Gender-Preferential Text Mining of E-mail Discourse

Malcolm Corney[1] Olivier de Vel[2] Alison Anderson[1] George Mohay[1]

(1) Faculty of Information Technology
Queensland University of Technology
Brisbane QLD 4001
Australia
m.corney, a.anderson, g.mohay @qut.edu.au

(2) Defence Science and Technology Organisation
P.O. Box 1500
Edinburgh SA 5111
Australia
Olivier.deVel@dsto.defence.gov.au

Abstract

This paper describes an investigation of authorship gender attribution mining from e-mail text documents. We used an extended set of predominantly topic content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features together with a Support Vector Machine learning algorithm. Experiments using a corpus of e-mail documents generated by a large number of authors of both genders gave promising results for author gender categorisation.

1. Introduction

With the rise in the use of computers and computer networks for illegal activities (e.g., fraud, money laundering etc.), the area of computer forensics has become increasingly important. Computer forensics has rapidly evolved over the past few years and with a range of different end-applications (e.g., data recovery, law enforcement, e-commerce), with each application having different requirements. For example, in the traditional law enforcement area, the primary focus is the prosecution of the perpetrator. Once the crime has been perpetrated, the *post-mortem* collection and preservation of the chain of custody of evidence, data analysis, interpretation etc. are undertaken subject to strict established prosecutorial guidelines. On the other hand, e-commerce is more concerned with the continual availability of the on-line business service, so that the focus is on the timeliness of the cycle of detection, forensic analysis and reaction.

Computer forensics investigations have to increasingly deal with e-mail as this is becoming an important form of communication for many computer users, for both legitimate and illegitimate activities. E-mail is used in many legitimate activities such as message and document exchange. Unfortunately, it can also be misused, for example, in the distribution of unsolicited junk mail, unauthorised conveyancing of sensitive information, mailing of offensive or threatening material. E-mail evidence can be central in cases of sexual harassment or racial vilification, threats, bullying and so on.

Some researchers have stated that e-mail is much like spoken communication. However, there are some important differences. For example, e-mail is more rarefied than normal spoken communication. With e-mail, participants cannot see each other's faces, hear each other's voices, or identify gestures or other visual cues. The information content in an e-mail can include simple text as well as mark-up text to convey additional information. Some senders of e-mail use only natural language text to formulate the content of the transmitted information, other users have developed an electronic "para-language" to mark-up their message and convey affective and socio-emotional information. Such informal language codes, called "emotext," include intentional misspelling (e.g., "u r ssoooo kooool"), lexical surrogates for vocalisations (e.g., "hmm"), grammatical markers (e.g., excessive use of upper-case letters, repeated question marks), and visual arrangements of text characters into "emoticons" (short combinations of normal and rotated characters to resemble facial expressions of joy, sadness etc.).

In this paper we are particularly interested in determining the gender of the author of an e-mail, based on the gender-preferential language used by the author. The paper is organised as follows. Firstly, we outline the current status of work in the area of author attribution in Section 2. We then focus our discussion on gender-preferential e-mail mediated communication in Section 3. Sections 4 and 5 briefly outline the Support Vector Machine learning algorithm used in our experiments, describe the e-mail corpus used, and present the methodology employed in the experiments. Validation of the method is then undertaken by presenting results of gender-based e-mail categorisation performance in Section 6. Finally, we conclude with some gen-

Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC'02)
1063-9527/02 $17.00 © 2002 IEEE

eral observations and present future directions for the work in Section 7.

