

Visualizing Authorship for Identification

Ahmed Abbasi and Hsinchun Chen

Department of Management Information Systems, The University of Arizona,
Tucson, AZ 85721, USA
aabbasi@email.arizona.edu, hchen@eller.arizona.edu

Abstract. As a result of growing misuse of online anonymity, researchers have begun to create visualization tools to facilitate greater user accountability in online communities. In this study we created an authorship visualization called *Writeprints* that can help identify individuals based on their writing style. The visualization creates unique writing style patterns that can be automatically identified in a manner similar to fingerprint biometric systems. *Writeprints* is a principal component analysis based technique that uses a dynamic feature-based sliding window algorithm, making it well suited at visualizing authorship across larger groups of messages. We evaluated the effectiveness of the visualization across messages from three English and Arabic forums in comparison with Support Vector Machines (SVM) and found that *Writeprints* provided excellent classification performance, significantly outperforming SVM in many instances. Based on our results, we believe the visualization can assist law enforcement in identifying cyber criminals and also help users authenticate fellow online members in order to deter cyber deception.

1 Introduction

The rapid proliferation of the Internet has facilitated the increasing popularity of computer mediated communication. Inevitably, the numerous benefits of online communication have also allowed the realization of several vices. The anonymous nature of the Internet is an attractive medium for cybercrime; ranging from illegal sales and distribution of software [11] to the use of the Internet as a means of communication by extremist and terrorist organizations [20, 1].

In addition to using the internet as an illegal sales and communication medium, there are several trust related issues in online communities that have surfaced as a result of online anonymity [13]. Internet-based deception is rampant when interacting with online strangers [5]. With widespread cybercrime and online deception, there is a growing need for mechanisms to identify online criminals and to provide authentication services to deter abuse of online anonymity against unsuspecting users.

We propose the use of authorship visualization techniques to allow the identification and authentication of online individuals. In order to accomplish this task we developed a visualization called *Writeprints*, which can automatically identify authors based on their writing style. Due to the multilingual nature of online communication and cybercrime, the visualization was designed to handle text in multiple languages. *Writeprints* is adept at showing long-term author writing patterns over larger quantities of text. We tested the

effectiveness of the technique on a test bed consisting of messages from three web forums composed of English and Arabic messages. Our results indicate that *Writeprints* can provide a high level of accuracy and utility which may greatly aid online users and law enforcement in preventing online deception and cybercrime.

2 Related Work

2.1 Visualizing Authors in Online Communities

Kelly et al. [8] suggested the notion that collecting user activity data in online communities and feeding it back to the users could lead to improved user behavior. Erickson and Kellogg [7] also contended that greater informational transparency in online communities would likely lead to increased user accountability. Due to the copious amounts of data available in online forums and newsgroups, information visualization techniques can present users with relevant information in a more efficient manner [17].

There have been several visualizations created using participant activity information for the purpose of allowing online users to be more informed about their fellow members [7, 14, 6, 17]. Most of the author information provided by each of these visuals is based on their interaction patterns derived from message threads. Hence, there is little evaluation of author message content. From the perspective of cybercrime and online deception, viewing author posting patterns alone is not sufficient to safeguard against deceit. Individuals can use multiple usernames or copycat/forge other users with the intention of harassing or deceiving unsuspecting members. Thus, there is also a need for authentication techniques based on message content. Specifically, authorship visualizations can provide additional mechanisms for identification and authentication in cyberspace.

2.2 Authorship Analysis

Authorship Analysis is grounded in Stylometry, which is the statistical analysis of writing style. Currently authorship analysis has become increasingly popular in identification of online messages due to augmented misuse of the Internet. De Vel et al. [4] applied authorship identification to email while there have been several studies that have applied the techniques to web forum messages [20, 1, 10].

Online content poses problems for authorship identification as compared to conventional forms of writing (literary works, published articles). Perhaps the biggest concern is the shorter length of online messages. Online messages tend to be merely a couple of hundred words on average [20] with great variation in length. In light of the challenges associated with cyber content, the ability to identify online authorship signifies a dramatic leap in the effectiveness of authorship identification methodologies in recent years. Much of this progress can be attributed to the evolution of the two major parameters for authorship identification which are the writing style markers (features) and classification techniques incorporated for discrimination of authorship.

