MIS Quarterly

SPECIAL ISSUE

CYBERGATE: A DESIGN FRAMEWORK AND SYSTEM FOR TEXT ANALYSIS OF COMPUTER-MEDIATED COMMUNICATION[1]

By: **Ahmed Abbasi**
Sheldon B. Lubar School of Business
University of Wisconsin–Milwaukee
Milwaukee, WI 53217
U.S.A.
abbasi@uwm.edu

Hsinchun Chen
Management Information Systems
Eller College of Management
University of Arizona
Tucson, AZ 85721
U.S.A.
hchen@eller.arizona.edu

Abstract

Content analysis of computer-mediated communication (CMC) is important for evaluating the effectiveness of electronic communication in various organizational settings. CMC text analysis relies on systems capable of providing suitable navigation and knowledge discovery functionalities. However, existing CMC systems focus on structural features, with little support for features derived from message text. This deficiency is attributable to the informational richness and representational complexities associated with CMC text. In order to address this shortcoming, we propose a design framework for CMC text analysis systems. Grounded in systemic functional linguistic theory, the proposed framework advocates the development of systems capable of representing the rich array of information types inherent in CMC text. It also provides guidelines regarding the choice of features, feature selection, and visualization techniques that CMC text analysis systems should employ. The CyberGate system was developed as an instantiation of the design framework. CyberGate incorporates a rich feature set and complementary feature selection and visualization methods, including the writeprints and ink blots techniques. An application example was used to illustrate the system's ability to discern important patterns in CMC text. Furthermore, results from numerous experiments conducted in comparison with benchmark methods confirmed the viability of CyberGate's features and techniques. The results revealed that the CyberGate system and its underlying design framework can dramatically improve CMC text analysis capabilities over those provided by existing systems.

Keywords: Computer-mediated communication, design framework, text analysis systems, design science, information visualization

Introduction

Computer-mediated communication (CMC) has seen tremendous growth due to the fast propagation of the Internet. Text-based modes of CMC include e-mail, listservs, forums, chat, and the World Wide Web (Herring, 2002). These CMC modes have redefined the fabric of organizational culture and interaction. With the persistent evolution of communication

[1] Roger Chiang was the guest associate editor for this paper.

Note: The figures in this paper are available in color on the *MIS Quarterly* website at http://www.misq.org/archivist/vol/no32/issue4/AbbasiFigures.pdf.

processes and constant advancements in technology, such metamorphoses are likely to continue. An important trend has been the increased use of online communities: communities interacting virtually via CMC (Cothrel 2000). Online communities provide invaluable support for various business operations including organizational communication, knowledge dissemination, transfer of goods and services, and product reviews (Cothrel 2000). Electronic communities (Wenger and Snyder 2000) and networks of practice (Wasko and Faraj 2005) enable companies to tap into the wealth of information and expertise available across corporate lines. Virtual teams and group support systems facilitate organizational operations regardless of physical boundaries (Fjermestad and Hiltz 1999; Montoya-Weiss et al. 2001). Internet marketplaces allow the efficient transfer of goods and services and offer a medium for consumer feedback equally useful to potential customers and marketing departments (Turney and Littman 2003).

In spite of the numerous benefits of CMC, it is not without its pitfalls. Two characteristics have proven to be particularly problematic: the lack of control on information quality and the enormity and complexity of data present in CMC. Newsgroups and knowledge exchange communities suffer from lurkers and agitators that decrease the signal to noise ratio in CMC, casting doubts onto the reliability of information exchanged (Wasko and Faraj 2005; Viegas and Smith 2004). Additionally, online communities encompass very large scale conversations involving thousands of users (Herring 2002; Sack 2000). The enormous information quantities make such places difficult to navigate and analyze (Viegas and Smith 2004).

CMC text analysis is the analysis of text-based modes of CMC. There is a need for analysis techniques that can evaluate, summarize, and present CMC text. Systems capable of navigation and knowledge discovery can enhance information transparency (Sack 2000; Wellman 2001). Tools supporting social translucence via the measurement of social accounting data in CMC may improve information quality and analysis capabilities, a condition mutually beneficial to online community members and researchers/analysts (Erickson and Kellogg 2000; Smith 2002). Consequently, numerous CMC systems have been developed to address these needs (Fiore and Smith 2002; Viegas and Smith 2004; Xiong and Donath 1999). These systems generally visualize data provided in the message headers, such as interaction (send/reply structure) and activity (posting patterns) based information. Little support is provided for analysis of information contained in the message body text. In the instances where text analysis is provided, simple feature representations such as those used in information retrieval systems are utilized (Mladenic 1999; Sack 2000). CMC text is rich in social cues including emotions, opinions, style, and genres (Hara et al. 2000; Henri 1992; Yates and Orlikowski 2002). Improved CMC text analysis capabilities based on richer text representations are necessary (Paccagnella 1997). CMC analysis systems often neglect message text due to representational and presentation complexities. Thus, design guidelines for CMC systems supporting text analysis are needed (Sack 2000). Using Walls et al.'s (1992) model, this paper proposes a design framework for CMC text analysis systems. Grounded in system functional linguistic theory, the framework calls for the development of systems that support various information types found in CMC text. Based upon it, we developed the CyberGate system. Cyber Gate incorporates various features, feature selection, and visualization techniques, including the writeprints and ink blots techniques.

The remainder of the paper is organized as follows. We first highlight the unique characteristics of text-based CMC and provide a review of systems developed to support CMC text analysis. We then describe challenges associated with CMC text and present an overview of our design framework. Subsequent sections elaborate on the components of the design framework. A description of the CyberGate system (developed as an instantiation of our framework) is then offered. The two ensuing sections provide application examples and experimental evaluations of the CyberGate system and its underlying framework. We conclude with a summary of our research contributions and potential future directions.

Background

Many studies have expounded upon the significance of CMC text analysis for analyzing organizations (Chia 2000). Online discourse via computer mediation has resulted in new forms of communicative practice worthy of in-depth analysis (Wilson and Peterson 2002). CMC text analysis is important for evaluating the effectiveness and efficiency of electronic communication in various organizational settings, including virtual teams and group support systems. Analysis of CMC text also plays a crucial role in facilitating the measurement of return on investment for various online communities including electronic communities and networks of practice. CMC and other Internet related technologies are not a source of business value by themselves. They require the utilization of ancillary IT resources (Barua et al. 2004). Paccagnella (1997) emphasized the need for such systems supporting CMC analysis (pp. 4-5), stating that

> deep, interpretative research on virtual communities could be greatly helped by an accurate use of new

> analytic, powerful yet flexible tools, exploiting the possibility of cheaply collecting, organizing and exploring digital data.

In the remainder of this section, we describe the unique characteristics of CMC text that differentiate it from other text documents. We also review prior CMC systems and emphasize the need for ones supporting enhanced text analysis.