2. Background to Author Attribution

The principal objectives of author gender attribution are to classify an ensemble of e-mails as belonging to a particular author gender and, if possible, obtain a set of characteristics or features that remain relatively constant for a large number of e-mails written by that gender cohort of authors. The question then arises; can characteristics such as language, structure, layout etc. of an e-mail be used, with a high degree of confidence, as a kind of author cohort phrenology and thus link the e-mail document with its author cohort? Also, can we expect the writing characteristics or style of an author cohort to evolve in time and change in different contexts? For example, the composition of formal e-mails will differ from informal ones (changes in vocabulary etc.). Even in the context of informal e-mails there could be several composition styles (e.g., one style for personal relations and one for work relations). However, humans are creatures of habit and have certain personal traits which tend to persist. All humans have unique (or near-unique) patterns of behaviour, biometric attributes, and so on. We therefore conjecture that certain characteristics pertaining to language, composition and writing, such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage (e.g., converting the letter "f" to "ph", or the excessive use of digits and/or upper-case letters), stylistic and sub-stylistic features will remain relatively constant. The identification and learning of these characteristics with a sufficiently high accuracy are the principal challenges in author cohort categorisation.

Related, but separate, areas of author cohort attribution are text categorisation and authorship attribution. The former attempts to categorise a set of text documents based on its contents or topic whilst the latter attempts to identify the author of the e-mail. Many methods have been proposed for text categorisation. Most of these techniques employ the "bag–of–words" or word vector space feature representation and use a learning algorithm such as decision trees [1], Bayesian probabilistic approaches [2], or support vector machines [3] to classify the text document. Work in e-mail text classification has also been undertaken by some researchers in the context of automated e-mail document filtering and filing (see, for example. [4]). Authorship attribution studies are also extensive and often controversial (for example, the authorship of the Federalist papers [5] and Shakespeare's works [6]). Almost all of these studies employ stylometric features ("style markers") for discriminating authors and all use large, formal texts as the source of documents. Over 1,000 stylometric features have been proposed [7], including word- or character-based stylometric features, function words, profanities, punctuation etc. Also, there exists a number of different techniques for performing the discrimination. These include statistical approaches (e.g., cusum [8], neural networks [9] and so on. Unfortunately, there does not exist a consensus on the existence of a set of uniquely discriminatory stylometric features, nor on a correct methodology as many of the mentioned techniques suffer from problems such as questionable analysis, inconsistencies for the same set of authors, failed replication etc.

A small number of studies in e-mail authorship attribution have been undertaken. Corney *et al* [10] used a set of stylometric and e-mail structural features and also studied the effect of text size and the number of e-mail documents per author on the author categorisation performance. They observed a relatively constant categorisation performance for text chunk sizes greater than approximately 100 words with, however, a significant drop-off for text sizes less than this. Also, they observed that as few as 20 documents may be sufficient for satisfactory categorisation performance. de Vel *et al* achieved satisfactory results with multi-topic and multi-author categorisation using a set of predominantly content-free e-mail document features such as structural characteristics and linguistic patterns [11].

3. Gender-Preferential E-mail Mediated Communication

Although computer-mediated communication (CMC) does inhibit some cues such as personal identity or individuating details (e.g., dress, location, demeanour, expressiveness), there is no evidence to suggest that all other cues are also inhibited. With e-mail mediated communication, some information about social categories or social identity, such as gender, or educational or first language background cues are likely to be inferred in the relative absence of interpersonal context cues [12].

Men and women use language and converse differently even though they technically speak the same language. Empirical evidence suggests that there exist gender differences in written communication, face-to-face interaction and in computer-mediated communication. It is thought that gender-preferential language is conveyed in all of these forms of communication due, in part, to the use of intersecting or generalised gender-preferential language attributes. Many studies have been undertaken on the issue of gender and language use (for example, see the bibliography at [13]). It has been suggested by various researchers that women's language makes more frequent use of emotionally intensive adverbs and adjectives such as "so", "terribly", "awfully", "dreadful" and "quite" and that their language is more punctuated with attenuated assertions, apologies, questions, personal orientation and support". On the other hand, male conversational patterns express "independence"

and assertions of vertically hierarchical power. Men are more "proactive" by directing speech at solving problems while women are more "reactive" to the contributions of others, agreeing, understanding and supporting. Some features of men's language are "strong assertions, aggressive, self-promotion, rhetorical questions, authoritative orientation, challenges and humor". In brief, men's on-line conversation resemble "report talk", rather than "rapport talk" which women tend to favour.

