Evaluating text visualization for authorship analysis

Victor Benjamin[1*], Wingyan Chung[2], Ahmed Abbasi[3], Joshua Chuang[4], Catherine A Larson[1] and Hsinchun Chen[1]

Abstract

Methods and tools to conduct authorship analysis of web contents is of growing interest to researchers and practitioners in various security-focused disciplines, including cybersecurity, counter-terrorism, and other fields in which authorship of text may at times be uncertain or obfuscated. Here we demonstrate an automated approach for authorship analysis of web contents. Analysis is conducted through the use of machine learning methodologies, an expansive stylometric feature set, and a series of visualizations intended to help facilitate authorship analysis at the author, message, and feature levels. To operationalize this, we utilize a testbed containing 506,554 forum messages in English and Arabic, source from 14,901 authors that participated in an online web forum. A prototype portal system providing authorship comparisons and visualizations was then designed and constructed in order to support feasibility analysis and real world value of the automated authorship analysis approach. A preliminary user evaluation was performed to assess the efficacy of visualizations, with evaluation results demonstrating task performance accuracy and efficiency was improved through use of the portal.

Keywords: Terrorism; Text visualization; Online forum; Authorship analysis; Cybercrime

Introduction

Authorship analysis is useful in any application context where authorship attribution is uncertain, unknown, or otherwise obfuscated. Such occurrences often arise in disciplines such as history and criminology. Traditionally, authorship analysis has been performed through manual analysis. However, manual analysis has become increasingly difficult with growing usage of electronic text (e.g. e-mail, websites) and social media (e.g. forums, blogs). Problems with manual analysis arise when processing large volumes of text content or adapting traditional stylometric analysis (e.g. handwriting style) to electronic text. As a result, researchers have become interested in developing techniques to conduct automated authorship analysis on electronic text across various languages. By borrowing perspectives and techniques from computational linguistics, many traditional features used to evaluate authorship have been operationalized for use with electronic texts.

In particular, there is great interest in developing authorship visualization tools to support greater user

accountability in online communities and social media. The anonymity provided by the Internet makes it an attractive platform for those wishing to conduct various forms of crime, including drug trafficking, piracy, cybercrime, and terrorism [1–3]. Additionally, there are several trust issues with online deception between individuals and organizations that could be mitigated with better authentication services; visualization tools could help deter abuse of the anonymity the Internet provides.

However, there are several basic hurdles researchers and practitioners must overcome to perform automated authorship analysis. First, a corpus or data must be contained for analysis to be performed. If attempting to conduct analysis on social media data, data must be collected through the usage of an API or by using an automated crawler. Second, stylometric features must be extracted from the text to conduct analysis. Features must be systematically chosen, often times by borrowing from previous research and also through rigorous feature selection. Techniques borrowed from linguistics are often utilized during feature extraction. Next, there is a need for a systematic mechanism to analyze and compare the writing styles of different authors. Different computational techniques rooted in machine learning

* Correspondence: vabenji@email.arizona.edu
[1]MIS Department, University of Arizona, Tucson, AZ, USA
Full list of author information is available at the end of the article

can be utilized. Finally, results must also be transformed into informative visualizations.

Due to the growing need for automated authorship analysis and visualization tools, the Arizona Authorship Analysis Portal (AzAA) was conceived as a platform on which analyses could be developed and assessed. The portal integrates machine learning methods, a robust stylometric feature set, and a series of visualizations designed to facilitate analysis at the feature, author, and message levels. It allows identification of "extreme" jihadi authors, and also supports analysis and comparison of author writing styles. Additionally, for the purpose of providing a testbed for analyses to be run against, the system contains two datasets from an Arabic and English forum identified to contain extremist content.

In this paper, we present the AzAA portal and discuss its relevance and effectiveness in the space of automated authorship analysis. To do this, we first describe the task of authorship analysis and cover previous related works. We then move on to more deeply discuss previous works that contributed directly to the AzAA portal's development. Next we provide an overview of the AzAA system and describe the many components necessary for the system to work. We demonstrate various case studies on how the system is of use, and present results of a preliminary user evaluation of the portal's text visualization function. We conclude with an outline of our next steps and future development.

Related work

To implement a system such as the AzAA portal, a review of important, recent works in relevant disciplines is necessary. First, an understanding of authorship analysis and its purpose must be established to provide trajectory in the design and implementation of the AzAA portal. Next, perspectives and methodology from previous research in text analysis provide direction for developing and operationalizing features for authorship analysis of electronic text. Visualizations can provide richer understanding during authorship analysis, and thus we briefly discuss previous practices in text visualization domain. Lastly, to build a corpus for the AzAA portal, computational approaches based on previous works are utilized to collect, transform, and store data from Arabic and English forums.

Authorship analysis

Authorship analysis is useful when authorship attribution is uncertain, unknown, or otherwise obfuscated. The goal of any such analysis is usually one of three purposes: authorship identification, authorship characterization, and authorship similarity detection. Authorship identification compares a particular author's known writings to a particular unattributed or mis-attributed document in order to determine the level of possibility that he or she is the

author. Traditionally, authorship analysis has been applied to domains such as history, the humanities, criminology, etc. Based on stylometric analysis, or the statistical analysis of writing style, authorship analysis has grown increasingly important to those wishing to examine the virtual space. As more individuals access the Internet and participate in social media, the volume of misconduct and abuse of cyber infrastructure becomes more frequent.

Previous researchers have experimented with authorship analysis techniques in some virtual contexts such as e-mail and web forums [4–7]. For example, Li et al. [6] explored and tested key features important to identifying authorship of online texts; Argamon et al. [8] developed methodology to construct various "profiles" of an author's characteristics (such as age, gender, personality) and analyzed which features were most effective for profiling each characteristic type. Similarly, Abbasi and Chen [9] focused on stylistic features for an authorship similarity detection experiment; they conducted a study where authors' identities were not known ahead of time, but writing samples of each author could be compared for similarity/dissimilarity to one another.

However, web content often poses difficulties for authorship analysis as compared to traditional forms of writing. The most often cited challenge of authorship experiments with web content is the shorter length of online messages, which tend to average no more than a couple of hundred words [7]. Additionally, online messages tend to vary greatly in length, adding yet another challenge to balanced analysis. The ability to automatically perform authorship analysis on web contents, despite challenges identified in research, is of great asset to both researchers and practitioners. A review of recent improvements to authorship analysis on web contents reveals that improvements have been largely grounded in the development and use of writing style markers (features) of electronic text, and also in machine learning classification techniques adopted for authorship identification and similarity comparisons.