CMC Text

Computer-mediated communication text has several unique characteristics that differentiate it from non-CMC documents (e.g., essays, reports, news articles, resumes, research papers). Three of these distinct properties are described here.

(1) The communicative nature of CMC text makes it rich in interaction (Sack 2000), while non-CMC documents are generally devoid of interaction information. Asynchronous and synchronous forms of CMC both contain high levels of interaction, with the specific discourse patterns and dynamics varying depending on the communication context and CMC mode (Fu et al. 2008; Herring 2002).

(2) CMC text also differs from non-CMC documents with respect to its informational composition. While non-CMC documents have a high concentration of topical information (Mladenic 1999), such information is less pervasive in CMC. Nigam and Hurst (2004) analyzed thousands of messages posted on USENET (a large collection of newsgroups), and found that only 3 percent of sentences contained topical information. In contrast, web discourse is rich in opinion and emotion related information (Nigam and Hurst 2004; Subasic and Huettner 2001).

(3) CMC text and non-CMC documents also differ linguistically, with new CMC technologies bringing about the emergence of novel language varieties (Wilson and Peterson 2002). CMC text encompasses a large spectrum of stylistic, genre-based, and idiosyncratic language usage attributable to age, gender, educational, cultural, and contextual differences (Herring 2002; Sack 2000).

CMC Text Analysis Features

The richness of CMC has brought about the emergence of many types of CMC text analysis. These include analysis of participation levels, interaction, social cues, topics, user roles, linguistic variation, types of questions posed, and response complexity (Hara et al. 2000; Henri 1992). The features (i.e., attributes) utilized for CMC text analysis can be broadly categorized as either structural or text-based.

Structural features are attributes based on communication topology. These features are extracted solely from message headers, without any use of information contained in the message body (Sack 2000). Structural features support activity and interaction analysis. Posting activity related features include number of posts, number of initial messages, number of replies, and number of responses to a particular author's posts (Fiore and Smith 2002). These features can be used to represent an authors' social accounting metrics (Smith 2002). Analysis of activity based attributes also provides insight into different roles played by online community members, such as debaters, experts, and disseminators (Viegas and Smith 2004; Zhu and Chen 2002). Features used for interaction analysis include the frequency of incoming and outgoing messages. These features are used as input for the construction of social networks based on who is talking to whom (Sack 2000; Smith and Fiore 2001).

Text features are attributes derived from the message body. Although the informational richness of CMC text was previously questioned (Daft and Lengel 1986), numerous studies have since demonstrated the opulence of CMC text (Lee 1994; Yates and Orlikowski 2002). In addition to topical information, CMC text is rich in social cues (Henri 1992), power cues (Panteli 2002), and genres (Yates and Orlikowski 2002). Social cues are elements not related to formal content or subject matter (Henri 1992). Examples include self-introductions, expressions of feelings, greetings, signatures, jokes, use of symbolic icons, and compliments (Hara et al. 2000). CMC text also contains evidence of power cues; stylistic indicators of one's position/ rank within an organization or online community (Panteli 2002). Genres are types of writing based on purpose and form (e.g., memos, meetings, reports, etc.). Highly prevalent in CMC, they serve as sources of organizing structures and communicative norms (Yates and Orlikowski 2002).

Inclusion of structural and text-based features is critical for CMC text analysis. For instance, online community sustainability analysis requires the use of communication activity, interaction, and text content attributes (Butler 2001). Similarly, Cothrel (2000) incorporated structural features (activity measures) and text features (discussion topics) into his model for measuring an online community's return on investment. He noted that activity measures describe the general health of a community while discussion topic metrics "assess the ongoing insights that the community offers into the business's products or processes" (p. 19).

Table 1. Previous CMC Systems

		Feature Types		
System Name	Reference	Structural	Text-based	Feature Descriptions
Chat Circles	Donath et al. 1999	√	√	length, headers
Loom	Donath et al. 1999	√	√	terms, punc., headers
People Garden	Xiong and Donath 1999	√		headers
Babble	Erickson and Kellogg 2000	√		headers
Conversation Map	Sack 2000	√	√	semantic, headers
NetScan	Smith and Fiore 2001	√		headers
Coterie	Donath 2002	√		headers
Newsgroup Treemaps	Fiore and Smith 2002	√		headers
Communication Garden	Zhu and Chen 2002	√	√	noun phrases, headers
PostHistory	Viegas et al. 2004	√		headers
Social Network Fragments	Viegas et al. 2004	√		headers
Authorlines	Viegas and Smith 2004	√		headers
Newsgroup Crowds	Viegas and Smith 2004	√		headers

CMC Text Analysis Systems

CMC systems can be sorted into two categories based on functionality: those that support the communication process and those that support analysis of communication content (Sack 2000). While it is certainly possible for a single system to support both functions (e.g., Erickson and Kellogg 2000), we focus only on the analysis functionalities provided by these systems due to their relevance to CMC text analysis. Table 1 provides a review of prior CMC systems supporting analysis of text-based CMC. The review is based on the analysis features incorporated by these systems.

A plethora of CMC systems have been developed to support structural features. Several tools visualize posting activity patterns, such as Loom (Donath et al. 1999) and Authorlines (Viegas and Smith 2004). PeopleGarden and Communication Garden both use garden metaphors with flower glyphs to display author and thread activity (Xiong and Donath 1999; Zhu and Chen 2002). The number of petals and thorns, petal colors, and stem lengths are used to represent activity features such as the total number of posts and number of threads in which an author has participated. Babble (Erickson and Kellogg 2000) and Coterie (Donath 2002) are both geared toward showing activity patterns in persistent conversation. In these systems, all participants are displayed in a two-dimensional space. More active authors are shown in the center while participants with fewer postings gradually shift to the perimeter. The visual effect is a good method for identifying active participants versus lurkers (Donath 2002). Systems displaying interaction information also exist. Conversation Map visualizes social networks based on send/ reply patterns (Sack 2000). NetScan displays message and author interactions (Smith and Fiore 2001), while Loom shows thread-level interaction structures (Donath et al. 1999).

Previous CMC systems offer limited support for text-based features. Loom shows some content patterns based on message moods. Moods are assigned by taking into consideration the occurrence of certain terms and punctuation in the message text. Chat Circles displays messages based on body text length. Conversation Map and Communication Garden provide more in-depth topical analysis. Conversation Map uses computational linguistics to build semantic networks for discussion topics while Communication Garden performs topic categorization using noun phrases.

Text systems are a related class of systems that are used for either information retrieval (IR systems) or general text categorization and analysis (text mining systems). However, IR systems are more concerned with information access than analysis (Hearst 1999). Mladenic (1999) presented a review of 29 IR systems, all of which used bag-of-words to represent text document topics. Similarly, Tan (1999) reviewed 11 commercial text mining systems and found IBM's Intelligent Miner to be the most comprehensive. However, this system also utilizes limited feature representations (i.e., bag-of-words, named entities) and only performs topic categorization and analysis (Dorre et al. 1999).