Many gender-preferential CMC studies have been undertaken in recent years. However, very few studies in the area of e-mail CMC have been performed (for example, Thomson *et al* [12]) and no studies, to the authors' knowledge, in automated e-mail gender-preferential author cohort attribution have been undertaken to date. In our study we use a combination of stylometric, structural and gender-preferential features, together with a Support Vector Machine classifier as the learning algorithm.

4. Support Vector Machine Classifier

The Support Vector Machine's (SVM) concept is based on the idea of structural risk minimisation which minimises the generalisation error (i.e. true error on unseen examples). This true error is bounded by the sum of the training set error and a term which depends on the Vapnik-Chervonenkis (VC) dimension of the classifier and on the number of training examples. SVMs belong to the class of the more general basis expansion and regularisation problem to which methods such as smoothing splines, multidimensional splines (eg, MARS, wavelet smoothing) belong. One advantage of SVMs is that they do not require a reduction in the number of features in order to avoid the problem of over-fitting, which is useful when dealing with large dimensions as encountered in the area of text mining. See [14] for more background information on SVMs.

Some researchers have applied SVMs to the problem of text document categorisation and author attribution concluding that, in most cases, SVMs outperform conventional classifiers (see, for example, [3]). SVMs have been used for automatic filing of e-mails as well as for classifying e-mail text as spam or non-spam [15][16].

5. E-mail Corpus and Methodology

We describe the process of generating the e-mail corpus and the selection of attributes for the gender-preferential language author categorisation experiment. We also briefly describe the sampling methodology used and calculation of the categorisation performance.

5.1. E-mail Corpus Generation

The generation of a suitable corpus of e-mails for the study was complicated by various factors. Firstly, the process of generating any e-mail corpus is constrained by privacy issues and ethical considerations. It is not possible to use e-mails from other people's inboxes without their consent. Unfortunately, obtaining a person's consent is an almost impossible exercise. Secondly, even though it is possible to use publicly available e-mail corpuses such as newsgroups, mailing lists etc., it is not always easy to validate the gender of the sender of each e-mail in the corpus. For example, it is not sufficient to use the sender's name as this could be an alias, indeterminate, spoofed etc.. Thirdly, it is generally difficult to obtain a sufficiently large and "clean" (i.e., void of cross-postings, off-the-topic spam, empty bodied e-mails with attachments etc.) corpus of e-mails. Finally, it is important not to generate an e-mail corpus that is biased towards, for example, a particular cohort (e.g., author's language style) or e-mail topic as these may affect the categorisation results of the gender-preferential language author attribution experiment. A judicious, and time-consuming, selection of e-mails for model building is therefore paramount.

The corpus of e-mail documents used in the experimental evaluation of the gender author categorisation study was sourced from the inbox of a member of a large (greater than 15,000 users) academic organisation[1]. The senders of the e-mail messages were selected based on the fact that they belonged to the organisation and their gender easily checked. All other senders (external) were not considered as it was not possible to confirm their gender reliably. Any cross-postings, re-quoted spammed e-mails (e.g., jokes, stories), general notification or broadcast e-mails relating to the organisation etc. were purged from the corpus. An initial total of 8820 e-mail documents sourced from 342 authors (approx. equally distributed between the two genders) were selected. The gender of each author was confirmed for all e-mail documents. This document set was subsequently pared down to 4369 e-mail messages (for 325 authors) to ensure only email messages with a minimum number of words equal to 50 are used (see [10] for suggested guidelines on the choice of e-mail document size). The body of each e-mail document was then parsed using an e-mail grammar, and the relevant e-mail body features were extracted. The body was pre-processed to remove (if present) any salutations, reply text and signatures. However, the existence, position within the e-mail body and type of some of these were retained as inputs to the categoriser (see below).