2.2.1 Authorship Features

Writing style features are characteristics that can be derived from a message in order to facilitate authorship attribution. Numerous types of features have been used in

previous studies including n-grams and the frequency of spelling and grammatical errors, however four categories used extensively for online material are lexical, syntactic, structural, and content specific features [20].

Lexical features include total number of words, words per sentence, word length distribution, vocabulary richness, characters per sentence, characters per word, and the usage frequency of individual letters. *Syntax* refers to the patterns used for the formation of sentences. This category of features is comprised of punctuation and function/stop words. *Structural* features deal with the organization and layout of the text. This set of features has been shown to be particularly important for online messages [4]. Structural features include the use of greetings and signatures, the use of phone numbers or email addresses for contact information, and the number of paragraphs and average paragraph length. *Content-specific* features are key words that are important within a specific topic domain.

2.2.2 Authorship Techniques

The two most commonly used analytical techniques for authorship attribution are statistical and machine learning approaches. Many multivariate statistical approaches such as principal component analysis [2, 3] have been shown to provide a high level of accuracy. Statistical techniques have the benefit of providing greater explanatory potential which can be useful for evaluating trends and variances over larger amounts of text.

Drastic increases in computational power have caused the emergence of machine learning techniques such as support vector machines, neural networks, and decision trees. These techniques have gained wider acceptance in authorship analysis studies in recent years [16, 4]. Machine learning approaches provide greater scalability in terms of the number of features that can be handled and are well suited to cope with the shorter lengths of online messages. Specifically, SVM has emerged as perhaps the most powerful machine learning technique for classification of online messages. In comparisons, it outperformed other machine learning techniques for classification of web forum messages [20, 1].

2.3 Authorship Visualization

There has been a limited amount of work on authorship visualization. Kjell et al. [9] used statistical techniques such as principal component analysis (PCA) and cosine similarity to visualize writing style patterns for Hamilton and Madison's Federalist papers. The study created writing style patterns based on usage frequencies of the ten n-grams with the greatest variance between the two authors. Shaw et al. [15] used latent semantic indexing (LSI) for authorship visualization of biblical texts based on n-gram usage. Their visualization tool, called SFA, allows authorship tendencies to be viewed based on visualization of eigenvectors (principal components) in multidimensional space. Ribler and Abrams [12] used an n-gram based visualization called Patterngrams to compare the similarity between documents for plagiarism detection in student computer programs.

There are several important things to note about these studies: (1) they all used n-gram features to discriminate authorship, (2) they all used manual observation to evaluate the visualizations, (3) none of them were applied to online media, and (4) there is no indication of whether the techniques can be successfully applied in a multi-lingual setting.

3 Visualization Requirements

Based on gaps in previous visualization tools created for online communities and authorship analysis, we have identified some requirements for our authorship visualizations. We would like to create visualizations that can automatically identify authors based on writing style patterns, using a set of features specifically designed and catered towards identification of authorship in cyberspace. A detailed description of our requirements is given below:

3.1 Visualizing Cyber Authorship

As we previously mentioned, there has been significant work relating to the creation of visualizations to improve user awareness and accountability in online settings. However, there have not been any visualization tools created specifically with the intention of diminishing cybercrime and online deception. We believe that there is a dire need for such tools and that authorship visualizations can facilitate identification of cyber criminals and lessen online deception.

3.2 Automatic Recognition

Previous authorship visualization studies used observation to determine authorship similarity. Manual inspection was sufficient since these studies only compared a few authors; however this is not possible for online authorship identification. It is infeasible to visually compare large numbers of writing style patterns manually. In addition to being overly time consuming, it is beyond the boundaries of human cognitive abilities. To overcome these deficiencies, Li et al. [10] called for the creation of authorship visualizations that can be automatically identified in a manner analogous to fingerprint biometric systems.