Overall, the features used in existing CMC text analysis systems are insufficient to effectively capture text-based content in CMC (Sack 2000). Paccagnella suggested that computer programs to support CMC text analysis would be helpful, yet do not exist. He noted numerous ways in which automated systems could benefit CMC text analysis, including data linking, content analysis, data display, and graphic mapping. Without appropriate CMC text analysis systems, text features are often overlooked (Panteli 2002). There has been limited analysis of CMC text since manual methods are time consuming (Hara et al. 2000). Cothrel (2000) stated that discussion content is an essential dimension of online community success measurement, yet proper definition and measurement remains elusive.

A Design Framework for CMC Text Analysis

Given the need for CMC text analysis and a lack of systems that address this need, an important and obvious question arises. Why do most CMC systems support structural features but neglect text content features? There are three major differences that are likely responsible for the disparity between the numbers of systems representing these feature types, including feature definitions, extraction, and presentation. Structural features are well defined, easy to extract, and easy to visualize. Appropriate activity (Fiore and Smith 2002) and interaction based features have been established in the sociology literature. These features are also easy to extract and visualize using bar chart variants for activity frequency (Viegas and Smith 2004; Xiong and Donath 1999) and networks for interaction (Donath et al. 1999; Smith and Fiore 2001). In contrast text features are loosely defined, difficult to extract, and harder to present to end users. The richness of CMC text necessitates a complex set of text features (Donath et al. 1999). For example, over 1,000 text features have been used for analyzing style, with no consensus (Rudman 1997). Additionally, text feature extraction can be challenging due to high noise levels in CMC text (Nasukawa and Nagano 2001). Finally, the informational richness of text requires multiple complementary presentation views (Keim 2002; Losiewicz et al. 2000). Different techniques have been developed to support various facets of text visualization with no ideal solution (Huang et al. 2005; Miller et al. 1998; Wise 1999).

In light of these challenges, Sack (2000) argues for a new CMC system design philosophy that incorporates automatic text analysis techniques. He states

> it is necessary to formulate a complementary design philosophy for CMC systems in which the point is to help participants and observers spot emerging groups and changing patterns of communication (p. 86).

Design guidelines are needed due to the lack of previous tools supporting in depth CMC text analysis, the complexity associated with properly representing CMC text, and the lack of consensus regarding appropriate text features and presentation formats.

According to the design science paradigm, design is a product and a process (Hevner et al. 2004; Walls et al. 1992). Development of a design framework for CMC text analysis systems requires consideration of the design product and design process. The design product is the set of requirements and necessary design characteristics that should guide IT artifact construction. An IT artifact can be a construct, method, model, or instantiation (Hevner et al. 2004). The design process is composed of the steps and procedures taken to develop the artifact.

Information systems development typically follows an iterative design process of building and evaluating (March and Smith 1995), which is analogous to the generate/test cycle proposed by Simon (1996). Such an approach is particularly important in design situations involving complex or vaguely defined user requirements (Markus et al. 2002). We believe that the ambiguities associated with CMC text analysis also warrant the use of an iterative design process. Hence, we focus on the design product. Walls et al. (1992) presented a model for the formulation of information systems design theories (ISDTs). Their model incorporates four components guiding the design product aspect of an ISDT. These include the kernel theories, meta-requirements, meta-design, and testable hypotheses (shown in Table 2). The kernel theories govern meta-requirements for the design product. The meta-design is anticipated to fulfill these meta-requirements by providing detailed specifications for the class of IT artifacts addressed by the design product. Testable hypotheses are used to evaluate how well the meta-design satisfies meta-requirements. A good example of an ISDT design product is the relational database (Walls et al. 1992). Using relational database theory as a kernel theory, meta-requirements are the elimination of insertion, update, and deletion anomalies. The meta-design consists of a set of tables in at least third normal form. The testable hypotheses are theorems and proofs validating the normalized database tables as being devoid of any anomalies.

Using Walls et al.'s model, we propose a design framework for CMC text analysis systems (shown in Table 3). Employ-

Table 2. Components of an ISDT Design Product (Adapted from Walls et al. 1992)

1. Kernel theories	Theories from natural or social sciences governing design requirements
2. Meta-requirements	Describes a class of goals to which theory applies
3. Meta-design	Describes a class of artifacts hypothesized to meet meta-requirements
4. Testable hypotheses	Used to test whether meta-design satisfies meta-requirements

Table 3. Components of the Proposed Design Framework for CMC Text Analysis Systems

1. Kernel theory	Systemic functional linguistic theory (SFLT)
2. Meta-requirements	Support for various information types found in CMC text that represent the ideational, textual, and interpersonal meta-functions.
3. Meta-design	The incorporation of a rich set of text features coupled with appropriate feature selection and visualization methods; collectively capable of representing the ideational, textual, and interpersonal meta-functions. Specific meta-design elements are as follows: • Utilization of extended feature set comprised of language and processing resources. • Use of ranking and projection based feature selection techniques. • Inclusion of multidimensional, text overlay, and interaction visualization methods.
4. Testable hypotheses	Empirical evaluation of the features and selection/visualization techniques' ability to accurately represent information types associated with the three meta-functions. Specific testable hypotheses are as follows: • Ability of the features, feature selection, and visualization methods to characterize information types associated with the three meta-functions. • Ability of the features, feature selection, and visualization methods to discriminate information types associated with the three meta-functions.

ing systemic functional linguistic theory as our kernel theory, we propose meta-requirements and a meta-design necessary to support CMC text analysis. We also present hypotheses intended to evaluate how well the meta-design satisfies our meta-requirements. The ensuing sections elaborate on the components of our design framework.

Kernel Theory

Perhaps the most important characteristic of CMC is the language complexities it introduces as compared to other forms of text (Wilson and Peterson 2002). Effective analysis of CMC text entails the utilization of a language theory that can provide representational guidelines. Grounded in functional linguistics, systemic functional linguistic theory (SFLT) provides an appropriate mechanism for representing CMC text information (Halliday 2004). SFLT states that language has three meta-functions: ideational, interpersonal, and textual. The three meta-functions are intended to provide a comprehensive functional representation of language meaning by encompassing the physical, mental, and social elements of language (Fairclough 2003).

The *ideational* meta-function states that language consists of ideas. According to Halliday (2004), the ideational meta-function of language suggests that a message is "about something" or "construing experience" (p. 30). It pertains to the use of "language as reflection" (p. 29). The ideational meta-function relates to aspects of the "mental world" which include attitudes, desires, and values (Fairclough 2003; Halliday 2004).