Stylometric & text analysis

The two most important analytical techniques for authorship analysis of web contents are stylometric analysis and text analysis based on machine learning approaches; both are grounded in statistics. Stylometric analysis refers to utilizing domain-specific features (i.e. characteristics) in statistical analyses to compare and distinguish one text document from another. Statistical techniques have the benefit of providing greater explanatory potential which can be useful for evaluating trends and variances over larger amounts of text; in particular, various multivariate statistical approaches have been tested and shown to provide a high level of accuracy [10,11]. Similarly, recent years have seen the usage of statistical machine learning-based

text analysis techniques grow in authorship analysis studies [4,5,7,12]. Such techniques provide scalability and performance helpful when conducting analyses on web forum messages.

Stylometric features, or writing style features, are characteristics that help in comparing and distinguishing between two documents or bodies of text. Depending on the domain that authorship analysis is applied to, features are categorized in different ways. In the context of web content, text style features, HTML features, and content-specific features are often used [9,13,14]. Text style features generally include structural features, syntactic features, and lexical features. Structural features generally include features that describe the overall organization of a document or text. When referring to web content, structural features include the usage of HTML-encoded text, which includes the ability to format text with word bolding, italics, font coloring, font size, etc. [15]. Syntactic features refer to the sentence-level of a document, including patterns used for formulating sentences. This category often includes features such as punctuation usage, function and stop word usage, etc. Syntactic features have been found to be quite useful in many different studies analyzing web content [3,4]. Lastly, lexical features are associated with word-level characteristics such as word frequency, vocabulary richness, frequency, word length distribution, etc. [13]. Lexical features are particularly important as they help establish content-specific differences among authors when conducting authorship analysis. Keywords relevant to different topics can be mapped to different authors when conducting analysis utilizing lexical features.

Computational analyses of text have grown in popularity with the increase of computational power available to researchers and practitioners. Techniques such as support vector machines, neural networks, genetic algorithms, and decision trees are all useful for text analysis tasks across [4,12,14,16]. The wider acceptance of such techniques has enabled authorship analysis to adapt to electronic text and web contents. Machine learning provides great scalability in terms of the number of features used for analysis, as well as the number of documents analyzed. These benefits over other methods greatly improves the authorship analysis task when applied to web contents, as online messages are often abundant in volume, involve classification of many authors, and provide large feature sets to utilize for analysis.

Two particular types of text analyses that are of interest to authorship analysis are sentiment analysis and affect analysis, both of which can be used to identify attitudes, emotions, moods, and polarity of a document and its author. Additionally, both analyses borrow much from natural language processing, linguistics, and machine learning techniques [1,17,18]. The analyses, often

implemented by automated machine learning classifiers, are commonly used to scrutinize a text to potentially reveal the author's opinions and affect state concerning multiple items. Such opinions and affect state can serve as additional important features that are helpful for attributing authorship of text when attribution is difficult.

Text visualization

Text visualization is the representation of large amount of text using visual metaphors [19–21]. It is concerned with getting insight into information obtained from one or more textual documents without users having read those documents. Examples of text visualization applications include generating high-quality keyphrases from text collections [22] and visualizing networks of business stakeholders on the web [23]. Despite the importance, scarce work is found in the analysis and visualization using various aspects of authorship styles and features.

In the context of web content, visualizations have traditionally been often used to create information concerning user activity of web forums, blogs, or other social media, allowing users to be more informed of their own activities and also those of fellow participants [24–27]. Most of the information projected in such visuals is entirely derived from participant activity patterns, and thus there is little evaluation of actual author-message content [13]. From the perspective of authorship analysis, activity patterns alone are not enough to accurately assign attribution to text. Thus, there exists a need for visualizations which utilize the data within message content.

Specifically, visualizations that could help researchers and practitioners assign authorship attribution in the virtual space would be of great asset. Visualizations can be used to help compare different writing samples, emphasize differences between authors, etc. The information projected by visualizations would be based entirely on the lexical, syntactic, and structural features identified in the text where attribution is in question.

There have been a few notable works on authorship visualization. Some research has used statistical techniques such as cosine similarity and principal component analysis to visualize writing style patterns [28]. Writing style patterns were rooted in word usage frequencies, and comparisons were drawn between authorship styles by observing the variance between top n-gram usage among different authors. Another study chose to use latent semantic indexing based on n-gram usage for authorship visualization [29]. Essentially, patterns in the relationships between terms and concepts contained in text are identified; this allows for individuals' authorship styles to be represented as an eigenvectors (i.e. principal components), allowing for further comparison and analysis between different authors. Further, the use of n-gram-based visualizations referred to as

Patterngrams can be used to compute document similarity [30]. Later, the visualization technique *Writeprint* was developed as a method to visualize web content [13]. *Writeprints* are useful for improving authorship identification and attribution by identifying individuals based on their writing style, including syntactic, structural, and lexical features. The visualization technique accounted for each category of features, and allowed researchers and practitioners to view authorship styles through different lenses offered by each feature category. The technique was also used to successfully attribute authorship on multilingual text. Overall, n-gram based techniques and those that account for syntactic, structural, and lexical features appear to be the most effective for authorship analysis of web contents.

Data collection & processing

To conduct authorship analysis and visualization of web content, data must first be collected and processed for use in research. Many recent studies utilizing web-based content commonly make use of automated crawlers for data collection [31,32]. Automated crawlers allow for large amounts of text to be collected very rapidly when compared to manual approaches. After web pages containing text are collected, automated parsers and feature extraction programs can be developed to strip relevant text out of web pages and compute feature usage values [3]. Feature usage values are often times stored permanently in a database and/or transformed into vectors for further analysis utilizing statistical techniques.

The AzAA portal

The AzAA portal was initially designed as an extension of the Dark Web Forum Portal (DWFP), a large archive of international terrorist and extremist web forums. The DWFP containing over 15 M messages in several different languages and supports search and analysis over a dataset of archived forum postings [13,32,33]. Currently, searching and browsing functions, multilingual translation, and social network analysis are supported, but the most recent version of the DW portal did not include the ability to perform authorship analysis, which can be important both for cybercrime investigation and counter-terrorism [34,35]. The AzAA portal was designed and implemented to fill this gap and provide additional tools for researchers and practitioners.

Research testbed and feature Set

The AzAA portal was conceptualized to help support identification of "extreme" authors of postings in forums from the Dark Web Forum Portal. The portal was designed to allow comparisons of writing samples from multiple authors, helping users identify differences and similarity in authorship style. A design framework for the portal can be viewed in Figure 1.

To construct our data set, automated crawlers were employed. We utilized a popular web crawling package called *Offline Explorer*, but any similar crawling software would work. The crawling program was used to automatically collect web pages from identified Dark Web forums, or forums that contain potentially dangerous, extremist contents. Two forums were selected for analysis; forum contents were in English and Arabic, respective to each forum. After collection, text parsing programs were written in Java to extract relevant message data embedded within collected web pages. Extracted messages could be further processed to develop lexical, syntactic, and structural features for authorship analysis. Messages are also used to identify extreme authors through their language usage.

