the course required for higher education. Interaction between institutions and students plays an important role in E-learning.

Flexible delivery can be defined as presenting online, the educational material, including theory guides, reading material, exams and facilities for the interaction between tutor and students and also between students. This flexible delivery of course material becomes an issue of prime importance for educational institutions. As this is Internet-based, face-to-face interactions are rare and often not practical due to distance. While many institutions of higher education offer courses via distance education, even if this is not the primary mode of delivery, assessment is one aspect which cannot be realized using the Internet exclusively. Traditional identity checks are expensive, e.g. if the student has to be present physically or has to take the exam at a local institution (other than the online university). In addition, the student may have to sign documents that certify his/her identity and as a result, legal costs may occur. Without any online authentication of the student's identity, flexible delivery breaks down at this point[1]. Here comes the role of author identification with the help of which plagiarism involved in online education system can be avoided. Author Identification also plays an important role in E-mails, short text messages and in many other domains for avoiding duplication of author identity.

This paper summarises different methods which have proved to be beneficial for authorship identification.

II. STYLOMETRY

Every author has a unique style of writing a document consisting of certain features that distinguishes him from other authors. Stylometry deals with identifying this peculiarity which every author imbibes. Earlier studies related to stylometry introduced the concept of counting features in a document and then applying this to both "word lengths" and "sentence lengths". Due to the difference in writing style of an author, variations are observed in sentence length provided these may change over time and also with the domain of a text. There are also differences in word length distributions in the prose and verse of the same author. Other features are counts of words beginning with a vowel or counts of words with certain lengths. The 'richness' or diversity of an author's vocabulary is a powerful criterion of stylometry. It is also observed that the number of words that occur a certain number of times depends on the age and intelligence of an author. In order to remove the dependency of vocabulary size from the text length, alternate features are also proposed. These range from the simple type-token ratio to more complex measures. An interesting feature is the comparison of the number of words that occur exactly j-times in the training data and the number of words which occur exactly j-times in a new text. Many studies found differences in the size of the vocabulary of authors, but also that vocabulary size is not a constant for any given writer. Hence, features such as vocabulary size are easy to calculate but have limited value for authorship attribution. The collection of different features, which may include vocabulary size in different word fields or the knowledge of specific words, has a power to discriminate authors.[1]

III. RELATED WORK

According to Ephratt [1] the author attribution comes from the following premises:
1. There is a specific single author.
2. There are choices to be made.
3. The author is consistent in his/her preferred choices, and
4. These choices are present and can be detected in all end products of that creator.

Bailey[1] defined three rules for authorship attribution in a forensic context:
1. The number of authors should constitute a well-defined set.
2. The lengths of the writings should be sufficient to reflect the linguistic habits of the author of the disputed text and also of each of the candidates.
3. The texts used for comparison should be commensurate with the disputed writing.

In his survey, Rudman[1] observed that "results of most non-traditional authorship attribution studies are not universally accepted as definite." Author attribution (also called "stylistics" or "stylometrics") is part of the judicial system of Great Britain, Canada and Australia, but not the United States. Rudman generally complains about a lack of experimental rigor, nevertheless, there are a number of statistical techniques which have been imported from other fields and which dominate the field of computer-based authorship attribution. Most notable were the Efron-Thisted Test, which were originally applied to ecology, and QSUM (or cusum), originally from industrial process and quality control monitoring.

IV. METHODS

Following are the review of the methods that can be used for authorship identification:-

A. Type/Token Ratio

The ratio of the number of unique words (type) to the total number of words (token) is called as Type/Token Ratio. This method can be considered as a measure of vocabulary richness which is one of the features for identifying author. However, this method fails to fully identify the author from accuracy point of view.

B. Readability Measures

Readability measures calculates the complexity of a document, and these calculations are based on sentence length and word length. It has been found that readability scores had high resemblance across all writing samples written by the same author. This method seems to be useful in author identification.

C. Content Analysis

Every document has specific content bounded to a particular domain. Analysis of this content plays an important role in identifying the author. This method classifies each word in the document by semantic category, and statistically analyses the distance between documents.