[1] In order to preserve anonymity, all third parties (such as any member of the DSTO) that were involved in the experiment were only presented with the summary statistics of the experiment and not with the contents of the e-mails in the corpus.

Table 1. Summary statistics of some of the e-mail corpus used in the experiment for gender-based author cohorts.

Minimum Number of Words	*Number of Authors*		
	Male Cohort	*Female Cohort*	*Total*
50	117	208	325
100	104	176	280
150	91	135	226
200	83	99	182

Minimum Number of Words	*Number of E-mails*		
	Male Cohort	*Female Cohort*	*Total*
50	2071	2298	4369
100	1257	1072	2329
150	842	585	1427
200	564	384	948

Attachments were excluded, though the e-mail body itself was used.

In order to study the impact of the number of words in an e-mail on the categorisation performance (see later), the e-mail corpus was further divided into multiple subsets. The subsets were generated by first creating a root-level subset with a minimum number of 50 words per e-mail, and then recursively generating lower-level subsets from their parent subsets with a minimum of 100, 150, 200 etc. words per e-mail. A summary of some of the e-mail document corpus statistics measured in terms of the number of authors in each gender cohort and the number of e-mails as a function of the minimum number of words per e-mail, is shown in Table 1.

5.2. Attribute Selection

The attributes/features selected for the experiment were members of two sets namely, a baseline set and a gender-specific set. The total number of attributes used in the experiment was 222.

A baseline set of attributes/features that was identified in previous authorship attribution experiments (see [10][17][11]) for e-mail authorship discrimination was extracted from each e-mail body document. These attributes included both a mix of character- and word-based style markers as well as structural features. A total of 211 baseline attributes, comprising 183 style marker attributes and 28 structural attributes, were employed in the experiment (see Table 2). Note that = total number of *tokens* (i.e., words), = total number of *types* (i.e., distinct words), C = total number of characters, and = total number of HTML tags in the e-mail body. Also, attribute $_{21}$ is the total number of characters in words, including apostrophes and hyphens, divided by C. The *hapax legomena* count is defined as the number of types that occur only once in the e-mail text. Attributes $_8$ to $_{20}$ are defined in Tweedie *et al* [7]. For example, Rubet's K value is computed as ( )/ ( ).

We briefly clarify how we derive some of the attributes shown in Table 2. Firstly, the set of short words in each e-mail document consists of all words of length less than or equal to 3 characters (e.g., "all", "at", "his" etc.). Only the total count of short words is used as a feature. The short word frequency distribution may be biased towards e-mail content and was therefore not used in our experiments. Secondly, the set of all-purpose function words ("a", "about", "after", "all", "also", ..., "yet", "you", "your", "yours") and its frequency distribution is obtained and also used as a sub-vector attribute. The number of function words used is 122. Finally, a word length frequency distribution consisting of 30 features (up to a maximum word length of 30 characters) is employed.

The re-quoted text position refers to the reply status of e-mail. A reply text can generally be placed in any position in the e-mail document and each line is usually prefixed with a special character (e.g., " "). In our experiment, the position of re-quoted text allowed for 6 different possibilities (e-mail body text interspersed with the re-quoted text, e-mail body text preceded by re-quoted text etc.). Due to some e-mailers using HTML formatting, we include the set of HTML tags as a structural metric. The frequency distribution of HTML tags was included as one of the 28 structural attributes.

The set of basic gender-specific language attributes were selected from the literature presented in Section 3. These are listed in Table 3 (attributes $_{211}$ to $_{221}$). The selected attributes attempt to measure the frequency of use of adjectives, adverbs (mainly through the presence of suffixes) and apologies. This attribute set is a small subset of possible gender-preferential language attributes listed in the literature.

Though our choice of attributes is specifically biased towards features that have been shown to be able to effectively discriminate between authors and, hopefully, between author gender, rather than discriminating between topics, some of the style marker attributes may have a combination of author and content bias as, for example, *hapax legomena* as defined in attributes $_6$ and $_7$ in Table 2 (see [18]).