3.3 Online Authorship Features

Previous studies have used n-grams. It is unclear whether or not such an approach would be scalable when applied to a larger number of authors in an online setting. In contrast, lexical, syntactic, structural, and content-specific features have demonstrated their ability to provide a high level of discriminatory potential in online settings [4] featuring multilingual message corpora [20, 1]. Thus, the authorship visualizations created for identification of online messages should incorporate a feature set encompassing these feature types.

4 Process Design

We propose the creation of a visualization techniques designed for authorship identification and authentication called *Writeprints*. The visualization uses dimensionality reduction to create a more coherent representation of authorship style. *Writeprints* is a principal component analysis based visualization technique that uses a dynamic feature-based sliding window algorithm. Principal component analysis is a powerful

technique that has been used in several previous authorship analysis and visualization studies [3, 9]. The *Writeprint* visualization is aimed at providing greater power for lengthier amounts of text by visualizing the writing style variation, which can create powerful patterns that are highly differentiable.

Figure 1 provides a description of the processes entailed towards the creation of *Writeprints* beginning with the collection of messages and extraction of writing style features from the collected content. The authorship visualizations are created by transforming the feature usage values into writing style patterns using dimensionality reduction techniques and specific algorithms designed to accentuate the important writing style patterns of the various authors. These patterns can uniquely identify authors using automatic recognition techniques.

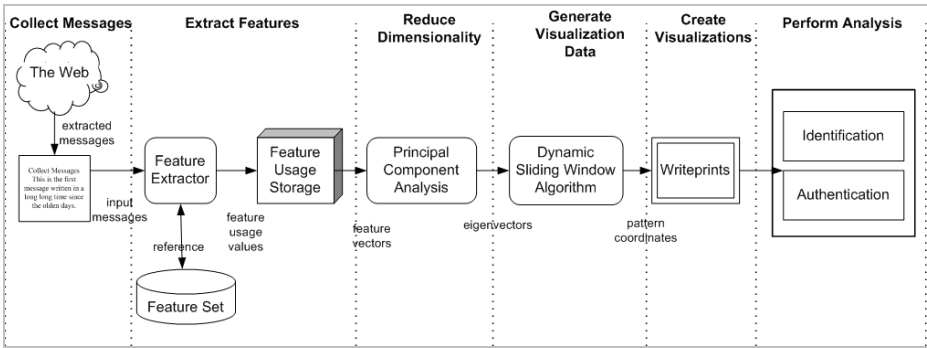


Fig. 1. Authorship Visualization Process Design

4.1 Collection and Extraction

Messages are automatically collected from the forums using web spidering programs that can retrieve all the messages for a particular author or authors. The messages are then cleaned to remove noise such as forwarded and re-quoted content. Automated feature extraction programs can then compute the feature usage values and extract the feature usage vectors. The list of features designed for online messages (based on previous online authorship studies) is presented below.

4.1.1 Feature Set: English and Arabic

We believe that the use of a more in depth feature set can provide greater insight for the assessment of content for authorship identification as compared to simple word or n-gram based features. Our English feature set consists of 270 writing style markers including lexical, syntactic, structural, and content-specific features. Figure 2 shows an overview of the complete feature set. In this particular feature set, the content specific words belong to messages related to software sales and distribution; however these words differ depending on the domain. A description of each feature can be found in Zheng et al. [20].

