Computational methods to detect plagiarism in assessment

Joachim Diederich

Abstract—**While many institutions of higher education offer courses via distance education, there is one aspect which is difficult to realise by use of the Internet only: assessment. If exams are performed online, how can the course provider guarantee that the student participating in the exam is the person enrolled? Without any Internet-based form of authenticating the student's identity, flexible delivery can break down at this point. As a consequence, traditional identity checks are introduced such as requiring the student to be physically present and to take the exam at a local institution, or requiring the student to sign documents that certify his/her identity.**

This paper discusses assessment in flexible delivery and how plagiarism can be detected. It presents a method for testing the identity of a student (or more generally, author) online, without any interference with the examination process. Recent advances in computational text analysis allow authorship identification with high reliability. That is, the original author of a document submitted for assessment can be determined successfully with an accuracy and precision of well above 90 percent. The computational methods include machine learning techniques such as "support vector machines", which are highly successful in text classification and a range of other practical applications.

***Index Terms*— plagiarism, authorship identification, machine learning, support vector machines**

I. Introduction

As the Web-based delivery of course material becomes more and more important in higher education, so does the use of advanced "Web-mining" technology to facilitate the interaction between institutions and students. This paper summarises previous work on authorship attribution and explores the use of "text mining" methods for the identification of students in an exam situation. Data mining techniques based on "support vector machines" (SVMs) offer the online analysis of student responses by use of machine learning, and as a result, allow the identification of a student. The method uses available information only (i.e. texts written by students prior to their enrolment and during the course taken) and does not require any additional hardware. Hence, data and text mining methods offer cost-effective technology for the identification of students in Web-based education.

Manuscript received March 14, 2006.

J. Diederich is with the American University of Sharjah, Sharjah, U.A.E. and the University of Queensland (corresponding author to provide phone: +971 6 515 2559; fax: +971 6 515 2979; e-mail: jdiederich@aus.edu).

A. Online authorship identification

The flexible delivery of course material becomes more and more important for universities and other educational institutions. Flexible delivery refers to the online presentation of educational content, including study guides, reading material, exams and facilities for the interaction between lecturer and students as well as between students. Flexible delivery is Internet-based and as a result, face-to-face interactions are rare and often not practical due to distance.

At this point in time, two types of organisations emerge which are based on the flexible delivery of course material or support online education. (1) Online universities which are fully Internet-based and offer courses such as the MBA via distance education, and (2) Support companies which provide the infrastructure: hardware (e.g. generally accessible server, modem banks) and software (e.g. Web design, organisation of chat rooms etc). The technology introduced here is important for both types of organisation.

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 realised using the Internet exclusively. If exams are performed online, e.g. an interactive Q&A session, how can the institution guarantee that the student participating in the exam is the person enrolled? Without any online authentication of the student's identity, flexible delivery breaks down at this point. 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.

Assessment is a major problem for online universities. At this point in time, the flexible mode of delivery is not available for exams or continuing assessment without the introduction of additional, costly checks of student identification. The following sections outline a method that allows verifying the identity of a student (or more generally, author) online without any interruption of the examination process. Students will not be aware that authorship of texts or documents is controlled and course providers have reliable, online and real-time fraud detection methods.

Consider the following scenario: Students are taking an online exam via the Internet. A number of questions are presented to one or more students and a machine learning system classifies the answers. This system takes written material in various forms as input and produces as output the name of the student. The accuracy of the classification (the identity of the student) will increase as the exam continues. If, for instance, the system is producing a name as output that is different from the student's name, the lecturer or instructor may take action that seems appropriate. Recent advancement in machine learning and in particular text classification brings this vision very close.

In summary, the data mining system would be trained to associate textual input with an author. This training or learning must be based on the presentation of written material from one or more authors, and answers to exam questions are most probably not sufficient for the learning process to be successful. In our scenario, the student would submit documents written earlier in his/her career at the time of enrolment. This could include assignments, essays, project reports and even letters; practically anything written by the student in the past. The data mining/machine learning system would be trained by use of this material and would be ready for use online at the time of the exam.