By referring to past research, we were able to identify a total of 4,000 lexical, syntactic, and structural features to extract for authorship analysis. Lexical features included words and terminology that may indicate potentially extremist contents. User messages can be broken into word vectors, which each unique word mapping to a unique feature that may help with authorship identification [9]. Additionally, as many features can be derived from author messages, lexical features compose the majority of the 4,000 features used in our research. Structural features of web content generally consist of usage of HTML; relevant features include image usage, hyperlink usage, font colors, font type, font size, text alignment, text bolding, italics, etc. Extracted syntactic features included punctuation usage, sentence patterns, etc.

In the interest of performing analysis on "extreme" authors, it is useful to measure author sentiments and affect states as a text analytics-based approximation [36,37]. Thus, we perform sentiment and affect analysis using a J48 classifier [37,38]. We extract content-specific features based on feature frequency and classifier information gain. Such features include religious/cultural terms, sentiment cues, and words associated with violence, anger, hate, and racism [36–38]. These techniques provide some means for identifying authors with the highest sentiment and affect intensities for anger, violence, hate, etc. to be identified and selected through filters. In prior benchmarking, the method has yielded affect intensity mean percentage errors of 5% or less on similar Dark Web forums [36]. Thirty of the authors with the highest average intensities for these affect classes (as well as a minimum of 100 postings) were identified by the classifier. This approach undertaken is consistent with prior work on the use of affect/sentiment analysis to identify highly relevant forum members in the Dark Web [36–39].

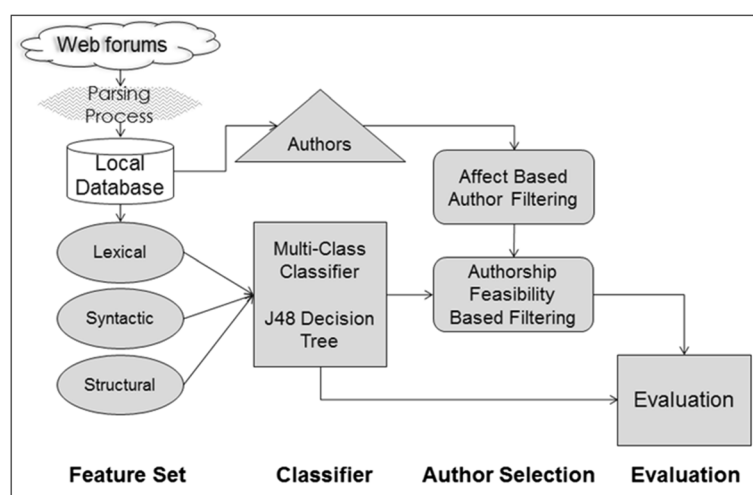All authors do not necessarily have a unique pattern, however; some exhibit writing patterns that are erratic

Figure 1 Authorship analysis design framework. Web forum pages are collected, with relevant data extracted and archived in a database. Features are generated from author messages, which are used for both author feasibility testing. Extreme authors are chosen based on affect analysis; identified authors are then tested for their feasibility in this experiment, as not all authors have unique writing styles. Authorship analysis is then performed using a decision tree classifier. Test results are then evaluated.

and/or include considerable reposting, quoting, plagiarism, non-sequiturs, short responses, etc. To evaluate the author selection, feasibility analysis was conducted to determine which authors were detectable. The latest 100 posts for each of the 30 identified authors were used for evaluation purposes. Recent postings were selected to avoid writing style changes that may occur naturally over time. For each author, 50 messages were used for classifier training and the other 50 are for testing. The authorship analysis methods employed were similar to ones utilized in prior studies using supervised machine learning classifiers such as a multi-class decision tree and established stylometric identification feature sets encompassing lexical, syntactic, structural, and content-specific attributes [4–7]. Only authors for which at least 90% of test messages were correctly classified were retained in the AzAA portal. The rational being that if the underlying patterns/insights and descriptive analytics are only meaningful if the associated classification performance is high. Consistent with prior work, using all three feature sets together (text style, HTML, and content-specific features) for authorship identification yielded the best overall performance, with over 90% macro-level accuracy [5,7].

System architecture

The AzAA portal is operationalized with many modern technologies and computing standards. At the core of the AzAA portal is an Apache webserver with Tomcat for JavaServer Pages support. Both of these Java applications are open source software created and maintained by the Apache Foundation. We employ the traditional Model-View-Controller (MVC) perspective for implementing the AzAA portal. In particular, we make use of the Struts2

Framework and Spring framework, two popular enterprise-level open-source frameworks, were adopted for scalability, flexibility, compatibility, and extendibility. With these two frameworks, we can easily apply the MVC perspective, thus allowing the portal to be more easily integrated into other local projects sharing the same frameworks. The front-end design (i.e. "view") and implementation were through JSP and HTML5 elements such as Javascript (JQuery and Bootstrap), HTML and CSS. Feature extraction and analysis (i.e. "model") for each author is calculated and stored in a Microsoft SQL Server database for quick recall at run-time. A NoSQL server for storage is also a viable alternative. The interface allows the user to quickly select various visualizations to view data through, supporting the controller functionality of the MVC perspective.

Use of the AzAA portal

Use of the AzAA portal allows for easy and quick analysis of authorship styles. Specifically, the AzAA portal provides users with multiple perspectives and visualizations for viewing different data, and comparisons between authorship styles can be performed with minimal input required on part of the user. An integrated tool to help support the identification of extreme authors 0Here we showcase and detail portal functionality.