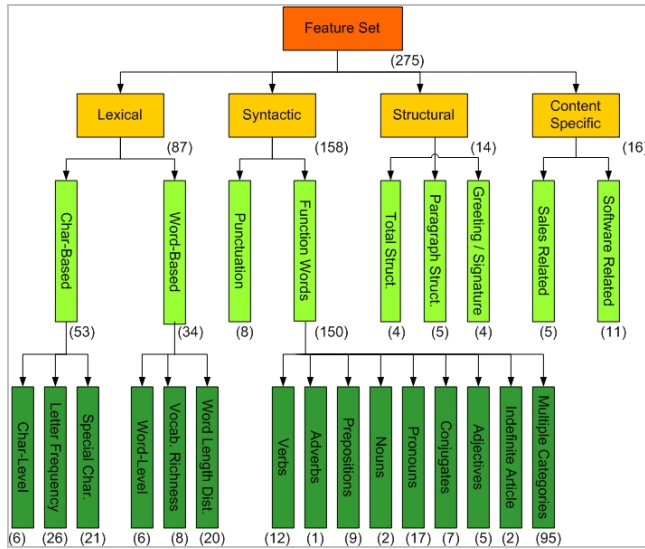


Fig. 2. Online Authorship Features

4.2 Dimensionality Reduction

Principal Component Analysis (PCA) is an unsupervised learning approach for feature selection and dimensionality reduction. It is identical to the self-features information compression variation of Karhunen-Loeve transforms that is often used in pattern recognition studies [18, 19]. The feature usage vectors for a specific group of features (e.g., punctuation) were transformed by using the two principal components (the two eigenvectors with the largest eigenvalues). Only the first two principal components were used since our analysis found that this many components were sufficient to capture the variance in the writing explained by each feature group. The extracted eigenvectors were then used by the dynamic sliding window algorithm to generate data points for the purpose of creating the *Writeprints*.

4.3 Dynamic Sliding Window Algorithm

The sliding window algorithm, originally used by Kjell et al. [9], is an iterative algorithm used to generate more data points for the purpose of creating writing style patterns. The algorithm can create powerful writing patterns by capturing usage variations at a finer level of granularity across a text document.

Figure 3 shows how the sliding window algorithm works. Once the eigenvectors have been computed using principal component analysis (Step 1) the algorithm extracts the feature usage vector for the text region inside the window, which slides over the text (Step 2). For each window instance, the sum of the product of the principal component (primary eigenvector) and the feature vector represents the x-coordinate of the pattern point while the sum of the product of the second component (secondary eigenvector) and the feature vector represents the y-coordinate of the data point (Step 3).

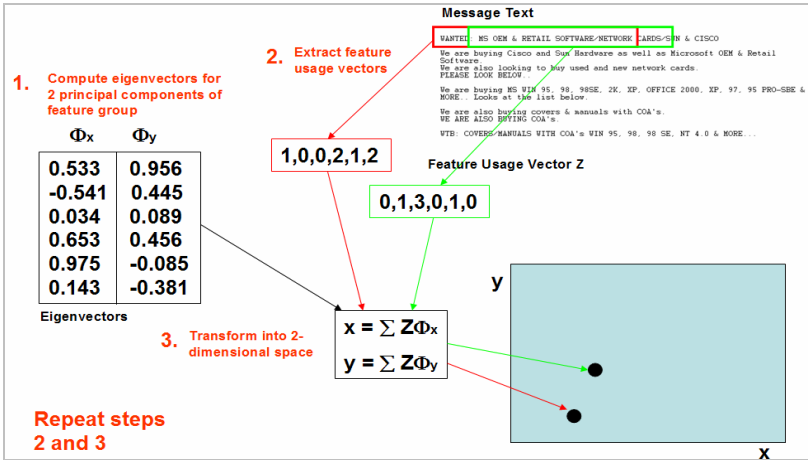


Fig. 3. Sliding Window Algorithm Illustration

Each data point generated is then plotted onto a 2-dimensional space to create the *Writeprint*. Steps 2 and 3 are repeated while the window slides over the text.

The dynamic sliding window algorithm features a couple of amendments over the original version [9], added specifically for the purpose of dealing with online material. Firstly, due to the challenging nature of cyber content and the shorter length of online messages, we applied the algorithm to our feature set sub-groups (e.g., punctuation, content-specific words, word length distribution, letter usage, etc.). Using multiple feature groups may provide better scalability to cope with the larger pool of potential authors associated with online messages. Secondly, we used dynamic window and overlap sizes based on message lengths. Online message lengths pose a challenge for authorship identification, with messages typically ranging anywhere from 20 to 5,000 characters in length (as compared to literary texts which can easily surpass lengths of 200,000 characters).