II. BACKGROUND

A. Stylometry

Central to the analysis of "style" is the assumption that every author has certain features that are *inaccessible to conscious manipulation*. Therefore, these features provide the most reliable basis for the identification of the author. However, the style of an author may vary because of differences in topic or genre and the personal development over time. The author may also use the *explicit imitation of literary styles*. Ideally, stylometry should be based on features that are invariant to the above mentioned effects but are expressive enough to discriminate an author from other writers.

Early stylometric studies introduced the idea of counting features in a document and applied this to both word lengths and sentence lengths [38]. There are differences in sentence length distributions for an author; these may change over time but also with the genre 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 [14].

A powerful criterion of stylometry is the 'richness' or diversity of an author's vocabulary. Zipf [32, 33] observed 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 have been 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. Thisted and Efron [27] estimated the size of Shakespeare's vocabulary by asking "How many new words would Shakespeare use if he were to write another play?"

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. It is clear that a *collection of different features*, which may include vocabulary size in different word fields or the knowledge of specific words, has a larger discriminatory power.

Instead of using word counts directly, it is possible to employ features derived from words [38]. An example is the syntactic class of words. Compared to the *use of syntax*, which is difficult to manipulate consciously, word use is more easily influenced by choices which are under the control of the author. As the discourse structure of texts from the same author and the corresponding vocabulary can be quite different, syntax-based features are more reliable for the purpose of authorship identification [38]. Charniak [31] discusses other techniques for enhancing statistical language processing with syntactic information.

According to Rudman [24; p.361] "approximately 1,000 style markers have already been isolated." There is clearly no agreement on significant style markers or "features". It seems that in text categorization nearly all words contain some information. Joachims [16] ranked 10000 word stems of a large corpus according to their information gain with respect to some classification. It turned out that a model using features with ranks 201-500 performed nearly as well as the best features in the top 1-200, and similar to feature set 400-9962. Hence even features ranked lowest still contain considerable information and are somewhat relevant. In the following section, statistical techniques for authorship attribution are introduced. While the conventional techniques rely on a few carefully selected features, newly developed approaches such as support vector machines allow the use of many hundred or even thousands of inputs features and alleviate the need for a careful selection. In addition, SVMs are machine learning techniques that built classifiers automatically. SVMs are the method of choice is section III.

B. Authorship attribution studies: Introduction

According to Ephratt [7] the rationale for authorship 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 [in 29] defines 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 [24] observes that "results of most non-traditional authorship attribution studies are not universally accepted as definite." Authorship attribution (also called "stylistics" or "stylometrics") is part of the judicial system of Great Britain, Canada and Australia, but not the United States. Non-traditional authorship attribution is a term reserved for computer-based author analysis and is the focus of the next sections.

Rudman [24] 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 notably the

- Efron-Thisted Test, originally applied to ecology, and
- QSUM (or cusum), originally from industrial process and quality control monitoring

These statistical methods are briefly introduced before the discussion continues with artificial intelligence (AI) techniques for authorship attribution.

C. Authorship attribution studies: Statistical techniques

QSUM is based on the observation that every author has a unique set of habits which s/he follows consistently when communicating [14]. QSUM includes a number of measurements, starting with the "average sentence length" in a sample of a person's utterance [9]. 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 [9]. 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" [14]. Farringdon [9] writes: Cusum 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 or speaker. QSUM was used in a number of court cases and received significant public attention, however, a number of independent investigations found the method unreliable [14]. Hilton & Holmes [12] and Rudman [24] challenge the QSUM technique: "When put to the same tests … [Morton's (the original author)] own writings seemed to bear the stamp of multiple authorship." (Philips, 1965, in Rudman [24]).