When users log into the system, they are greeted with a welcome screen from which a user may select how to proceed. The welcome screen contains text to introduce the user to the system and to explain different ways to use the system. At this point, the user may choose to view authorship styles at an author-level or message-level perspective (Figure 2). Both perspectives ultimately display the similar data to users, but offer more focus on
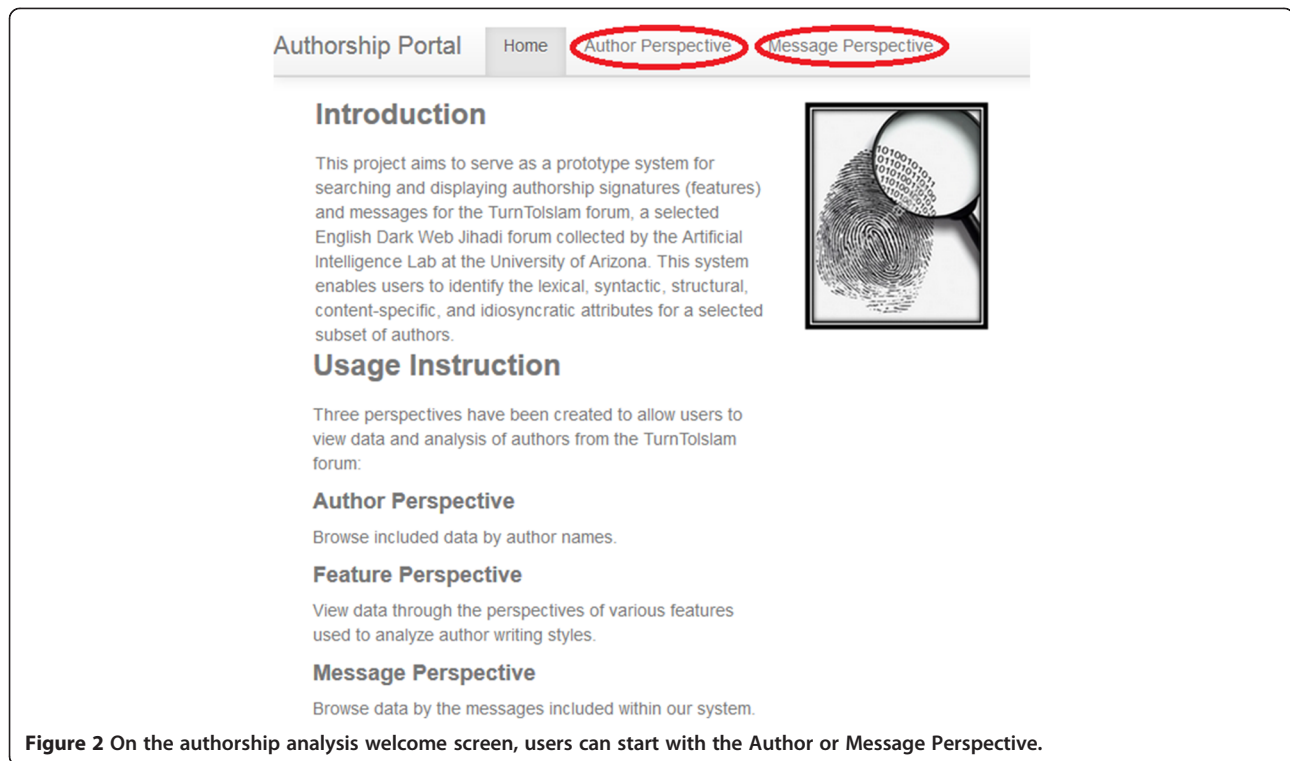
Figure 2 On the authorship analysis welcome screen, users can start with the Author or Message Perspective.

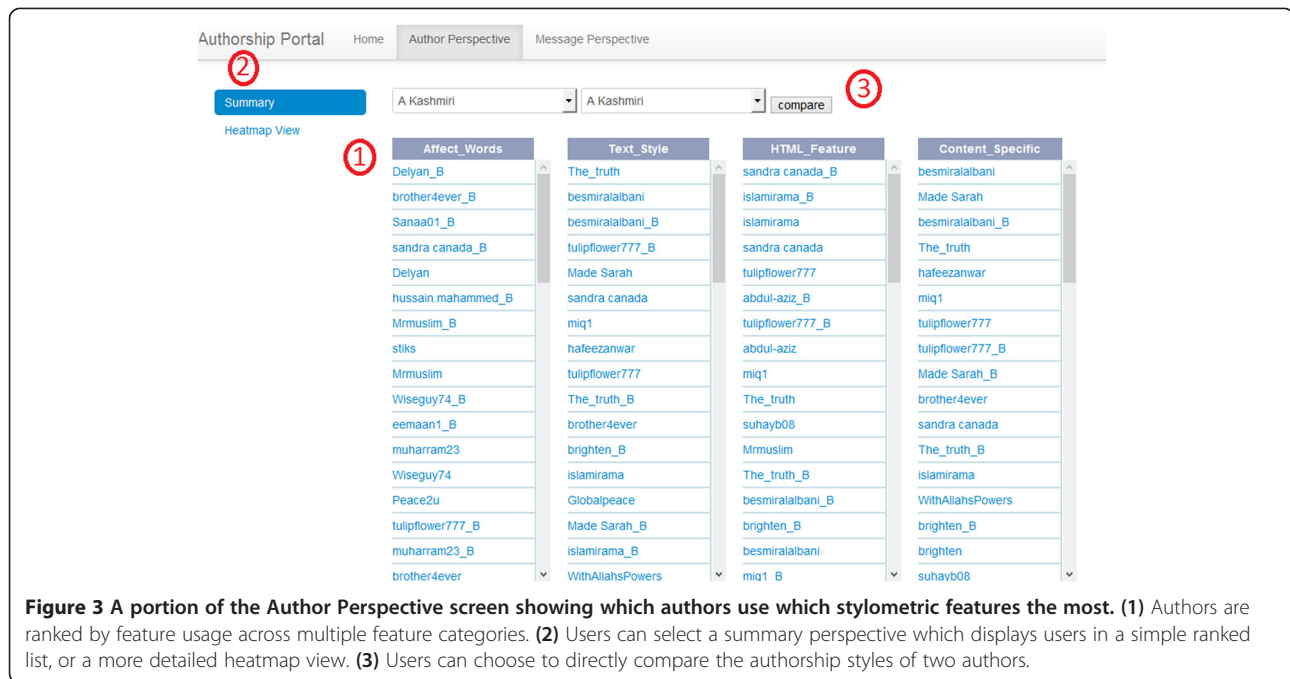observing author postings at a cumulative level and at an individual-message level.

In the Author-Based Perspective, users can view the authorship styles of individual forum participants captured within our dataset. The author-based perspective is especially useful for identifying which authors use specific stylometric features the most (Figure 3). On this screen, users are presented with some author information as well as options to proceed. Users are initially presented with columns containing ranked lists of author feature usage for feature categories such as affect words, HTML features, content features, etc. Users can select to keep the columns in a simple summary view, or choose a more detailed heatmap view (Figure 4) for deeper scrutiny. Users also are supplied with dropdown menus to select two authors for which to compare authorship style.

Visualizations provide quick context on author stylometric features. Heatmaps (Figure 4) can show differences in data by coloring data points in various shades to show frequent feature usage. The darker and more intense shades used to color datapoints are representative of greater feature usage or frequency. For example, Figure 4 shows a portion of the heatmap where authors are sorted by their usage of racist terminology. A gradient is formed that visually presents feature usage per author, relative to other authors. The user Muharram23 has the most intensely shaded cell for the racism feature, and thus we can conclude from this visualization that

this author uses racist terminology the most frequently out of all authors in our dataset.