The *textual* meta-function indicates that language has organization, structure, flow, cohesion, and continuity (Halliday 2004). It relates to aspects of the "physical world" pertaining to the manner in which ideas are communicated (Fairclough 2003; Halliday 2004). The textual meta-function therefore serves as a facilitating function enabling the conveyance of the ideational and interpersonal meta-functions. It can be present via information types such as style, genres, and vernaculars (Argamon et al. 2007). For instance, an author biography and vita may convey similar ideational meaning about one's educational background and career accomplishments using contrasting textual functions, in this case due to genre differences.

The *interpersonal* meta-function refers to the fact that language is a medium of exchange between people (Sack 2000). It pertains to the use of "language as action" (Halliday 2004, p. 30). The interpersonal meta-function is concerned with the enactment of social relations; it relates to aspects of the "social world" (Fairclough 2003; Halliday 2004). It is generally represented using CMC interaction information.

Meta-Requirements

Analysis of CMC text requires the inclusion of all three language meta-functions described by SFLT: ideational, textual, and interpersonal. "Any summary of a very large scale conversation is incomplete if it does not incorporate all three of these meta-functions (ideational, interpersonal, and textual)" (Sack 2000, p. 75). Therefore, effective depiction of CMC text entails consideration of information types capable of representing these three meta-functions.

The ideational meta-function in CMC text can be manifested in the form of various information types, including topics, events, opinions, and emotions. Topics are the most commonly represented information type in text (Mladenic 1999; Tan 1999). Events are specific incidents with a temporal dimension. While "hurricane" is a topic, "Hurricane Katrina" is an event. Event detection has garnered significant attention in recent years, although it continues to present challenges since effective representation of events in text remains elusive (Allan et al. 1998). Additional information types representing the ideational meta-function include opinions and emotions. Opinions include sentiment polarities (e.g., positive, neutral, negative) and intensities (e.g., high or low) about a particular target (Pang et al. 2002). Popular applications of opinion-related information include mining online movie and product reviews for consumer preference information (Turney and Littman 2003). CMC text is also rich in emotional information (Picard 1997). Emotions encompassed in online communication consist of various affects such as happiness, sadness, horror, and anger (Subasic and Huettner 2001).

Styles, genres, and vernaculars are information types representing the textual meta-function. Style is based on the literary choices an author makes, which can be a reflection of context (who, what, when, why, where) and personal background (education, gender, etc.). Example styles are formal (use of greetings, structured sentences, paragraphs) and informal (no sentences, no greetings, erratic punctuation, use of slang). Stylistic information is utilized in numerous forms of analysis. Authorship analysis identifies and characterizes individuals based on their writing style (Zheng et al. 2006). Deception detection attempts to determine if an individual's writing is deceitful (Zhou et al. 2004), while power cue identification explores the writing style differences between superiors and subordinates in organizational settings (Panteli 2002). Genres are classes of writing. Genres found in CMC include inquiries, informational messages, memos, reports, interview transcripts, and feedback comments (Santini 2004; Yates and Orlikowski 2002).

The interpersonal meta-function is generally represented by CMC interaction information (i.e., who is communicating with whom). Interaction information can be derived from message headers for certain CMC modes such as e-mail and blogs. In e-mail, the "RE:" in the message subject coupled with the presence of quoted content are salient interaction cues. However other CMC modes (e.g., chat rooms, instant messaging, web forums) require the use of text interaction cues inherent in the body text. Text-based interaction cues include direct references to fellow users' names, references to previously posted content, and conjunction and ellipsis based cues indicating continuation of an existing conversation between users (Fu et al. 2008; Sack 2000). Interaction information is useful for social network analysis and evaluation of conversation streams based on communication thread patterns (Smith and Fiore 2001).

The three meta-functions and their associated information types are interrelated and should not be considered in isolation from one another. For instance, an analyst may be concerned with opinions regarding a particular topic, or the stylistic tendencies for two interacting participants' text. Table 4 shows examples for information types that represent the three meta-functions, and their related analysis applications. The following section presents a meta-design for how the three meta-functions can be supported by accurately representing their corresponding information types.

Meta-Design

While meta-requirements are derived from the kernel theories, the objective of the meta-design is to introduce a class of artifacts hypothesized to meet the meta-requirements (Walls et al. 1992). Three critical elements of any text mining, text analysis, or information retrieval system are their features,

Table 4. Various Information Types for the Three Meta-Functions

Meta-Function	Info. Types	Analysis Types	References
Ideational	Topics	Topical Analysis	Chen et al. 2003; Mladenic 1999
	Events	Event Detection	Allan et al. 1998
	Opinions	Sentiment Analysis	Argamon et al. 2007; Turney & Littman 2003
	Emotions	Affect Analysis	Picard 1997; Subasic & Huettner 2001
Textual	Style	Authorship Analysis Deception Detection Power Cues	Abbasi & Chen 2006; Zheng et al. 2006; Zhou et al. 2004 Panteli 2002
	Genres	Genre Analysis	Santini 2004; Yates & Orlikowski 2002
	Vernaculars	Semantic Networks	Koppel & Schler 2003; Sack 2000
Interpersonal	Interaction	Social Networks	Sack 2000; Viegas et al. 2004
		Conversation Streams	Smith & Fiore 2001

feature selection methods, and visualization techniques (Chen 2001; Cunningham 2002; Mladenic 1999; Tan 1999). For CMC text analysis, the meta-design requires the incorporation of an extended set of linguistic features (Cunningham 2002; Mladenic 1999) capable of representing various information types associated with the ideational, textual, and interpersonal meta-functions. Feature selection methods present features in a ranked and/or reduced state for improved knowledge discovery (Guyon and Elisseeff 2003). Feature selection techniques are necessary for enhancing the representational richness of various information types present in CMC text (Mladenic 1999). Although a large number of linguistic features are needed for CMC text analysis, only a subset of these may be relevant or useful for a particular information type (Forman 2003; Hearst 1999). Furthermore, visualization techniques are needed for effective analysis of CMC text (Chen 2001; Tan 1999). Such methods are capable of presenting important CMC text information in a concise and informative manner (Keim 2002; Wise 1999). In the subsequent sections, we review the merits of potential meta-design alternatives for features, feature selection, and visualization.

Features for CMC Text Analysis

Text features are linguistic attributes used to represent various information types. They can be classified into two broad categories: language resources and processing resources (Cunningham 2002). Language resources are data-only resources such as lexicons, thesauruses, and word lists. These self-standing features exist independent of their application context and provide powerful discriminatory potential. However, language resource construction is often manual, and features may be less generalizable across information types (Pang et al. 2002).