Thisted & Efron [27] estimated the size of Shakespeare's vocabulary by asking "How many new words would Shakespeare use if he were to write another play?" This is similar to the question: If a butterfly collector has already trapped x different species, what is the likelihood he will catch a new species on his next expedition? In 1985 Thisted & Efron tested their method by use of a newly discovered Shakespeare poem and approved it as original. Again, this result was questioned by Valenza (1990; in [14]).

The following is a brief outline of the method. Assume that n_1 is the number of word types in a corpus that occur exactly once and n_2 the number of word types that occur exactly twice etc. [14]. If a sample text is considered which is not part of the baseline corpus, m_j is defined as the number of word types in the sample which occur exactly j times in the baseline corpus. Therefore, m_0 is the number of word types that do not appear in the baseline corpus, m_1 is the number of word types that occur exactly once in the corpus etc [30]. That is, m_j does not depend on the sample text alone but also on the baseline corpus.

Obviously, n_j and m_j are directly observable. The Thisted & Efron technique consist of

1. an estimation of m_j and n_j based on modelling assumptions explained below,
2. a series of tests based on deviations of the observations from the estimates, and
3. various significance tests [30].

The core assumption of Efron & Thisted [5] is that word selection of an author is a Poisson process. For each word type a relative frequency is associated (between 0 and 1) independent of context. The main calculation is based on the estimation of word counts (by frequency of usage) expected in the sample from the corresponding counts of the baseline corpus. Only rare word types (that occur fewer than 100 times in the baseline corpus) are considered. The test that Thisted & Efron introduced compares the estimated number of new word types (not in the baseline corpus) to the observed number. This is not a direct comparison but takes into account the richness of the vocabulary of the sample [30].

Valenza [30] applied the Thisted & Efron tests to the works of Shakespeare and Marlowe and found good consistency for the Shakespeare plays but poor consistency between Shakespeare poems and plays or among Marlowe's plays.

D. Neural network and artificial intelligence techniques

Learning is the most important feature of artificial neural networks (ANN). 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. It provides a simple mathematical model for synaptic modification in biological networks. Several important modifications of this simple weight change rule have been proposed. The basic principle, i.e. local weight change depending on the outputs/states/potentials of the connected units, is accepted since.

Since their renaissance in the mid-1980s, ANN techniques have been successfully applied across a broad spectrum of problem domains such as pattern recognition and function approximation. However despite these capabilities, to an end user an ANN is an arcane web of interconnected input, hidden, and output units. 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. This lack of a capacity to provide a "human comprehensible" explanation is seen as a clear obstacle to a more widespread acceptance of ANNs.

In order to address this situation, recently 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].

Tweedie et al. [29] used a standard feedforward artificial neural network (also called multi-layer perceptron) to attribute authorship to the disputed Federalist papers. Rather than counting relatively rare words, Tweedie et al. decided to count the number of times that a set of predetermined words occur, in this case eleven function words (an, any, can, do, every, from, his, may, on, there, upon). 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.

The data was normalised so that each word had a rate that was 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 (Hamilton and Madison) are the output of the network. The network was trained with conjugate gradient and tested by use of k-fold cross-validation. The network unambiguously classified the disputed Federalist papers as being by Madison, which is consistent with the results of authors using other methods.

Lowe & Matthews [22] used an RBF-type neural network for stylometric analysis. An RBF-network uses a linear transfer function at the output nodes and alternative, non-linear 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 feedforward neural networks. They are easier to interpret and a number of rule-extraction from neural network techniques are available for RBF networks [1]. Also, prior knowledge can be used to initialise weight vectors. This is important for relatively small data sets, a situation that can easily arise in authorship attribution.

Five function word descriptors were extracted from each play of Shakespeare and Fletcher. These function word descriptors correspond to the occurrence of common 'scaffolding' words (are, in, no, of, the) drawn from examples of whole acts of plays ([22] p.455). These words are not particularly context-sensitive and since the authors try to find statistically reliable estimators, commonly occurring words are better for relatively small text samples.