Each attribute is also scaled as follows:

Table 2. E-mail document body style marker and structural attributes.

Attribute Type, ($= 0, \ldots, 10$)	
Document-based:	
A_0:	Number of blank lines/total number of lines
A_1:	Average sentence length (number of words)
Word-based:	
A_2:	Average word length
A_3:	Vocabulary richness i.e., V/M
A_4:	Number of function words/M
A_5:	Number of short words/M (word length 3)
A_6:	Count of hapax legomena/M
A_7:	Count of hapax legomena/V
A_8:	Guirad's R
A_9:	Herdan's C
A_{10}:	Herdan's V
A_{11}:	Rubet's K
A_{12}:	Maas' A
A_{13}:	Dugast's U
A_{14}:	Lukjanenkov and Neistoj's measure
A_{15}:	Brunet's W
A_{16}:	Honore's H
A_{17}:	Sichel's S
A_{18}:	Yule's K
A_{19}:	Simpson's D
A_{20}:	Entropy measure
Character-based:	
A_{21}:	Number of characters in words/C (see text)
A_{22}:	Number of alphabetic characters/C
A_{23}:	Number of upper-case characters in words/C
A_{24}:	Number of digit characters in words/C
A_{25}:	Number of white-space characters/C
A_{26}:	Number of spaces/C
A_{27}:	Number of spaces/Number white-space chars
A_{28}:	Number of tab spaces/C
A_{29}:	Number of tab spaces/Number white-space chars
A_{30}:	Number of punctuation characters/C
Function Words:	
$A_{31\ldots152}$:	Function word frequency distribution (122 features)
Other:	
$A_{153\ldots182}$:	Word length frequency distribution/M (30 features)
Structural:	
A_{183}:	Reply status
A_{184}:	Has a greeting acknowledgement
A_{185}:	Uses a farewell acknowledgement
A_{186}:	Contains signature text
A_{187}:	Number of attachments
A_{188}:	Position of re-quoted text within e-mail body
$A_{189\ldots210}$:	HTML tag frequency distribution/H (22 features)

Table 3. E-mail document gender-preferential language attributes.

Attribute Type, ($= 11, \ldots, 1$)	
Gender-Preferential:	
A_{211}:	Number of words ending with *able* /M
A_{212}:	Number of words ending with *al* /M
A_{213}:	Number of words ending with *ful* /M
A_{214}:	Number of words ending with *ible* /M
A_{215}:	Number of words ending with *ic* /M
A_{216}:	Number of words ending with *ive* /M
A_{217}:	Number of words ending with *less* /M
A_{218}:	Number of words ending with *ly* /M
A_{219}:	Number of words ending with *ous* /M
A_{220}:	Number of *sorry* words /M
A_{221}:	Number of words starting with *apolog* /M

$$^{(\text{scaled})} = (\quad - \quad _{\min}) S \quad _i \quad _i$$

so as to ensure all attributes are treated equally in the classification process. The scaling factor, $S \quad _i$, is computed as:

$$S \quad _i = \frac{_i - \quad _i}{_{\max} - \quad _{\min}}$$

with $_{\min}$ and $_{\max}$ being the minimum and maximum values of the attribute , respectively. Also, $_i$ and $_i$ are the defined lower and upper bounds of the scaled attribute, respectively (we have used $_i = 0.0$ and $_i = 1.0$).

5.3. Performance Evaluation Methodology

The ^{ht} Support Vector Machine classifier developed by T. Joachims from the University of Dortmund [19] was used in the experiments. ^{ht} is an implementation of Vapnik's Support Vector Machine [14], as described in Section 4 . It (^{ht}) scales well to a large number of sparse instance vectors as well as efficiently handling a large number of support vectors. In our experiments we explored a number of different kernel functions for the SVM classifier namely, the linear, polynomial, radial basis and sigmoid *tanh* functions. We obtained maximal $_1$ classification results (see below for the definition of $_1$) on our data set with a polynomial kernel of degree 3. The "LOQO" optimiser was used for maximising the margin.