D. QSUM Technique

This technique is based on the observation that every author has a unique set of words which the author follows frequently through his writing. QSUM includes a number of measurements, starting with the "average sentence length" in a sample of a person's document. Here, names are treated as single symbols. Each sentence is compared to the average of the sample and marked with a + if it is longer and a – if it is shorter. This generates a sentence length profile. The next step includes the calculation of the deviations of each sentence from the average. Taking these final values for the sentences, it is possible to visually inspect the sample in form of a graph. QSUM continues by analysing the use of function words by an author as well as "shorter" words, e.g. "vowel words" (words beginning with a vowel) and combinations such as "short + vowel word" [1]. It was analysts have found that there are nine tests which can be applied to samples. The three most common are the use of the 2 and 3 letter words, words starting with a vowel (initial vowel words); and the third is the combination of these two. This combination often proved the most useful identifier of consistency. The other tests involve the use of words with four letters.One of these nine tests—and sometimes more than one— will prove consistent for a writer.It was used in a number of court cases and received significant public attention, however, a number of independent investigations found the method unreliable [1].
A brief outline of the method is given below:
Assume that n1 is the number of word types in a corpus that occur exactly once and n2 the number of word types that occur exactly twice etc. If a sample text is considered which is not part of the baseline corpus, mj is defined as the number of word types in the sample which occur exactly j times in the baseline corpus. Therefore, m0 is the number of word types that do not appear in the baseline corpus, m1 is the number of word types that occur exactly once in the corpus etc . That is, mj does not depend on the sample text alone but also on the baseline corpus.From this it is clear that nj and mj are directly observable [1]. This method plays a vital role for author identification.

E. Neural network and artificial intelligence techniques

Artificial Neural Networks(ANN) are characterized by its important feature known as learning ANNs are adaptive, i.e. they can change internal representations as a response to training data, sometimes combined with a teaching input. Since all knowledge in ANNs is encoded in weights, i.e. numeric values associated with links connecting network nodes (units), learning is performed by weight change. A weight represents the strength of association, i.e. the co-occurrence of the connected features, concepts, propositions or events represented by a unit during a training or learning period. On the network level, a weight represents how frequent the receiving unit has been active simultaneously with the sending unit. Hence, weight change between two units depends on the frequency of both units having positive output simultaneously.

This form of weight change is called Hebbian learning which provides a simple mathematical model for synaptic modification in biological networks[1]. The basic principle, here is the local weight change depending on the outputs/states/potentials of the connected units. ANN techniques have been successfully applied across a broad spectrum of problem domains such as pattern recognition and function approximation. Moreover an ANN solution manifests itself entirely as sets of numbers. As such a trained ANN offers little or no insight into the process by which it has arrived at a given result nor, in general, the totality of "knowledge" actually embedded therein. Clear obstacle to a more widespread acceptance of ANNs is its lack of capacity to provide a "human comprehensible" explanation . To overcome this limitation, considerable effort has been directed towards providing ANNs with the requisite explanation capability. In particular a number of mechanisms, procedures, and techniques have been proposed and developed to extract the knowledge embedded in a trained ANN as a set of symbolic rules which in effect mimic the behaviour of the ANN [1]. A standard feed forward artificial neural network (also called multi-layer perceptron) has been used to attribute authorship to the disputed Federalist papers. In this method rather than counting relatively rare words, the number of times a set of predetermined words occur, is counted. For this network eleven function words (an, any, can, do, every, from, his, may, on, there, upon) are considered. These words are believed to be good discriminators as their rate of use should be relatively constant for each author and each author should have a distinguishable rate. In this case The data is normalised so that each word has a rate that is normally distributed with a mean of 0 and a variance of 1. The total set of Federalist papers was split into three groups:

the joint author papers, the disputed papers and the undisputed single-author papers. The joint and the disputed papers are the test set while the undisputed papers are the training set. The neural network had an "eleven input, three hidden and two output nodes" architecture. The eleven function words are input to the neural network and the two possible authors are the output of the network. This network was trained with conjugate gradient and tested by use of k-fold cross-validation. The network unambiguously classified the disputed Federalist which was found to be consistent with the results of authors using other methods[1].

For stylometric analysis RBF-type neural network has been used. An RBF-network uses a linear transfer function at the output nodes and alternative, nonlinear functions at the hidden nodes. An RBF-network is a generalised Gaussian classifier or predictor, the hidden nodes represent local response functions; i.e. the distance between a weight and a pattern vector presented to the network. A hidden node's activation decreases as the distance between the input vector and the weights (the centre of the node's response) increases. RBF-networks have a number of advantages over standard feed forward neural networks. They are easier to interpret and a number of rule-extraction from neural network techniques are available for RBF networks . Also, prior knowledge can be used to initialise weight vectors[1]. This method gives favourable results when used for small data sets for author identification.