A total of 50 samples were used for each author. Each set of ratios of occurrence from each author was normalised to zero mean and unit variance. The resulting data sets were used for the application of standard statistical methods as well as RBF-networks. The trained RBF-networks produced classifications in agreement with conventional scholarship and the application of computational methods such as multi-layer neural networks.

Holmes & Forsyth [13] used genetic algorithms to determine authorship of the disputed Federalist papers. Genetic algorithms are rule-based machine learning techniques that generate human comprehensible results. An example of a pro-Hamilton rule using rates of occurrences of function words is '(ON – THERE) <2,832> ([14] p.115). The rules produced by Holmes & Forsyth [13] use only eight function words, yet they correctly classify the disputed papers.

Elman [6] emphasised that language and speech unfolds in time and therefore, recurrent artificial neural networks which can accept a sequence of input patterns are the preferred choice for many natural language processing tasks. Schellhammer et al. [26] present preliminary results of experiments with recurrent neural networks for a natural language learning task. The strategy in these experiments is to start with simple children's texts and to step-wise increase the complexity of these texts to explore the learning characteristics of recurrent neural networks.

Schellhammer et al. [26] used two types of recurrent neural networks: Elman-networks and Recurrent Cascade Correlation (RCC) are trained on the text of a first-year primary school reader. The networks perform a one-step-look-ahead task, i.e. they have to predict the lexical category of the next following word. Elman-networks with 9 hidden units give the best training results (72% correct) but score only 63% when tested for generalisation using a "leave-one-sentence-out" cross-validation technique. An RCC network can learn 99.6% of the training set by adding 42 hidden units but achieves best generalisation (63%) with four hidden units only. The results are compared to probabilistic approaches, i.e. ngrams (bi-, tri-, 4- and 5-gram performance. In summary, the networks perform well, i.e. above 4-gram performance.

Towsey et al. [28] focus on the extraction of grammatical rules from trained Artificial Neural Networks and, in particular, Elman-type recurrent networks. The extracted grammatical rules do not only represent a portion of the English language, the also capture idiosyncrasies of individual speakers and authors. Exactly these idiosyncrasies can be used to identify authors [28].

E. Style marker and their role in authorship identification

There is clearly no agreement of significant style markers; techniques such as QSUM are a very heterogeneous collection of methods, some poorly justified. The artificial neural network studies described above very much rely on an ad hoc decision about relevant features: in one case five function words were used, in other cases eight or eleven. None of the available feature selection techniques in statistics, databases, data mining or neural networks were used. It is immediately apparent why the neural network techniques outlined above limited input to a few features only: feedforward neural networks allow for fixed-length vectors as input only and cannot represent the text as a whole.

This calls for the application of techniques such as recurrent neural networks and in particular "support vector machines" which are described in the next section.

F. Support vector machines

Support vector machines are based on the Structural Risk Minimisation (SRM) principle [16]. 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), 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 [16]. In their basic form, SVMs learn linear threshold functions. However, it is possible to "plug-in" kernel functions so SVMs can be used to learn polynomial classifiers, radial basis function (RBF) networks and three or more layered neural networks.

The most important property of SVMs for text mining and authorship attribution is that learning is independent of the dimensionality of the feature space [16]. 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.

Joachims [16] lists the following reasons why SVMs are a 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 has been empirically determined by a number of authors (Joachims, 1998).

G. *SVMs for authorship identification*

Diederich et al. [38] performed a number of experiments with texts from a German newspaper. With nearly perfect reliability the SVM was able to reject non-authors and detected the target author in 60-80% of the cases. In a second experiment, Diederich et al. [38] investigated a more content-free summary of a text, including counts of grammatical tags combined with bigrams to capture morphologic details of language patterns. This resulted in slightly reduced performance. Authorship detection with SVM on full word forms is remarkably robust even if the author writes about a number of different topics such as sports, politics, business etc [39].