The Support Vector Machine computes two-way categorisation. Therefore, in our experiments on author gender categorisation, only a single two-way classification model with a two-way confusion matrix needed to be generated.

Table 4. $_1$ categorisation performance results (in %) for different e-mail document sizes and for different e-mail cohort sizes. See text for explanation. Values indicated by "-" correspond to insufficient e-mail document size/word count population.

Number of E-mails per Gender Cohort Class	*Minimum Word Count*			
	50	100	150	200
50	64.4	62.2	57.1	59.8
100	68.4	64.0	56.8	65.0
200	64.8	61.5	62.2	63.8
300	66.4	67.6	66.6	67.3
400	67.5	68.7	70.2	-
1000	69.4	71.1	-	-

The training-testing sampling methodology used was a 10-fold cross-validation of the entire e-mail document set.

To evaluate the categorisation performance on the e-mail document corpus, we calculate the accuracy, recall (R), precision (P) and combined $_1$ performance measures commonly employed in the information retrieval and text mining literature (for a discussion of these measures see, for example, [20]), where:

$$_1 = \frac{RP}{(R \quad P)}$$

6. Results and Discussion

We present our author gender-preferential language attribution results and report the $_1$ statistic using the Support Vector Machine (SVM) classifier. The results are given for different e-mail document sizes (measured as the minimum word count) and for different e-mail author gender cohort sizes (number of e-mail documents per female and male author cohort). These are displayed in Table 4.

As observed in Table 4, $_1$ categorisation performance results indicate that, in general, the SVM classifier combined with the style markers, structural attributes, and gender-preferential language attributes is able to satisfactorily discriminate between the author gender cohorts. As expected, there is a general improvement, though not dramatic, in performance as the the number of e-mails in each gender cohort class increases. However, the improvement in performance as a function of the minimum word count is not as consistent as the e-mail count performance results. A noticeable improvement is only achieved when the number of e-mails in each gender cohort class is not too small

Table 5. Effect of the attribute type on the $_1$ categorisation performance results.

Feature Set Type	*Operation*	$_1$ (%)
Character-based attributes	Removed	70.0
Word-based attributes	Removed	69.6
Word length distribution	Removed	67.4
Structural attributes	Removed	68.1
Function words	Removed	64.0
All baseline attributes	-	70.1
All attributes (baseline + gender-based)	-	70.2

(00) and/or the number of authors in each author gender class increases. These results indicate that a small number of e-mails per author cohort class is generally sufficient for satisfactory gender classification. This result compares favourably with similar observations made in authorship attribution studies [10].

Some preliminary analysis of the impact of the different types of attributes (stylistic, structural, gender-preferential) on the author gender categorisation performance was also undertaken. Each type of attribute set was removed from the feature set and the performance results calculated. These are shown in Table 5.

Though preliminary at this stage, the results in Table 5 show that the full combination of attributes gives the best author gender categorisation. Removal of any of the attributes gives rise to a reduced performance value, though some more importantly than others. In particular, the set of function words (attributes $_{31}$ $_{152}$) are seen to be an important gender discriminator. This is as expected since function words has been shown to be a good author discriminator [11] as well as containing words that could belong to gender-preferential language (such as "so", "very" etc.). However, we also note that the gender-preferential attributes used in the experiment only give a marginal improvement in the categorisation performance. This indicates that the current set of gender-based attributes are insufficient and a more selective and/or more extensive set of gender-preferential attributes will need to be used to achieve better categorisation performance.

7. Conclusions

In this paper, we have investigated the learning of the author gender categories from e-mail documents. We used an extended set of predominantly content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features together with

a Support Vector Machine learning algorithm. Experiments on a number of e-mail documents generated by over 800 authors of both genders gave promising results for author gender categorisation. We observed an improvement in performance with increasing number of e-mails in both gender cohort classes.