Processing resources require algorithms for computation. Parts-of-speech tags, n-grams, statistical features (e.g., average word length), and bag-of-words are all examples of processing resources. The majority of processing resource features are context-dependent; they change according to the text corpus. However, the extraction procedures remain constant, making processing resources highly generalizable across information types. Consequently, features such as bag-of-words, part-of-speech tags, and n-grams are used to represent numerous information types including topics, events, opinions, style, and genres (Pang et al. 2002; Santini 2004).

Using language and processing resources in conjunction can improve text categorization and analysis capabilities since processing resources provide breadth across information types while language resources offer depth within specific information types (Cunningham 2002). Table 5 provides a summary of numerous language and processing features as well as the information types these feature groups have been used to represent. The table can be read as follows: syntactic language resources, including function words, punctuation, and special characters have been used to represent opinion, style, genre, and interaction information in text. While most of the feature descriptions are straightforward, certain categories (e.g., lexical) are more involved. Interested readers can attain further details about these feature groups from prior studies (Koppel and Schler 2003; Zheng et al. 2006).

Table 5. Various Linguistic Features Used for Text Analysis

Resource	Category	Feature Group	Examples	Info. Type
Language	Syntactic	Function Words	of, for, the, on, who, what, because	Opinions Style Genres Interaction
		Punctuation	!, ?, :, ”	
		Special Characters	$,@,#,*,&	
	Structural	Technical Structure	file extensions, font colors, sizes	Style
	Lexicons	Sentiments	positive/negative term lists	Opinions
		Affect Classes	happiness, anger, hate, etc. terms	Emotions
		Idiosyncrasies	misspelled word lists, vernaculars	Style
		Geographic	lists of places (e.g., states, cities)	Events
		Temporal	time references (e.g., day, month)	Events
	Thesaurus	Synonyms	synonymy information for words	Opinions Emotions Style
Processing	Lexical	Word Lexical	total words, % char. per word	Opinions Style Genres
		Character Lexical	total char., % numeric char.	
		Vocabulary Richness	hapax legomana, Yules K	
		Word Length Dist.	frequency of 1-20 letter words	
		Character N-grams	at, att, atta, attai	
		Digit N-grams	12, 94 192	
	Syntactic	POS Tag N-Grams	NNP_VB VB,VB ADJ	Topics Events Opinions Style Genres Interaction
		Word N-grams	went to, to the, went to the	
	Semantic	Noun Phrases	account, bonds, stocks	
		Named Entities	Enron, Cisco, El Paso, California	
		Bag-of-Words	all words except function words	
	Structural	Document Structure	has greeting, url, quoted content	Style

Feature Selection Techniques for CMC Text Analysis

Two categories of feature selection techniques commonly applied to text are ranking and projection based methods (Guyon and Elisseeff 2003). Ranking techniques rank attributes based on some heuristic (Hearst 1999). Examples include information gain, chi-squared, and Pearson’s correlation coefficient (Forman 2003; Koppel and Schler 2003). Projection methods are transformation based techniques that utilize dimensionality reduction (Huang et al. 2005). Examples are principal component analysis (PCA), multidimensional scaling (MDS), and self-organizing map (Chen et al. 2003; Huang et al. 2005). Ranking and projection based methods each have their advantages and disadvantages.

Ranking methods have been used to analyze several information types, including topics, style, and opinions (Abbasi and Chen 2005; Pang et al. 2002). They offer greater explanatory potential than projection methods since they preserve the original feature set and simply rank/sort attributes (Seo and Shneiderman 2005). Ranking methods also offer simplicity and scalability. However, they typically consider only an individual feature’s predictive power; resulting in the potential loss of information stemming from feature interactions (Guyon and Elisseeff 2003).

Projection methods have been used to transform text feature spaces into lower dimensional projections for style and topic categorization (Abbasi and Chen 2006; Allan et al. 2001; Chen et al. 2003). Projection methods are highly robust against

Table 6. Examples of Ranking and Projection Based Feature Selection Methods Applied to Text

Selection Method	Example Technique	Info. Type	Reference
Ranking	Information Gain	Topics	Koppel & Schler 2003
	Chi-Squared	Topics	Forman 2003
	Decision Tree Model	Style	Abbasi & Chen 2005
	Minimum Frequency	Opinions	Pang et al. 2002
Projection	Principal Component Analysis	Style	Abbasi & Chen 2006
	Multidimensional Scaling	Topics	Allan et al. 2001
	Self-Organizing Map	Topics	Chen et al. 2003

noise, making them useful for text analysis. They can uncover important underlying patterns (Abbasi and Chen 2006). However, the transformation process from original features to projections can also diminish explanatory potential (Seo and Shneiderman 2005). Projection methods may describe important high-level patterns but have difficulty explaining details about specific features.

The rank-by-feature framework states that systems designed to support complex analysis tasks should incorporate divergent feature selection methods to enhance analysis capabilities (Seo and Shneiderman 2005). For instance, using ranking and projection methods in unison (i.e., independently applying them to the same data) can facilitate analysis of overview (projection methods) and specific feature details (ranking methods). Therefore, CMC text analysis systems should employ both categories of feature selection techniques**.** Table 6 shows examples of ranking and projection based methods applied to various information types.

Visualization Techniques for CMC Text Analysis

CMC text analysis systems should present interaction information using network and tree representations as done in prior systems (Sack 2000; Smith and Fiore 2001). However, visualization of text information derived from message bodies is challenging since text cannot easily be described by numbers (Keim 2002). Visualization of complex high dimensional information can be enhanced using coordinated views, that is, multiple complementary presentation formats (Andrienko and Andrienko 2003; Losiewicz et al. 2000). Wise (1999) noted that text analysis should

> provide a basis for altered visualization of the information for different users and purposes…why should we preconceive that there is only one "correct" visualization of text information in a document corpus? (p. 1230).

For instance, text itself is one-dimensional, textual features are multidimensional (Huang et al. 2005), and the relation between features and the text they represent is often established using two- or three-dimensional text overlay (Cunningham 2002). CMC text analysis systems can dramatically benefit from complementary presentation formats including multidimensional and text overlay methods (Keim 2002; Wise 1999). These two categories of visualization techniques are described below.

Multidimensional techniques used for text visualization include graphs and reduced dimensionality views. Graphical formats such as radar charts, parallel coordinates, and scatter plot matrices have been applied to topic, affect, and style information (Huang et al. 2005; Subasic and Huettner 2001). Reduced dimensionality visualizations decrease the feature space to show essential patterns. These techniques are typically used in conjunction with projection-based feature selection techniques to create two or three dimensional views. Examples include writeprints (Abbasi and Chen 2006), ThemeRiver© (Havre et al. 2002), and Themescapes™ (Wise 1999). Text overlay methods combine text with feature occurrence patterns to provide greater insight. The Stereoscopic Document View in Topic Islands™ uses wavelet transformations to show key topical patterns, superimposed onto the document text (Miller et al. 1998). Text annotation highlights feature occurrences in text (Cunningham 2002).