III. Evaluating SVMs for Authorship Identification

Business news articles from the Persian Gulf are utilised and SVM experiments on 10 authors are performed [39]. The pre-processing for this experiment consists of two processes: text extraction and feature selection. Text extraction is performed by lexical analysis to strip out all non-word annotations and to convert the text into a list of words or tokens. The pre-processing used in this experiment can be summarised as follows: (1) upper case letters are converted to lower case, (2) all words containing non-letter characters are removed including hyphenated words and words with underscore in the middle of them, (3) all punctuation are replaced with space characters to be treated as token delimiters.

After the text extraction process, a fixed length vocabulary is built from the set of all extracted news articles through a feature selection process. Firstly, stopword removal and stemming are performed on each extracted text. Secondly, document frequency thresholding is used to reduce further the feature vector space. In this experiment, words occurring once only are removed.

After the pre-processing steps, for each class the set of all extracted texts is mapped to one SVM data file of which each line represents a news article. Each line contains a label that indicates whether the article belongs to the class (i.e. the target author to be identified) or not. These data files are used to generate SVM models.

Figure 1 shows a ROC-curve for an SVM trained on one of the authors of business news articles. The performance of the SVM was tested by standard statistical evaluation techniques (cross-validation) and criteria widely used in information retrieval (accuracy, precision and recall). For this particular author, the cross-validation error is 0.69%, the precision 100% and the recall value is 96%. That is, there performance of the SVM on new cases after completion of the learning process is near perfect

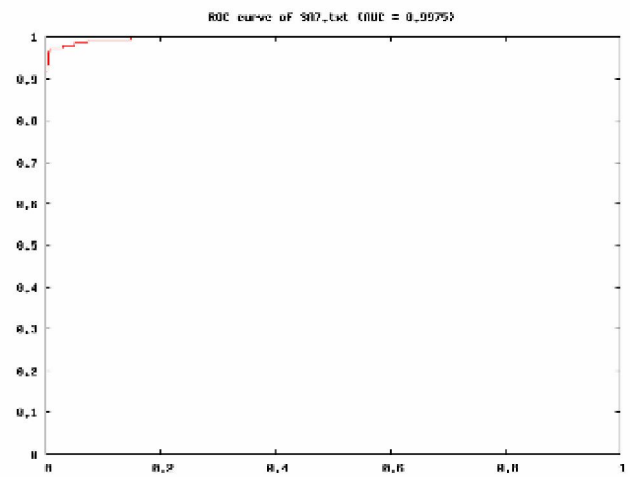

Figure 1: The ROC-curve for an SVM trained on detecting a target author as part of a binary classification task (i.e. target other vs. all other others in the experiment). Since the performance of the SVM is near optimal, false positives do not increase with the number of true positives and *the curve is parallel to the right and top axis* with exception of the top left corner. This is regarded as an indication of very good classification performance.

The ROC-curve (Figure 1) also demonstrates that the classification of documents into categories such as "target author or not" succeeds at a high level. This is consistent with the performance achieved in other experiments [38, 39] and indicates that the technique is suitable for the detection of plagiarism.

IV. Conclusion

Authorship attribution has previously suffered from the problem that the important features in a document are unknown and that a text as a whole cannot be analysed. The use of a limited set of function words or "short words" is clearly restrictive and there is an ongoing discussion on the relevance of appropriate style marker. SVMs for authorship attribution and text mining can process documents of significant length, databases with a large number of texts and they do not require pre-determining relevant features. SVM technology is firmly grounded in computational learning theory and training times compare favourably with other methods such as neural networks. It is therefore proposed to explore SVMs for authorship attribution in the context of plagiarism detection.

V. Acknowledgements

The SVM data file used in the experiment outlined above has been prepared by Insu Song. The ROC curve was provided by Imran Aslam. Other participants in the text classification project are Aqeel Al Ajmi, Jinhan Zhu and Mark Pedersen.