From the author-perspective, users can also select to directly compare the authorship styles of two authors. Users are presented with a table comparing authors through individual feature-level comparisons (Figure 5), while also generating a radar chart to summarize authorship differences and similarities for the user (Figure 6). Comparison of authors on the individual-feature level is particularly useful if the user of the system is interested in evaluating how some particular authors differ on a specific subset of features. If the user of the system has very focused questions, this perspective to view authorship differences may be useful.
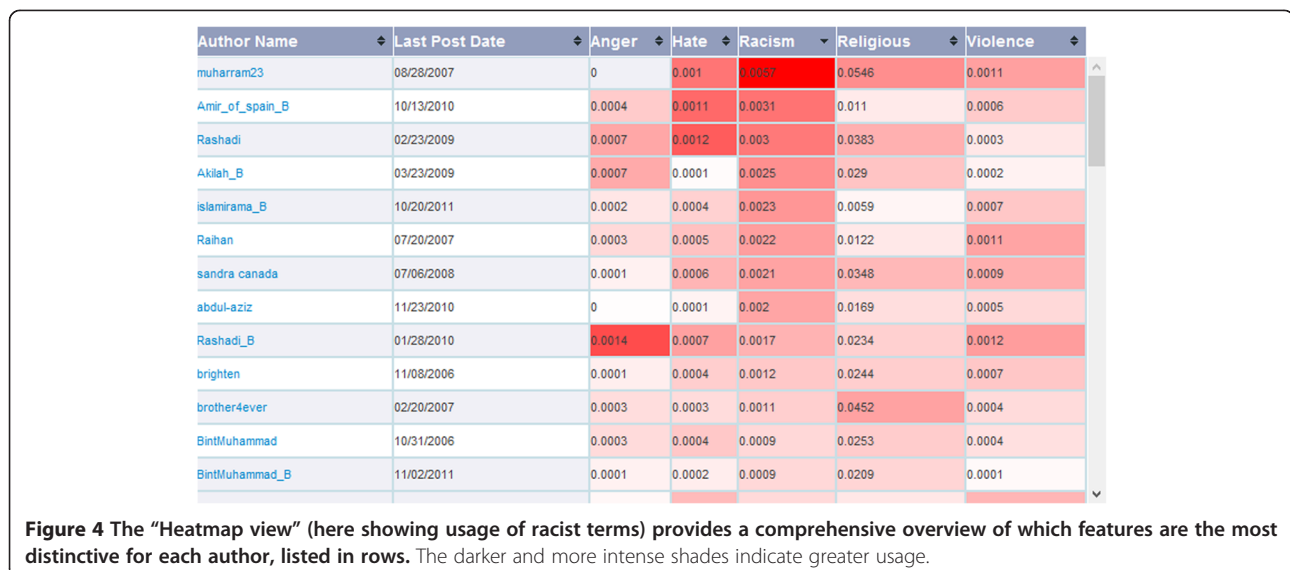
Conversely, the radar chart shown in Figure 6 is more suited for providing fast, high-level summaries of authorship style comparisons. In our examples, we compare the author Muharram23 against the author Brother4ever. Differences for various stylometric features can be seen in Figure 5, while Figure 6 highlights differences across the major feature categories. The radar chart supports such comparisons at both the feature category and subcategory level. Here we show a comparison at the subcategory level betweenMuhahrram23 and Brother4-ever. Muharram23 uses a great deal of racist terminology within his messages; conversely, Brother4ever appears to discuss a wider range of topics, particularly religion and culture.

Figure 3 A portion of the Author Perspective screen showing which authors use which stylometric features the most. (1) Authors are ranked by feature usage across multiple feature categories. **(2)** Users can select a summary perspective which displays users in a simple ranked list, or a more detailed heatmap view. **(3)** Users can choose to directly compare the authorship styles of two authors.

Users may also browse the raw message contents in their original form or plaintext form, supported by feature highlighting within messages. To do this, users must switch from the author-perspective to the message-perspective, which can be performed as seen in Figure 2. When users select the message-perspective, they land on a screen in which they can select authors from a drop-down menu to view individual messages of. We select the user 'Ahmed ibn Ibrahim' as an example to walk through this section of the system; after choosing the 'Ahmed ibn Ibrahm' via the author drop-down menu, the page is populated with a list the author's messages (Figure 7). From here, one can select to

view a message in plain text or with its original HTML formatting. Additionally, users may also employ a series highlighting tools to easily identify specific features within messages.

Feature highlighting within text can help draw user attention towards more interesting aspects of authorship style. It is particularly useful for aiding in quick identification of lexical features. In Figure 8, one of the author's messages has been selected for viewing, and many feature highlighting options have been toggled on. Specifically, some affect-related features that imply hate, violence, anger, etc. on part of the author are highlighted in the

Figure 4 The "Heatmap view" (here showing usage of racist terms) provides a comprehensive overview of which features are the most distinctive for each author, listed in rows. The darker and more intense shades indicate greater usage.

Feature	Category	Sub-Category	Dominant Author	Mean Usage for muharram23	Mean Usage for brother4ever	Max Mean Usage
force	Affect Words	Violence	muharram23	0	0	0.0004
destroy	Affect Words	Violence	muharram23	0	0	0.0006
battle	Affect Words	Violence	muharram23	0	0	0.0003
defeat	Affect Words	Violence	brother4ever	0	0	0.0001
hit	Affect Words	Violence	muharram23	0.0002	0	0.0002
fight	Affect Words	Violence	brother4ever	0	0.0002	0.0007
torture	Affect Words	Violence	muharram23	0.0002	0.0001	0.0005
war	Affect Words	Violence	muharram23	0.0004	0	0.0016
%	Text Style	Special Characters	muharram23	0.0001	0.0001	0.0004
_	Text Style	Special Characters	muharram23	0.0002	0.0001	0.0015
]	Text Style	Special Characters	brother4ever	0.0001	0.0003	0.0008
[	Text Style	Special Characters	brother4ever	0.0001	0.0003	0.0008
alaykum	Content Specific	Social and Communities	muharram23	0.86	0.04	0.98
BISMILLAH	Affect Words	Religious	brother4ever	0	0.0001	0.001
alaikum	Affect Words	Religious	brother4ever	0	0.0002	0.0156

Figure 5 Feature-level comparisons between two authors, Muharram23 and Brother4ever. This type of comparison is useful to identify differences in individual feature expression between authors.

message with "warm" tones (i.e. red, orange, pink). Features indicating content about topical associations including religion, education, politics, daily life, etc. are highlighted in "cooler" tones (i.e. blues, greens, purples). The interface also provides a simple within-message search, assisting users in locating specific keywords in lengthy messages.

As with any system development, it is helpful to conduct user studies in order to evaluate the effectiveness, usefulness, and usability of a system. To measure the value and effectiveness of the Authorship Portal, we conducted a user study with 31 participants. The overall goal of the experiment was to evaluate the performance of the portal's visualization functionalities, including feature highlighting on the message-level, the stylometric feature radar chart for author-level comparisons, and the stylometric heatmap found within the author-perspective.