4.3.1 Selecting Feature Groups

Certain groups of features are not suitable for use with the dynamic sliding window algorithm. Ideal features are “frequency-based” features, which can easily be measured across a window, such as the usage of punctuation. Vocabulary richness, average, and percentage-based features (e.g., % characters per word, average word length) are not suitable since these are more effective across larger amounts of data and less meaningful when measured across a sliding window. Based on these criteria, we incorporated punctuation, letter frequencies, special characters, word length distributions, content specific words, and structural features. Structural features could not be captured using the sliding window, so they were transformed using feature vectors at the message level. This exception was made for structural features since they have been shown to be extremely important for identification of online messages [4]. Excluding them would result in weaker classification power for the *Writeprint* visualization.

4.3.2 Dynamic Window and Overlap Sizes

Our algorithm uses dynamic window and overlap sizes based on message lengths. This feature was incorporated in order to compensate for fluctuations in the sizes of online messages and to provide better support for shorter messages. We varied our window size between 128-1024 characters and interval between 4-16 characters.

4.4 Writeprints

Figure 4 shows an example of an author *Writeprint*. Each of the two regions shows the pattern for a particular feature group created using principal component analysis and the dynamic sliding window algorithm. Each point within a pattern represents an instance of feature usage captured by a particular window.



Fig. 4. Example Author Writeprint

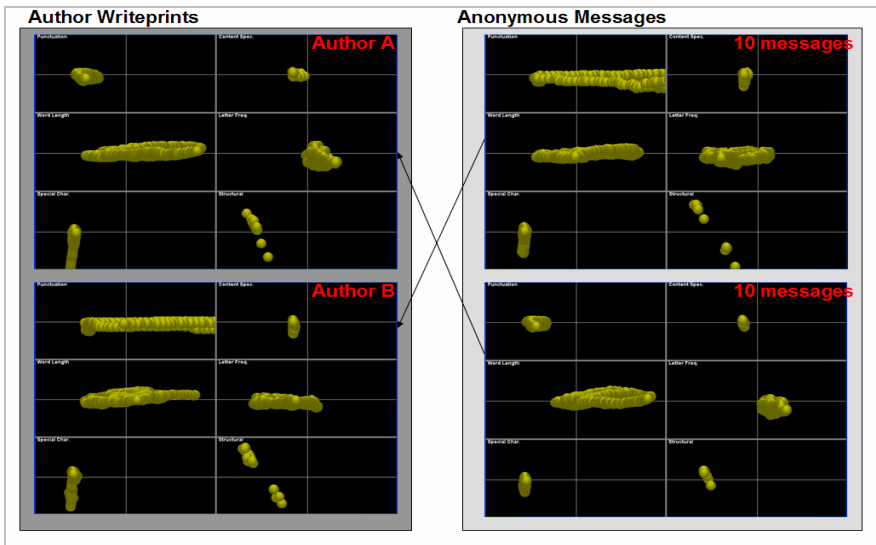


Fig. 5. Identification Example using Writeprints

4.4.1 Using Writeprints for Identification

Figure 5 presents an example of how *Writeprints* can be used for identification. The left column shows the writing style patterns for two authors (authors A-B) created

using 20 messages. After creating these *Writeprints*, we then extracted 10 additional messages for each author as anonymous messages. By comparing the anonymous patterns to the ones known to belong to authors A and B, we can clearly attribute correct authorship of the anonymous messages. The anonymous messages on top belong to Author B and the ones on the bottom belong to Author A.

4.4.2 Limitations of Writeprints

While *Writeprints* provides a powerful visualization technique for authorship identification when given larger amounts of text, it is constrained when dealing with shorter individual messages (i.e., messages less than 30-40 words). This is due to the minimum length needs of the sliding window algorithm.

5 Evaluation

We believe that the *Writeprints* is better suited for larger quantities of information. In order to evaluate the viability of this visualization, we need to create automatic recognition mechanisms and conduct experiments on messages from different forums to evaluate our hypotheses.

5.1 Automatic Recognition

An automatic recognition technique is essential for the use of *Writeprints* in an online authorship identification setting. We created a Writeprint Comparison Algorithm in order to compare author/message writing style patterns created by *Writeprints*. The Writeprint algorithm compares the anonymous message(s) against all potential authors and determines the best author-message fit based on a similarity score.