Multidimensional views are often used to visualize text feature statistics such as frequency, variance, and similarity (Keim 2002). While these views provide important insight and summarization capabilities, they abstract away from the underlying nonnumeric content they are intended to represent. Multidimensional techniques can tell us what features are important but not how or why. In contrast, text overlay techniques serve an important complementary function. They have greater explanatory potential, allowing users to see exactly how and where features occur within their proper context. Hence it is important to include multidimensional presentation formats that can summarize feature statistics as well as text overlay illustrations that can bridge the gap between feature statistics and their actual occurrences in text.

Testable Hypotheses

Testable hypotheses are intended to assess whether the meta-design satisfies meta-requirements (Walls et al. 1992). For the proposed design framework, this entails evaluating the meta-design's ability to accurately represent information types associated with the three meta-functions, as outlined in the meta-requirements. In text mining, *representation* can imply data characterization or data discrimination. Data characterization is "a summarization of the general characteristics or features of a target class of data" (Han and Kamber 2001, p. 21). From a data characterization perspective, a good representation is able to derive important patterns, trends, or phenomenon of interest from text information (Tan 1999). Data discrimination is clustering or categorizing of information types into meaningful classes (Chen 2001; Tan 1999). With respect to data discrimination, a good representation is one capable of accurately categorizing text into various information classes (Han and Kamber 2001).

For the proposed CMC text analysis design framework, a suitable meta-design must incorporate features, feature selection, and visualization techniques capable of effectively characterizing and discriminating information types used to represent the meta-functions. Prior CMC systems used application examples or case studies to illustrate their systems' data characterization capabilities (e.g., Erickson and Kellogg 2000; Sack 2000; Smith and Fiore 2001; Viegas and Smith 2004; Zhu and Chen 2002). In contrast, the effectiveness of data discrimination is generally assessed using rigorous text categorization experiments for various information types (Argamon et al. 2007; Pang et al. 2002; Zheng et al. 2006).

In the following section, we describe the CyberGate system developed as an instantiation of our design framework. We use CyberGate to evaluate the effectiveness of our meta-design. A brief application example is used to illustrate the system's ability to characterize information types associated with the meta-functions. Text categorization experiments are also conducted to test the meta-design's effectiveness for discriminating information types used to represent the ideational, textual, and interpersonal meta-functions.

System Design: The CyberGate System

Based on our design framework, we developed the CyberGate system for text analysis of CMC (Figure 1). The system was developed using a cyclical design process involving several iterations of adding and testing system components (March and Smith 1995; Simon 1996). The testing phase encompassed experiments for performance evaluation and feedback solicitations from CMC researchers and analysts. CyberGate supports features for representing several information types associated with all three meta-functions. It also uses various feature selection and visualization techniques, including writeprints and ink blots. We first present an overview of the CyberGate system and then provide details about the writeprints and ink blots techniques.

Information Types and Features

CyberGate supports several information types for representing the ideational, textual, and interpersonal meta-functions. These include topics, opinions, affects, style, genres, and interaction information. In order to capture such a bevy of information, several language and processing resources were incorporated (i.e., most of the features shown in Table 5). The language resources encompass sentiment and affect lexicons, word lists, and the WordNet thesaurus (Fellbaum 1998). Embedded processing resources include n-grams, statistical features, parts-of-speech, noun phrases, and named entities (Koppel and Schler 2003; Zheng et al. 2006).

Feature Selection

CyberGate uses both ranking and projection based feature selection methods. For feature ranking, it uses information gain and decision tree models (Abbasi and Chen 2005; Forman 2003). PCA and MDS projections are used for dimensionality reduction (Abbasi and Chen 2006; Huang et al. 2005). Figure 2 shows examples of the feature selections techniques used in CyberGate. The table on the left (a) shows the complete set of features while (b) shows the top two dimensions of the PCA projections (Ex and Ey) and (c) shows the decision tree model rankings.

Visualization

CyberGate includes multidimensional, and text overlay based visual representations. Multidimensional visualizations in-

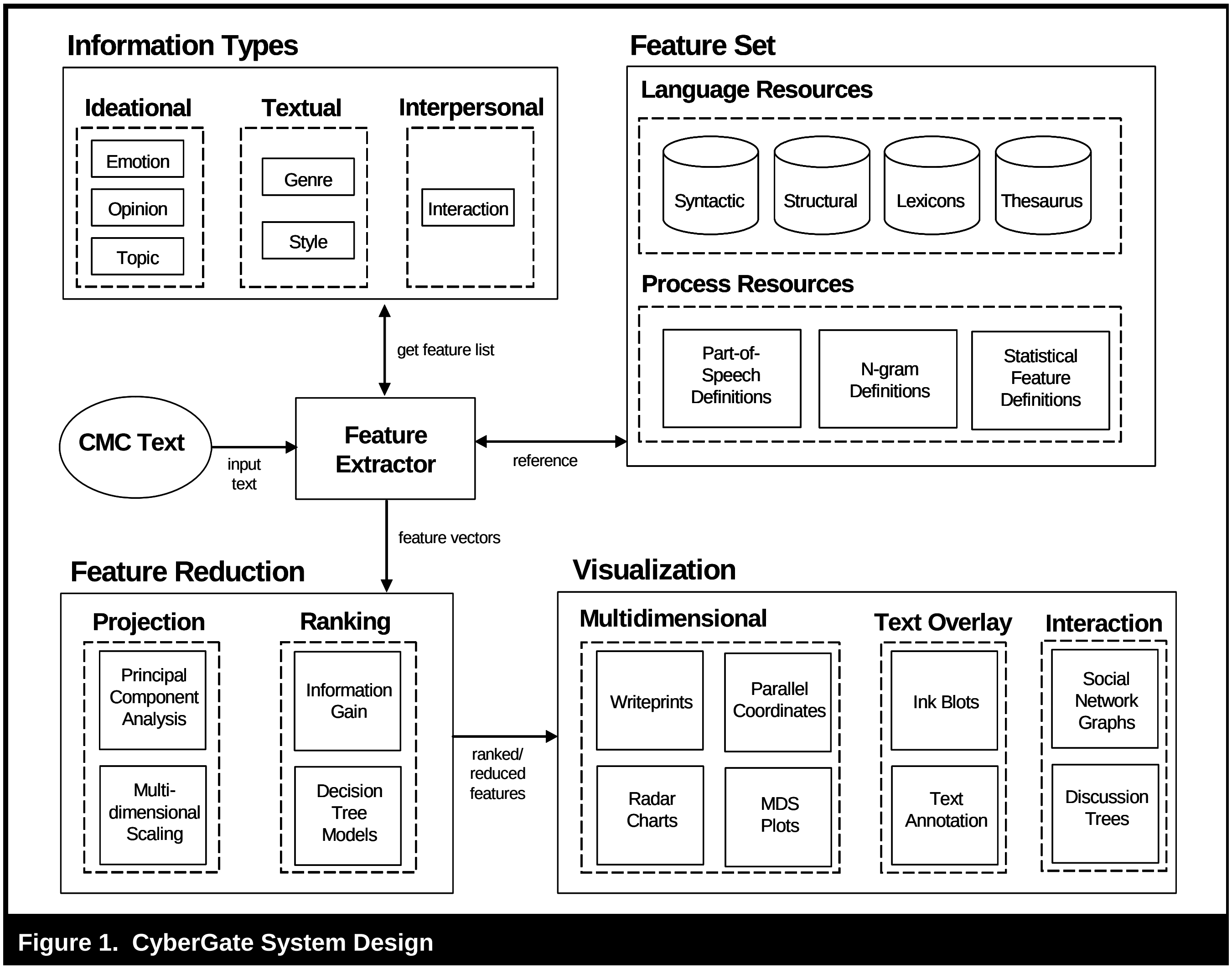
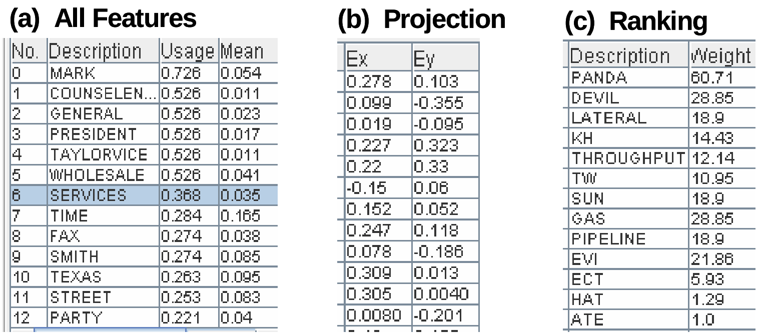