A. Experimental Setup

As described previously, the authorship analysis task is useful in any context where authorship attribution is uncertain. Traditional authorship analysis has relied on manual analysis of text and writing style, but manual analysis does not effectively scale to large Internet-based datasets. In the context of identifying extreme authors within virtual communities such as web forums, manual
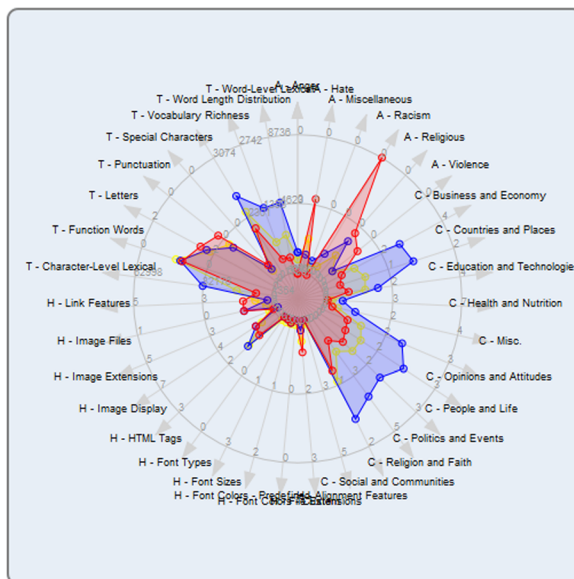

Figure 6 The radar chart visualization of the authorship styles of Muharram23 and Brother4ever. This visualization is particularly useful for quick summaries that highlight the main differences and similarities in authorship style between two authors.

Figure 7 System users can view individual messages of specific authors via the message-perspective feature of the portal. By selecting an author through the drop-down menu, users can view a list of messages written by the author. Users can then select individual messages to read in plaintext or with original HTML formatting, supported by feature-highlighting tools to draw attention to various aspects of authorship style.

analysis techniques are difficult and time-consuming to use. In these cases, automated analyses and integrated tools such as the AAzA portal can help practitioners conduct authorship analyses much more quickly at a greater scale. Further, systems such as the AAzA portal can help support practitioners by allowing for various visualization techniques and pre-programmed analyses that can be quickly executed. However, during and after an integrated tool to support authorship analysis, it is useful to evaluate the effectiveness of the tool in helping complete authorship-related tasks. In our case, the AAzA portal should be evaluated for its effectiveness in helping users identify the writing styles of different extreme authors.

The study participants were undergraduate students at different stages of their academic curriculum. The participants were tasked with using the portal to answer a series of simple questions pertaining to authorship style of specific authors within our dataset. The purpose of such tasks is intended to help evaluate performance of the portal in identifying and comparing "extreme" authors that are present on TurnToIslam online forum.

We adopted a one-factor repeated-measures approach in our experimental design, which has been shown to demonstrate a greater precision than designs that employ only between-subjects factors [40]. Each subject used the Authorship Portal to answer two sections of questions and then provide ratings on a number of

Figure 8 Message-perspective view with original HTML formatting and feature highlighting.

statements and their demographic data in a third section. In one of the first two sections, the participant used the portal's visualization functionality to answer the questions while in the other section the participant did not use the functionality. The subject used the portal to answer two or three questions in the section's three parts. When allowed to use the portal's visualization feature, the participant would be able to use the portal's feature highlighting, authorship comparison spider chart, and stylometric heatmap in the section's three parts respectively. A sample question in Part 1 is "How many times do "Opinions and Attitudes" words appear in the message?" A sample question in Part 2 is "Between BintMuhammad and Raihan, which author has a higher usage of affect words?" A sample question in Part 3 is "Which author has the highest usage of "Politics and Events" content in their authoring of forum messages?" Two different sets of questions were used in the two sections.

The whole experiment took about 60 minutes. In the first 10 minutes, a participant was given a tutorial in which the experimenter guided the use of the portal's functionality. Then the participant worked on the two sections of questions as described above (approximately 20 minutes per section). The order of using or not using the portal's visualization features (i.e., treatment vs. control) was assigned randomly to each participant to remove any bias on the results due to learning effect. The assignment of question sets (i.e., Set 1 or Set 2) was also random. Each question set contained identical items, but the order of items was changed. These two random assignments created four scenarios (T1-C2, T2-C1, C2-T1, and C1-T2 where "T" stands for "treatment," "C" stands for "control," and the two numbers stand for the respective sets of questions). A participant is thus randomly assigned to one of these four scenarios. Upon finishing the two sections of questions, the subject filled out a short questionnaire asking them to rate (on a five-point Likert Scale) their perception on the u6sability of the portal's visualization functionality and to provide demographic data.

The thirty-one participants were undergraduate students (20 males and 11 females) enrolled in a business software application course or a business statistics course offered by a regional university in the United States. The students were primarily college-age (between 18 and 25 years old), with an average age of 22.98 years old.

B. Performance Measures

1) *Accuracy:* The accuracy of the task performance was measured by how close the subject's answer was to the correct answer (for tasks in in Part 1 where a number was expected), as shown in the following formula. The accuracies in all tasks in Part 1 were averaged to obtained an overall accuracy of that part.

$$\text{Accuracy} = 1 - \min\left(\left|\frac{\text{Correct Answer} - \text{Subject's Answer}}{\text{Correct Answer}}\right|, 1\right)$$

For tasks in Parts 2 and 3 where written responses are expected, the accuracy was calculated by averaging the correctness of each task's performance (correct response = 1, incorrect response = 0).

2) *Efficiency:* The efficiency was measured by the time elapsed (in minutes) between the beginning of a task and the completion of the task.
3) *User Rating:* The user rating was measured in a five-point Likert Scale, where 5 means "strongly agree" and 1 means "strongly disagree."

C. Hypothesis Testing

We are interested in testing these hypotheses about the accuracy and efficiency of the Authorship Portal.

H1. The Authorship Portal's feature highlighting function enables users to achieve a significantly higher accuracy in authorship analysis (counting and relating category-specific words) than not using the function.

H2. The Authorship Portal's feature highlighting function enables users to achieve a significantly higher efficiency in authorship analysis (counting and relating category-specific words) than not using the function.

H3. The Authorship Portal's authorship comparison spider chart function enables users to achieve a significantly higher accuracy in authorship analysis (comparing authors' use of category-specific features, sub-category features, and similarity of authors' writing profiles) than not using the function.

H4. The Authorship Portal's authorship comparison spider chart function enables users to achieve a significantly higher efficiency in authorship analysis (comparing authors' use of category-specific features, sub-category features, and similarity of authors' writing profiles) than not using the function.

H5. The Authorship Portal's stylometric heatmap function enables users to achieve a significantly higher accuracy in authorship analysis (identifying authors' usage of category-specific features and sub-category features) than not using the function.

H6. The Authorship Portal's stylometric heatmap function enables users to achieve a significantly higher efficiency in authorship analysis (identifying authors' usage of category-specific features and sub-category features) than not using the function.

D. Experimental Results

The accuracy and efficiency of task performance using the Authorship Portal's visualization functions are generally higher than those without using the visualization functions. Table 1 shows the detailed performance levels and the mean differences in all three parts of each section (using or not using visualization). The figures show that the participants achieved higher efficiency in all three parts of the study, and obtained higher accuracy in all parts (except 1b) when they used the visualization functions of the Authorship Portal.