The evaluation algorithm, consists of three parts. The first part (Step 1) attempts to determine the degree of similarity based on the shape and location of patterns for each feature group. The second part (Step 2) attempts to account for differences in the sizes of the two patterns that may occur due to large variations in the number of underlying data points that exist in the two *Writeprints* being compared. The final part (Step 3) involves computing the overall score which is the sum of the average distances between points in the two patterns as calculated based on Steps 1 and 2 taken across all feature groups (e.g., punctuation, content specific words, etc.).

5.2 Test Bed

Our test bed consisted of data taken from three online forums that were used based on their relevance to cybercrime research. The forums used included a USENET forum consisting of software sales and distribution messages (misc.forsale.computers.*), a Yahoo group forum for Al-Aqsa Martyrs (an Arabic speaking extremist group), and a website forum for the White Knights (a chapter of the Ku Klux Klan). For each forum, 30 messages were collected for each of 10 authors (300 messages per forum).

5.3 Experiment

In order to evaluate the discriminatory potential of Writeprints, we conducted an experiment on our test bed which consisted of messages from the three forums. For each author, 20 messages were used for training. The remaining 10 messages per author were used to test the visualizations.

The experiment was designed to evaluate what we perceived to be the strengths of our visualization. The experiment tested the ability of Writeprints to discriminate authorship across a group of messages. Support Vector Machines (SVM) was used as a baseline since it has established itself as an effective technique for online authorship identification. The classification accuracy (# correctly classified / # total classifications) was used to evaluate performance.

Principal component analysis was performed on our training messages (20 per author) for each forum in order to compute our eigenvectors. The three sets of eigenvectors were then used with the sliding window algorithm to create our “existing” author Writeprints. The testing messages were then compared against these Writeprints in order to determine authorship.

Three different scenarios were used for the test messages. We assumed that the anonymous messages from each author were received in groups of 10 messages, 5 messages, and individual messages. Thus, given 100 test messages total, we had 10 groups of 10 messages (1 group per author), 20 groups of 5 messages (2 groups per author), or 100 individual messages (10 per author). Varying cluster sizes of messages were used in order to test the effectiveness of *Writeprints* when presented with different amounts of text. For each scenario, the classification accuracy of *Writeprints* was compared against SVM, with the results presented in Table 1.

Table 1. Writeprints and SVM Classification Accuracy

Forum	10-Message Groups		5-Message Groups		Single Messages*	
	WP	SVM	WP	SVM	WP	SVM
Software	100.00%	50.00%	95.00%	55.00%	75.47%	93.00%
White Knights	100.00%	60.00%	100.00%	65.00%	85.00%	94.00%
Al-Aqsa	100.00%	50.00%	90.00%	60.00%	68.92%	87.00%

It should be noted that for individual messages, *Writeprints* was not able to perform on messages shorter than 250 characters (approximately 35 words) due to the need to maintain a minimum sliding window size and gather sufficient data points for the evaluation algorithm.

Pair wise t-tests were conducted to show the statistical significance of the results presented. The t-test results indicated that all experiment results were significant at an alpha level of 0.01 or 0.05. Thus, *Writeprints* significantly outperformed SVM when presented with a group of 5 or 10 messages and SVM significantly outperformed *Writeprints* on individual messages.

Based on the results, it is evident that SVM is better suited for classifying individual messages while *Writeprints* performs better for a group of anonymous messages that are known to belong to a single author. Thus, *Writeprints* may be a better alternative when provided a group of messages belonging to an anonymous author, as is

quite common in computer mediated conversation. CMC conversations can result in a large group of messages written by an individual in a short period of time. Treating these messages as a single entity (as done in *Writeprints*) makes sense as compared to evaluating each as a separate short message, resulting in improved accuracy.

5.4 Results Discussion and Limitations

Based on the experimental evaluation of *Writeprints*, we believe that this technique is useful for authorship identification. Specifically, *Writeprints* is strong for identifying a group of messages. While the techniques isn't powerful enough to replace machine learning approaches such as SVM for authorship identification of individual online messages, it could provide significant utility if used in conjunction. Furthermore, *Writeprints* can provide invaluable additional information and insight into author writing patterns and authorship tendencies. While we feel that this work represents an important initial exploration into the use of information visualization for authorship identification, there is still a need for further experimentation to evaluate the scalability of these techniques across different online domains and languages, using a larger number of authors.