VI. References

1. Andrews, R.; Geva S. (1994), Rule extraction from a constrained error backpropagation MLP. Australian Conference on Neural Networks, Brisbane, Queensland, 9-12, 1994.

2. Andrews, R.; Diederich, J.; Tickle, A.B. (1995). A Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks. Knowledge-Based Systems 8, 373-389.

3. Bolz, N.: Gewinnung und Auswertung quantitativer Merkmale in der statistischen Stilforschung. In: Spillner, B. (Ed.): Methoden der Stilanalyse. Tuebingen: Guenther Naur, 1987.

4. Crain, C.: Donald Foster uses high-powered Computer Tests to Search for Shakespeare's Hidden Hand. His Critics Challenge Him on Every Move. Lingua Franca, 1998.

5. Efron, B.; Thisted, R.: Estimating the number of unseen species: How many words did Shakespeare know? Biometrika 63 (1976) 435- 448.

6. Elman, J.L.: Finding Structure in Time. Cognitive Science 14, 179-211, 1990.

7. Ephratt, M.: Authorship attribution - the case of lexical innovations. Joint International Conference of the Association for Computers and the Humanities and the Association for Literary \& Linguistic Computing, 1997.

8. Fahlman, S.E.: The Recurrent Cascade Correlation Architecture. (Tech. Rep. CMU-CS-91-110). Pittsburgh, PA.: Carnegie Mellon University, 1991.

9. Farringdon, Jill M. Analysing for Authorship: A Guide to the Cusum Technique. Cardiff: University of Wales Press, 1996.

10. Farringdon, J.M.: How to be a Literary Detective: Authorship Attribution. A brief introduction to cusum analysis. QsumIntroduction.html

11. Fish, S.E.: What is stylistics and why are they saying such terrible things about it? In: Freeman, D. (Ed.): Essays in modern stylistics.53-78.

12. Hilton, M.L.; Holmes, D.I.: An Assessment of Cumulative Sum Charts for Authorship Attribution. Literary and Linguistic Computing 8 (1993) 73-80.

13. Holmes, D.I.; Forsyth, R.S.: The Federalist Revisited: New Directions in Authorship Attribution". Literary and Linguistic Computing 10.2 (1995): 111-27.

14. Holmes, D.I.: The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing 13 (1998) 3, 111-117.

15. Joachims, T.: Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, Schoelkopf, B.; Burges, C.; Smola, A. (Ed.), MIT-Press, 1999.

16. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. European Conference on Machine Learning (ECML), Claire Nedellec and Celine Rouveirol (Ed.), 1998.

17. Karlgren, J.; Cutting, D.: Recognizing Text Genres with Simple Metrics using Discriminat Analysis. Coling 1994.

18. Lewis, D.D.: Evaluating Text Categorization. Proceedings of the Speech and Natural Language Workshop. San Mateo: Morgan Kaufmann, 312-318, 1991.

20. Lewis, D.D.: Feature Selection and Feature Extraction for Text Categorization. In: Speech and Natural Language: Proceedings of a Workshop held at Harriman, New York. San Mateo: Morgan Kaufmann, 212-217, 1992.

21. Lewis, D.D.: Text Representation for Intelligent Text Retrieval. In: Jacobs, P.S. (Ed.): Text-Based Intelligent Systems. Lawrence Erlbaum, 1992.

22. Lowe, D.; Matthews, R.: Shakespeare vs. Fletcher: A Stylometric Analysis by Radial Basis Functions. Computers and the Humanities 29 (1995) 449-461.

23. Matthews, R.A.J.; Merriam, T.V.N.: Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher. Literary and Linguistic Computing 8 (1993) 4.

24. Rudman, J.: The State of Authorship Attribution Studies: Some Problems and Solutions. Computers and the Humanities, 31(1998), 351-365.

25. Rudman, J.: Debates in Humanities Computing: Methodology in Authorship Studies. Computers and the Humanities 30:3 (1996). Special issue on statistical methods for authorship attribution.