Statistical tests of the differences between using and not using the visualization functions of the portal indicate significance at most parts of the study (alpha error = 0.05). The columns labeled "p-value$_A$" in Table 1 show that the participants achieved significantly higher accuracy in answering the questions of Parts 2a, 2b, 2c, 3a, and 3b when they used the Authorship Portal's visualization functions. Therefore, hypotheses H3 and H5 were confirmed. We believe that the portal's spider chart and stylometric heatmap provided accurate comparison of the authorship features, thus contributing to the significant results. The columns labeled "p-value$_E$" in Table 1 show that the participants used significantly less time in answering the questions in all parts when they used the Authorship Portal's visualization functions. Therefore, hypotheses H2, H4, and H6 were confirmed. We believe that the portal's spider chart and stylometric heatmap functions helped participants to quickly identify the information they needed to answer the questions, thus contributing to the superior efficiency. On the other hand, hypothesis H1 was not confirmed, even though participants obtained a higher accuracy in Part 1a on average when using the portal's feature highlighting. We

believe that it was because participants were able to count the category-specific words accurately in simple messages even without using feature highlighting of the portal. However, more complicated tasks such as comparing author style and identifying feature usage among all authors were shown to be more difficult that they must rely on advanced functions such as the Authorship Portal's visualization in order to achieve significantly higher accuracy.

Subjects rated the Authorship Portal very highly. Table 2 shows their ratings on three statements related to the three visualization functions of the portal. All these ratings are close to the maximum of 5 (strongly agree) along a Likert Scale. In particular, the mean rating of the Authorship Comparison Spider Chart is the highest (4.81) among the three, showing subjects' preference toward a novel visualization of the different writing feature values.

E. Discussion and implication

The highly positive results shown in the experimental findings illustrate the power of the Authorship Portal's visualization functions. Using these functions, subjects were able to achieve higher accuracies (in Parts 2 and 3) and efficiencies (in all parts) than using only ordinary message browsing. These results demonstrate a high usability of the portal in supporting authorship analysis. The portal can possibly save analysts' time and enhance accuracy in understanding online messages related to terrorist activities. Considering the increasing use of forum data and social-media-based analysis in security informatics (e.g., [41]), this study provides new empirical findings to confirm the usability and efficiency of using visualization in authorship analysis. The results should be relevant to

Table 1 Accuracy and efficiency of task performance

Pt.	Task (related function of the portal)	A(V)	A(~V)	Diff$_A$	p-value$_A$	E(V)	E(~V)	Diff$_E$	p-value$_E$
1a	Counting category-specific words (feature highlighting)	0.83	0.77	0.06	0.4367	1.45	4.45	−3.00	0.0020 **
1b	Relating category-specific words (feature highlighting)	0.46	0.47	−0.01	0.9286	1.90	2.74	−0.84	0.0084 **
2a	Comparing authors' use of category-specific features (spider chart)	0.90	0.48	0.42	0.0007 **	1.81	4.61	−2.80	0.0000 **
2b	Comparing authors' use of sub-category-specific features (spider chart)	0.94	0.68	0.26	0.0090 **	1.55	2.81	−1.26	0.0001 **
2c	Comparing similarity of authors' writing profiles (spider chart)	1.00	0.81	0.19	0.0118 **	2.00	3.16	−1.16	0.0061 **
3a	Identifying authors' usage of category-specific features (stylometric heatmap)	0.94	0.32	0.62	0.0000 **	1.52	4.03	−2.51	0.0000 **
3b	Identifying authors' usage of sub-category-specific features (stylometric heatmap)	0.90	0.19	0.71	0.0000 **	1.32	4.42	−3.1	0.0075 **

Note:
A(V) = Mean accuracy of task performance using visualization.
A(~V) = Mean accuracy of task performance without using visualization.
Diff$_A$ = Mean difference of A(V) − A(~V).
E(V) = Mean efficiency of task performance using visualization (in minutes).
E(~V) = Mean efficiency of task performance without using visualization (in minutes).
Diff$_E$ = Mean difference of E(V) − E(~V) (in minutes).
**p < 0.01.

Table 2 Subjects' rating of authorship portal

Statement	Mean rating	S.D.
I find the feature highlighting of the Authorship Portal to be more useful in identifying message features than manually reading the forum messages.	4.68	0.54
I find the authorship comparison spider chart of the Authorship Portal to be more useful in comparing authors' writing than manually reading and comparing the authors' messages.	4.81	0.48
I find the stylometric heatmap of the Authorship Portal to be more useful in feature usage than manually reading the forum messages.	4.74	0.51

terrorism and informatics researchers, visualization users, and security practitioners.

Conclusions and future work

The Arizona Authorship Analysis (AzAA) Portal was developed primarily to support efforts in authorship analysis for terrorism research, cybercrime investigation, and intelligence analysis. Presently, the portal supports identification of "extreme" authors, as well as comparison of authorship styles between authors to reveal writing style similarities and differences. Additionally, sentiment analysis incorporating various lexical features was operationalized via a machine learning classifier for deeper analysis. Results of such analysis were useful for formulating the web-based visualizations that serve as graphical summaries of authorship styles. Other useful functions such as message searching, message browsing, and feature highlighting were also implemented within the system interface, allowing users to explore authors' styles from a variety of perspectives and contexts.

A user evaluation was designed and executed to measure the effectiveness and performance of the AzAA portal in assisting with authorship analysis. Specifically, user evaluation participants were asked to complete a series of tasks involving use of portal text visualizations. Evaluation results demonstrate that the visualizations support greater task efficiency and accuracy for task performance. The AzAA portal demonstrates potential to possibly save analysts' time while also enhancing understanding of online messages related to potential terrorist activities.

Future efforts on the AzAA portal will include the operationalization of a larger testbed for investigation with the system. As large-scale, big data analysis has become an important topic in similar research, it is important to consider how the AzAA portal can be extended to handle larger authorship analysis tasks. Technical components of the authorship analysis algorithm itself may be also be improved; for example, the current decision tree classifier may be replaced with a more scalable SVM classifier in the future [2]. More types of visualizations can also be developed to add value and aid in the process of authorship analysis. Finally, further evaluation of the system's efficacy for authorship analysis tasks is always valuable for finding new directions in which to improve the AzAA portal.

Author details
[1]MIS Department, University of Arizona, Tucson, AZ, USA. [2]Department of Decision and Info. Sciences, Stetson University, DeLand, FL, USA. [3]McIntire School of Commerce, University of Virginia, Charlottesville, VA, USA. [4]HC Analytics, Tucson, AZ, USA.