F. Support Vector Machines

Support Vector Machines method relies on the Structural Risk Minimisation (SRM) principle. SRM includes a bound on the difference between the empirical and actual risk. The former is typically identified by the test error over some unseen data set (e.g. as part of cross-validation), while the latter is the actual error independent of data sets that have been sampled for training and testing of a classifier. The SRM principle states that the function identified by a learner with the smallest empirical error selected from a set of functions with the smallest VC-dimension (a measure for the complexity of the hypothesis space of a learner) will have the smallest difference between actual and empirical error. Support vector machines find the hypothesis h out of the hypothesis space H of a learning system which approximately minimises the bound on the actual error by controlling the VC-dimension of H. SVMs are very universal learning systems [1]. In their basic form, SVMs learn linear threshold functions. The most important property of SVMs for text mining and authorship attribution is that learning is independent of the dimensionality of the feature space [1]. SVMs evaluate hypotheses by use of the margin they use for separating data points, not the number of features or attributes. This allows good generalisation even in the presence of a large number of features[1]. Following features of SVM makes it the preferred method for learning text classifiers:

1. High-dimensional input spaces: If every word of a text is a feature, the input space can easily be larger than > 100,000.

SVMs control overfitting internally, and therefore, large feature spaces are possible.
2. Few irrelevant features: Feature selection is normally used to avoid input spaces of high dimensionality. In text classification, this is either not practical or many features are equally important. Therefore, SVMs are a convenient way to learn a text classifier with limited pre-processing.3. Document vectors are sparse: For the reasons mentioned above, SVMs are ideally suited for sparse input vectors of high dimensionality.
4. Most text categorisation problems are linearly separable.
This method is of prime importance for author identification.

V. CONCLUSION

In the process of Identification of Authors the important features in a document remains unknown and the text cannot be completely analysed. This has been one of the crucial problems faced in the field of author identification. However, the above methods discussed in this paper provides a way to overcome this problem. Of all the above methods mentioned, SVMs for author identification can process documents of relevant length, large databases which do not require pre-deciding features. It has been observed that SVM technology has its roots in machine learning as compared to other methods. Therefore it can be said that SVMs for author identification can be useful. It can also be concluded that author identification can be useful for plagiarism detection in the context of online learning(e-learning).

REFERENCES

1.Joachim Diederich “Computational methods to detect plagiarism in assessment” Paper No. 145: Diederich J.:Computational methods to detect plagiarism in assessment 2006ITHET.

AUTHORS

AKHIL SANJEEV GOKHALE
Pursuing B.E.(Computer Science & Engineering),University of Pune,
Vishwakarma Institute of Technology, Survey No. 2/3/4, Kondhwa (Bk),
Pune 411048, India.

RAJENDRA KRISHNAT DALBHANJAN
Pursuing B.E.(Computer Science & Engineering),University of Pune,
Vishwakarma Institute of Technology, Survey No. 2/3/4, Kondhwa (Bk),
Pune 411048, India.

RAJESH SHARADANAND PRASAD
Professor in Computer Engineering; from Muzaffarpur (Bihar), India. B.E. (Computer Science & Engineering), North Maharashtra University, Jalgaon, India in Jun 1996, M.B.A. (Marketing), North Maharashtra University, Jalgaon, India in May 1998, M.E.(Computer Engineering), University of Pune, in 2004, Ph.D. from SRTM, University, Nanded, India in Mar 2012. Lecturer, Department of Computer Engineering at S.S.B.T.’s College of Engineering and Technology, Jalgaon, 1996-98; Lecturer, Department of Computer Engineering, S.E.S. College of Engineering, Kopargaon, 1998-2000; Lecturer, Department of Computer Engineering, Sinhgad College of Engineering, Pune, 2000-2004; Head, Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, 2005-2012. Professor, ZES’s Dnyanganga College of Engineering & Research, Pune, since Jul 2012Author: Software Engineering, 2001;Principles of Programming Languages, 2002, Compiler Construction, 2001, Computer Graphics and Multimedia, 2006, Articles published in CSI Communications, Volume No. 31, Issue No. 9, ISSN: 0970-647X, Lecture Notes in IMECS-2007, ISBN: 978-988-98671-4-0, Lecture Notes in IMECS-2008, ISBN: 978-988-98671-8-8.Member of IEEE, ISTE, CSI and IAENG.Contributed expertise for Industry Institute interaction, Involved with curriculum development at University of Pune, member of LIC and UGC interview committees at University of Pune. Project Examiner, BVCOEW, April 2007,Pune; Member of the Advisory Committee, National Conference on Artificial Intelligence, May 2007, Baramati (India).Session Chairman, National Conference on Emerging Trends in Computational Sc. & Information Processing, JNEC Aurangabad (India), April 2006; Session Chairman, National Conference on Computing, Communication and Electronica, KCE’s College of Engineering and IT, Jalgaon, February 2008; Member of the International Program Committee, 5th International Conference on Soft Computing, CSTST ’08, University of Cergy, Pontoise, Paris (France); Member of the International Program Committee, 2nd International Conference on Digital Information Management, (ICDIM ’08), Lyon, (France).Office: Department of Computer Engineering, Vishwakarma Institute of Technology, Survey No. 2/3/4, Kondhwa (Bk), Pune 411048, India.