The current approach has several limitations. Firstly, as mentioned in Section 6, a larger set of gender-preferential language attributes needs to be used to improve the performance results further. Secondly, more studies on the usefulness of specific style markers, such as -graphs, for author gender identification should be investigated as it is conjectured that, for example, certain bi-graphs incorporating punctuation could be effective discriminators [21]. Finally, the diversity in author characteristics in the author cohort e-mail database is currently quite small owing to the type of organisation where the e-mails were sourced. Though it is not easy to obtain a sufficiently large set of e-mails from authors with varying cohort characteristics (educational level, language background etc.), we hope to be able to build up a suitable forensic database and further test our approach.

References

[1] C. Apte, F. Damerau and S. Weiss, "Text Mining with Decision Rules and Decision Trees", *Workshop on Learning from Text and the Web, Conference on Automated Learning and Discovery*, Pittsburgh, PA, 1998.

[2] Y. Yang and X. Liu, "A Re-examination of Text Categorisation Methods", *Proc. 22nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR99)*, Berkeley, CA, 1999, pp. 67–73.

[3] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", *Proc. European Conf. Machine Learning (ECML'98)*, Chemnitz, Germany, 1998, pp. 137–142.

[4] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, "A Bayesian Approach to Filtering Junk E-mail", *Learning for Text Categorization Workshop: 15th National Conf. on AI. AAAI Technical Report (WS-98-05)*, Madison, WI, 1998, pp. 55–62.

[5] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, MA, 1964.

[6] E. Elliot and R. Valenza, "Was the Earl of Oxford the True Shakespeare?", *Notes and Queries*, **38**, 1991, pp. 501–506.

[7] F. Tweedie and R. Baayen, "How Variable May a Constant Be? Measures of Lexical Richness in Perspective", *Computers and the Humanities*, **32**(5), 1998, pp. 323–352.

[8] J. Farringdon, *Analysing for Authorship: A Guide to the Cusum Technique*, University of Wales Press, Cardiff, 1996.

[9] F. Tweedie, S. Singh and D. Holmes, "Neural Network Applications in Stylometry: The Federalist Papers", *Computers and the Humanities*, **30**(1), 1996, pp. 1–10.

[10] M. Corney, A. Anderson, G. Mohay and O. de Vel, "Identifying the Authors of Suspect E-mail", *Computers and Security*, in press, 2001.

[11] O. de Vel, A. Anderson, M. Corney and G. Mohay, "E-mail Authorship Attribution for Computer Forensics", in D. Barbara and S. Jajodia, *Data Mining for Security Applications*, Kluwer Academic Publishers, Boston, MA, 2002.

[12] R. Thomson and T. Murachver, "Predicting Gender from Electronic Discourse", *British Journal of Social Psychology*, **40**, 2001, pp. 193–208.

[13] H. Schiffman, "Bibliography of Gender and Language", July 2002. http://ccat.sas.upenn.edu/ haroldfs/ popcult/bibliogs/gender/genbib.htm, 2002.

[14] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[15] H. Druker, D. Wu and V. Vapnik, "Support Vector Machines for Spam Categorisation", *IEEE Trans. on Neural Networks*, **10**, 1999, pp. 1048–1054.

[16] J. Diederich, J. Kindermann, E. Leopold and G. Paass, "Authorship Attribution with Support Vector Machines", *Applied Intelligence*, 2000, submitted.

[17] O. de Vel, "Mining E-mail Authorship", *Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000)*, Boston, MA, 2000.

[18] Chaski, C., "A Daubert-Inspired Assessment of Current Techniques for Language-Based Author Identification", US National Institute of Justice, 1998. http://www.ncjrs.org

[19] T. Joachims, "Making Large-Scale SVM Learning Practical", in B. Scholkopf, C. Burges and A. Smola, *Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 1999.

[20] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, CA, 2000.

[21] C. Chaski, "Empirical Evaluations of Language-Based Author Identification Techniques", *Forensic Linguistics*, **8**(1), 2001.