26. Schellhammer, I.; Diederich, J.; Towsey, M.; Brugman, C. (1998). Knowledge Extraction and Recurrent Neural Networks: An Analysis of an Elman Network trained on a Natural Language Learning Task. Computational Natural

Language Learning Conference. Australian Natural Language Processing Fortnight. Sydney: Macquarie University, 15-17 Jan 1998.

27. Thisted, R.; Efron, B.: Did Shakespeare write a newly discovered poem? Biometrika 74, (1987) 3, 445-55.

28. Towsey, M.; Diederich, J.; Schellhammer, I.; Chalup, S.; Brugman, C. (1998). Natural Language Learning by Recurrent Neural Networks: A Comparison with Probabilistic Approaches. Computational Natural Language Learning Conference. Australian Natural Language Processing Fortnight. Sydney: Macquarie University, 15-17 Jan 1998.

29. Tweedie, F.J.; Singh, S.; Holmes, D.I.: Neural Network Applications in Stylometry: The Federalist Paper. Computers and the Humanities 30 (1996) 1-10.

30. Valenza, R.J.: Are the Thisted-Efron Authorship Tests Valid? Computers and the Humanities, 25(1)(1991), 27-46.

31. Charniak, E.: Statistical Language Learning. MIT Press, Cambridge, Ma., 1993.

32. Zipf, G.K.: Selected studies of the principle of relative frequency in language. Cambridge MA: Harvard University Press, 1932.

33. Zipf, G.K.: Human behaviour and the principle of least effort. An introduction to human ecology; Cambridge; Boston: Houghton-Mifflin, 1935

34. Chitashvili, R.; Baayen, R. H.: Word frequency distributions. In: Altmann, G.; Hvrebicek, L. (Eds): Quantitative Text Analysis; (QL52); wvt: Trier, S.46-135, 1993.

35. Grotjahn, R.: Ein statistisches Modell fuer die Verteilung der Wortlaenge. In: Zeitschrift fuer Sprachwissenschaft 1, 44-75, 1982.

36. Herdan, G.: The advanced theory of language as choice and chance. Springer: Berlin, Heidelberg, New York, 1966.

37. Salton, G.; McGill, M. J.: Introduction to Modern Information Retrieval; McGraw Hill: New York et al., 1983.

38. Diederich, J., Kindermann, J., Leopold, E., Paass, G., Authorship Attribution with Support Vector Machines. *Applied Intelligence,* Vol. 19, No. 1 (July/Aug, 2003) 109–123.

39. Diederich, J., Song, I., Al-Ajmi, A., Zhu, J., Authorship attribution and search engine technology. Kasabov, N., Chan, Z.S.H. (Eds.), Proceedings of the Conference on Neuro-Computing and Evolving Intelligence, Auckland, New Zealand (13-15 December 2004). Auckland: Knowledge Engineering and Discovery Research Institute (KEDRI) (2004).

Joachim Diederich is serving as Professor and Chair, Department of Computer Science, American University of Shrajah, U.A.E. He is also an Honorary Professor in the School of Information Technology and Electrical Engineering as well as the Centre for Online Health at the University of Queensland, Australia. Dr. Diederich 's qualifications include a Habilitation from the University of Hamburg (Germany), a Doctorate (summa cum laude) from the University of Bielefeld (Germany), and a Master's degree from the University of Münster (Germany).

His professional career includes four years as a team leader at the German National Research Center for Information Technology (GMD) and more than two years at the International Computer Science Institute (ICSI) in Berkeley, California. He has also been a Full Professor at Queensland University of Technology (Australia) and Sohar University (Oman; affiliated with the University of Queensland).

Dr. Diederich has published eight books (three as single author), 50 book chapters and journal articles, and more than 100 peer-reviewed conference papers. As a recognized innovator, Dr. Diederich was selected as winner of the E-Health category in the prestigious "Secrets of Australian IT Innovation Competition 2003."