6 Conclusions and Future Directions

In this research we proposed a technique for authorship identification of online messages called *Writeprints*. The use of a writing style feature set specifically tailored towards multilingual online messages and automatic recognition mechanisms makes our proposed visualizations feasible for identification of online messages. The visualization provides unique benefits that can improve identification and authentication of online content. We believe that this technique represents an important and essential contribution towards the growing body of tools geared towards facilitating online accountability.

References

1. Abbasi, A. & Chen, H. Applying Authorship Analysis to Extremist-Group Web Forum Messages. *IEEE Intelligent Systems*, 20(5): (2005), 67-75.
2. Baayen, R. H., Halteren, H. v., & Tweedie, F. J. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 2: (1996), 110-120.
3. Burrows, J. F. Word patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing*, 2: (1987), 61 -67.
4. De Vel, O., Anderson, A., Corney, M., & Mohay, G. Mining E-mail content for author identification forensics. *SIGMOD Record*, 30(4): (2001), 55-64.
5. Donath, J. Identity and Deception in the Virtual Community. In *Communities in Cyberspace*, London, Routledge Press, 1999.
6. Donath, J., Karahalio, K. & Viegas, F. Visualizing Conversation. Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS, 99), Hawaii, USA, 1999.

7. Erickson, T. & Kellogg, W. A. Social Translucence: An Approach to Designing Systems that Support Social Processes. *ACM Transactions on Computer-Human Interaction*, 7(1): (2001), 59-83.
8. Kelly, S. U., Sung, C., Farnham, S. Designing for Improved Social Responsibility, User Participation and Content in On-Line Communities. *Proceedings of the Conference on Human Factors in Computing Systems (CHI '02)*, 2002.
9. Kjell, B., Woods, W.A., & Frieder, O. Discrimination of authorship using visualization. *Information Processing and Management*, 30 (1): (1994), 141-150.
10. Li, J., Zeng, R., & Chen, H. From Fingerprint to Writeprint. *Communications of the ACM*, (2006) Forthcoming.
11. Moores, T., & Dhillon, G. Software Piracy: A View from Hong Kong. *Communications of the ACM*, 43(12): (2000), 88-93.
12. Ribler, R. L., & Abrams, M. Using visualization to detect plagiarism in computer science classes. *Proceedings of the IEEE Symposium on Information Visualization*, 2000.
13. Rocco, E. Trust Breaks Down in Electronic Contexts but can be repaired by some Initial Face-to-Face Contact. *Proceedings of the Conference on Human Factors in Computing Systems (CHI '98)*, (1998), 496-502.
14. Sack, W. Conversation Map: An Interface for Very Large-Scale Conversations. *Journal of Management Information Systems*, 17(3): (2000), 73-92.
15. Shaw, C.D., Kukla, J.M., Soboroff, I., Ebert, D.S., Nicholas, C.K., Zwa, A., Miller, E.L., & Roberts, D.A. Interactive volumetric information visualization for document corpus management. *International Journal on Digital Libraries*, 2: (1999), 144-156.
16. Tweedie, F. J., Singh, S., & Holmes, D. I. Neural Network applications in stylometry: the Federalist papers. *Computers and the Humanities*, 30(1): (1996), 1-10.
17. Viegas, F.B., & Smith, M. Newsgroup Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS, 04)*, Hawaii, USA, 2004.
18. Watanabe, S. *Pattern Recognition: Human and Mechanical*. John Wiley and Sons, Inc., New York, NY, 1985.
19. Webb, A. *Statistical Pattern Recognition*. John Wiley and Sons, Inc., New York, NY, 2002.
20. Zheng, R., Qin, Y., Huang, Z., & Chen, H. A Framework for Authorship Analysis of Online Messages: Writing-style Features and Techniques. *Journal of the American Society for Information Science and Technology* 57(3): (2006), 378-393.