References
1. KKR Choo, Organized crime groups in cyberspace: a topology. Trends Organ Crime **11**(3), 270–295 (2008)
2. D Zimbra, H Chen, H, Scalable Sentiment Classification Across Multiple Dark web Forums, in *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI 2012)*, 2012, pp. 78–83
3. V Benjamin, H Chen, Securing Cyberspace : Identifying key Actors in Hacker Communities, in *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI 2012)*, 2012, pp. 24–29
4. O De Vel, A Anderson, M Corney, G Mohay, Mining e-mail content for author identification forensics. SIGMOD Record **30**(4), 55–64 (2001)
5. A Abbasi, H Chen, Applying authorship analysis to extremist-group web forum messages. IEEE Intell Syst **20**(5), 67–75 (2005)
6. J Li, R Zheng, H Chen, From fingerprint to writeprint. Commun ACM **49**(4), 76–82 (2006)
7. R Zheng, Y Qin, Z Huang, H Chen, A framework for authorship analysis of online messages: writing-style features and techniques. J Am Soc Inf Sci Technol **57**(3), 378–393 (2006)
8. S Argamon, M Koppel, J Pennebaker, J Schler, Automatically profiling the author of an anonymous text. Commun ACM **52**(2), 119–123 (2009)
9. A Abbasi, H Chen, Writeprints: a stylometric approach to identify-level identification and similarity detection in cyberspace. ACM Trans Inf Syst **26**, 2 (2008)
10. RH Baayen, H Halteren, FJ Tweedie, Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. Literary Ling Comput **2**, 110–120 (1996)
11. JF Burrows, Word patterns and story shapes: the statistical analysis of narrative style. Literary Ling Comput **2**, 61–67 (1987)
12. FJ Tweedie, S Singh, DI Holmes, Neural network applications in stylometry: the federalist papers. Comput Hum **30**(1), 1–10 (1996)
13. A Abbasi, H Chen, Visualizing Authorship for Identification, in *Proceedings of the 4th IEEE Symposium on Intelligence and Security Informatics*, 2006, pp. 60–71
14. V Benjamin, H Chen, Machine Learning for Attack Vector Identification in Malicious Source Code, in *IEEE Intelligence and Security Informatics*, 2013, pp. 21–23
15. T Urvoy, E Chauveau, P Filoche, T Lavergne, Tracking web spam with HTML style similarities. ACM Trans Web **2**(1), 3 (2008)
16. X Liu, H Chen, AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *Smart Health* (Springer, Berlin Heidelberg, 2013), pp. 134–150

17. A Abbasi, H Chen, CyberGate: a design framework and system for text analysis of computer mediated communication. MIS Q **32**(4), 811–837 (2008)
18. A Balahur, JM Hermida, A Montoyo, Detecting implicit expressions of emotion in text: a comparative analysis. Decis Support Syst **53**(4), 742–753 (2010)
19. W Chung, H Chen, JF Nunamaker, A visual framework for knowledge discovery on the web: an empirical study on business intelligence exploration. J Manag Inf Syst **21**(4), 57–84 (2005)
20. JA Wise, JJ Thoma, K Pennock, D Lantrip, M Pottier, A Schur, V Crow, Visualizing the non-Visual: Spatial Analysis and Interaction With Information from Text Documents, in *IEEE, Proceedings of Information Visualization*, 1995, pp. 51–58
21. V Benjamin, W Chung, A Abbasi, J Chuang, C Larson, H Chen, Evaluating Text Visualization: An Experiment in Authorship Analysis, in *IEEE International Conference on Intelligence and Security Informatics*, 2013, pp. 16–20
22. J Chuang, CD Manning, J Heer, Without the clutter of unimportant words: descriptive keyphrases for text visualization. ACM Trans Comput Hum Interact **19**(3), 1–29 (2012)
23. W Chung, Visualizing e-business stakeholders on the web: a methodology and experimental results. Int J Electron Bus **6**(1), 25–46 (2008)
24. T Erickson, WA Kellogg, Social translucence: an approach to designing systems that support social processes. ACM Trans Comput Hum Interact **7**(1), 59–83 (2001)
25. W Sack, Conversation map: an interface for very large-scale conversations. J Manag Inf Syst **17**(3), 73–92 (2000)
26. J Donath, K Karahalio, F Viegas, Visualizing Conversation, in *Proceedings of the 32nd Hawaii International Conference on System Sciences*, 1999
27. FB Viegas, M Smith, Newsgroup Crowds and Authorlines: Visualizing the Activity of Individuals in Conversational Cyberspaces, in *Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004
28. B Kjell, WA Woods, O Frieder, Discrimination of authorship using visualization. Inf Process Manag **30**(1), 141–150 (1994)
29. CD Shaw, JM Kukla, I Soboroff, DS Ebert, CK Nicholas, A Zwa, EL Miller, DA Roberts, Interactive volumetric information visualization for document corpus management. Int J Digit Libr **2**, 144–156 (1999)
30. RL Ribler, M Abrams, Using Visualization to Detect Plagiarism in Computer Science Classes, in *Proceedings of the IEEE Symposium on Information Vizualization*, 2000, pp. 173–178
31. TJ Fu, A Abbasi, H Chen, A focused crawler for dark web forums. J Am Soc Inf Sci Technol **61**(6), 1213–1231 (2010)
32. Y Zhang, S Zeng, CN Huang, L Fan, X Yu, Y Dang, CA Larson, D Denning, N Roberts, H Chen, Developing a Dark web Collection and Infrastructure for Computational and Social Sciences, in *Intelligence and Security Informatics*, 2010, pp. 59–64
33. H Chen, *Dark web: Exploring and Data Mining the Dark Side of the web* (Integrated Series in, Information Systems, 2011)
34. R Zheng, J Li, H Chen, H, Z Huang, A framework for authorship identification of online messages: writing-style features and classification techniques. J Am Soc Inf Sci Technol **57**(3), 378–393 (2006)
35. G Frantzeskou, S Gritzalis, SG MacDonnel, Source Code Authorship Analysis for Supporting the Cybercrime Investigation Process, in *Proceedings of the 1st International Conference on E-Business and Telecommunication Networks*, 2004, pp. 85–92
36. A Abbasi, H Chen, Analysis of affect intensities in extremist group forums. In *Terrorism informatics* (Springer, US, 2008), pp. 285–307
37. A Abbasi, H Chen, Affect Intensity Analysis of Dark Web Forums, in *Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics*, 2007, pp. 282–288
38. A Abbasi, H Chen, Analysis of Affect Intensities in Extremist Group Forums, in *Terrorism Informatics*, 2007, pp. 285–307
39. D Zimbra, H Chen, A cyber-archaeology approach to social movement research: framework and case study. J Comput-Mediat Commun **16**, 48–70 (2010)
40. J Myers, A Well, *Research design and statistical analysis* (Lawrence Erlbaum Associates, 1995)
41. W Chung, D Zeng, IMood: Discovering U.S. Immigration Reform Sentiment, in *Proceedings of the 23rd Workshop on Information Technology and Systems*, 2013