Figure 1. CyberGate System Design



(a) All Features

No.	Description	Usage	Mean
0	MARK	0.726	0.054
1	COUNSELEN...	0.526	0.011
2	GENERAL	0.526	0.023
3	PRESIDENT	0.526	0.017
4	TAYLORVICE	0.526	0.011
5	WHOLESALE	0.526	0.041
6	SERVICES	0.368	0.035
7	TIME	0.284	0.165
8	FAX	0.274	0.038
9	SMITH	0.274	0.085
10	TEXAS	0.263	0.095
11	STREET	0.253	0.083
12	PARTY	0.221	0.04

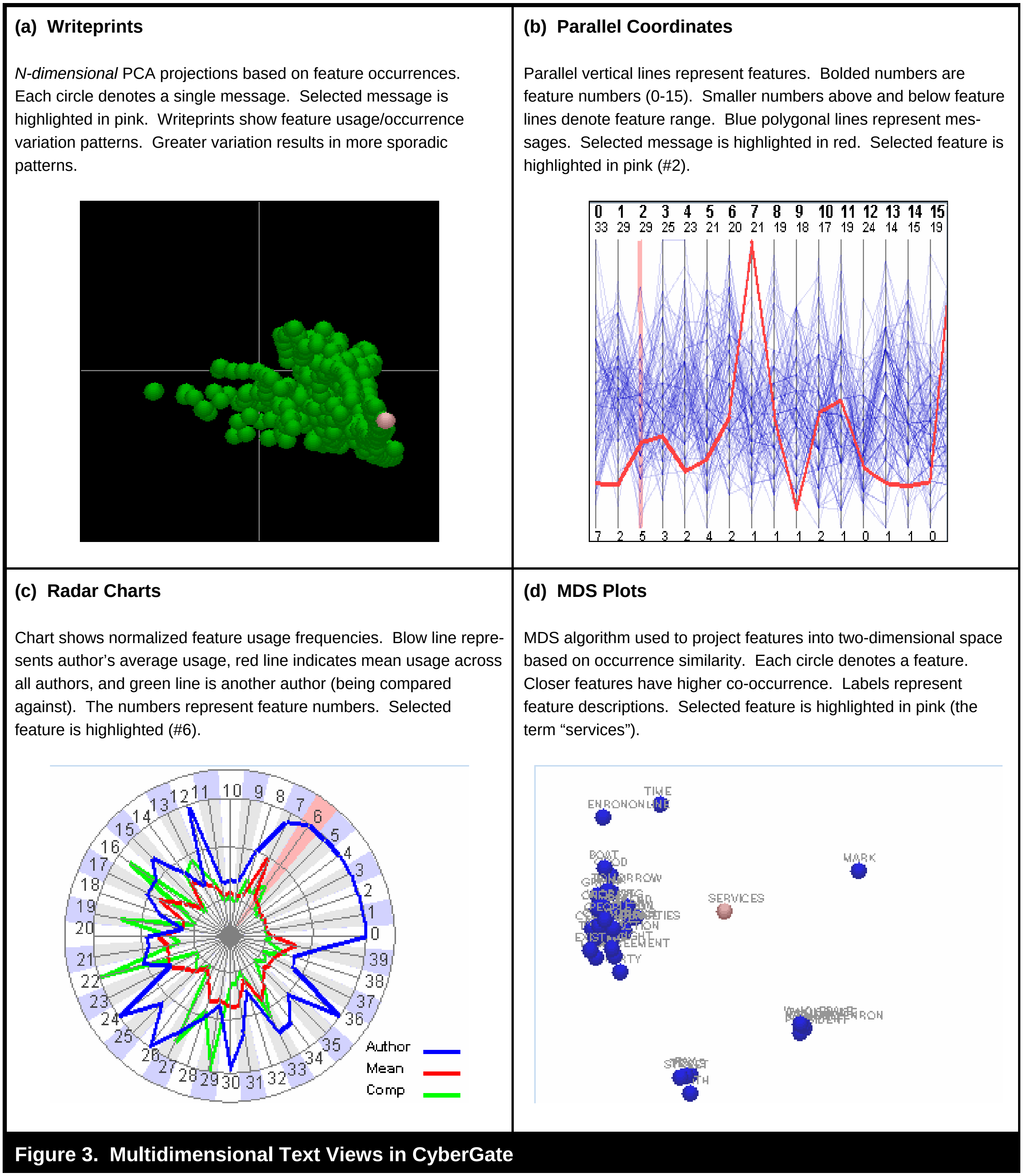
(b) Projection

Ex	Ey
0.278	0.103
0.099	-0.355
0.019	-0.095
0.227	0.323
0.22	0.33
-0.15	0.06
0.152	0.052
0.247	0.118
0.078	-0.186
0.309	0.013
0.305	0.0040
0.0080	-0.201

(c) Ranking

Description	Weight
PANDA	60.71
DEVIL	28.85
LATERAL	18.9
KH	14.43
THROUGHPUT	12.14
TW	10.95
SUN	18.9
GAS	28.85
PIPELINE	18.9
EVI	21.86
ECT	5.93
HAT	1.29
ATE	1.0

Figure 2. CyberGate Feature Selection Examples



(a) Writeprints

N-dimensional PCA projections based on feature occurrences. Each circle denotes a single message. Selected message is highlighted in pink. Writeprints show feature usage/occurrence variation patterns. Greater variation results in more sporadic patterns.

(b) Parallel Coordinates

Parallel vertical lines represent features. Bolded numbers are feature numbers (0-15). Smaller numbers above and below feature lines denote feature range. Blue polygonal lines represent messages. Selected message is highlighted in red. Selected feature is highlighted in pink (#2).

(c) Radar Charts

Chart shows normalized feature usage frequencies. Blow line represents author's average usage, red line indicates mean usage across all authors, and green line is another author (being compared against). The numbers represent feature numbers. Selected feature is highlighted (#6).

(d) MDS Plots

MDS algorithm used to project features into two-dimensional space based on occurrence similarity. Each circle denotes a feature. Closer features have higher co-occurrence. Labels represent feature descriptions. Selected feature is highlighted in pink (the term "services").

Figure 3. Multidimensional Text Views in CyberGate

clude writeprints, which shows usage variation, and parallel coordinates, which shows feature occurrences (Figures 3a and 3b). Each circle in writeprints denotes a single message or text window projected using principal component analysis. The blue polygonal lines in parallel coordinates also represent messages or text windows. The selected writeprints point corresponds to the selected parallel coordinates' polygonal line. The intersection between a polygonal line and a vertical axis in parallel coordinates represents the occurrence frequency of that feature in that particular message. For example, the selected message in Figure 3b has a high occurrence of feature #7 (occurs 21 times).

CyberGate also utilizes MDS plots (Figure 3d) to show overall feature similarities and radar charts (Figure 3c) for comparing feature occurrence statistics. The radar chart shown compares the selected author against another author and the mean normalized usage frequencies for a set of features (which are numbered along the perimeter). The MDS plot in Figure 3d shows features projected based on occurrence similarity for the bag-of-words features. We can see one large cluster and two smaller ones in addition to three or four features that are on their own. These features (e.g., services) do not frequently co-occur with any of the three clusters.

CyberGate's text overlay techniques are shown in Figure 4. Text annotation highlights key features in the text (Cunningham 2002). Figure 4a shows an example where the bag-of-words features are highlighted in blue while the selected feature (CounselEnron) is highlighted in red. Ink blots (Figure 4b) superimposes colored circles (blots) onto text for key features as identified by the underlying feature ranking method used. The size of the blot indicates the feature weight (based on the feature ranking technique). Features unique to a particular author have higher weights than ones that are equally common across authors. The color indicates the author's usage of the particular feature (red = high, blue = low, yellow = medium). The selected feature (again Counsel Enron) is highlighted with a black circle. This feature is represented with large red blots indicating that it has a high weight: it is unique to this author and frequently used.

CyberGate also includes graph and tree visualizations for viewing interaction information in CMC text. Author and thread social networks show the interaction between author nodes represented using links (Figures 5a and 5b). Discussion trees (Figure 5c) denote the interaction between subsequent message nodes within a thread. In addition to deriving interaction information from message headers, CyberGate also utilizes body text features (Fu et al. 2008). These features include the occurrence of user names and keywords that serve as indicators of user interaction. The use of text-based interaction features allows CyberGate to construct interaction patterns even when structural features are unavailable or insufficient (e.g., chat rooms and web forums).

Writeprints and Ink Blots

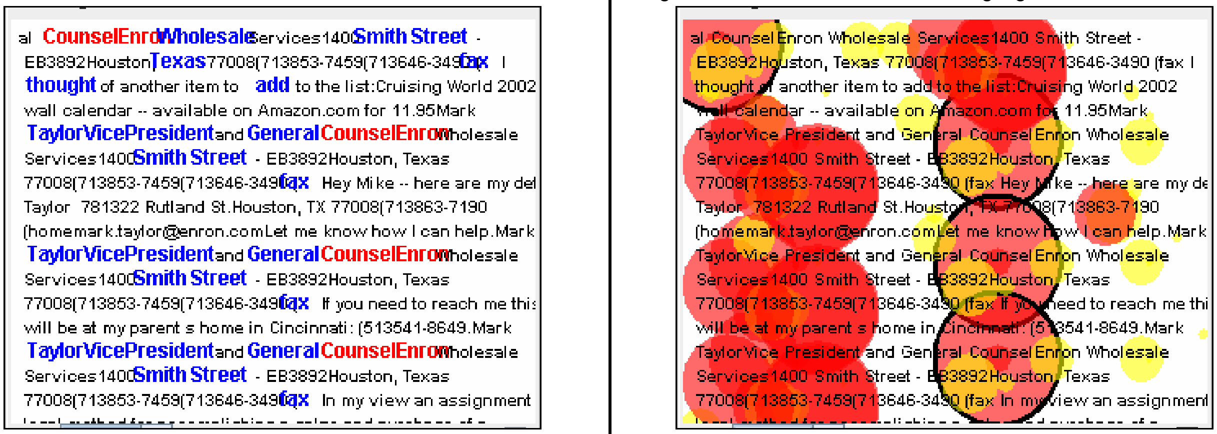
CyberGate includes the writeprints and ink blots techniques, which are the core components driving the system's analysis functions. These techniques epitomize the essence of the proposed design framework: representation of rich features using divergent feature selection and visualization techniques. Writeprints and ink blots can incorporate an array of features representing various information types. Both techniques also utilize complementary feature selection and visualization methods. Writeprints uses principal component analysis (PCA) with a sliding window algorithm to create lower dimensional plots that accentuate feature usage variation. Ink blots uses decision tree models (DTM) to select features that are superimposed onto text to show them as they occur. Writeprints is better suited for presenting a broad overview across large numbers of features. Ink blots is intended to show detailed examples of feature occurrences. Both techniques can be used for text characterization and discrimination (i.e., analysis and categorization). Specific details about the two methods are presented below.

Writeprints

The steps for the writeprints technique are

(1) Derive *n* primary eigenvectors (ones with largest eigenvalues) from the feature usage matrix where *n* is determined by the stopping rule or end user.
(2) Extract feature vectors for sliding window instance.
(3) Compute window instance coordinates by multiplying window feature vectors with *n* eigenvectors.
(4) Plot window instance points in *n* dimensional space.
(5) Repeat steps 2 through 4 for each window.

A sliding window of length *L* with a jump interval of *J* characters is run over the messages. The feature occurrence vector for each window is projected to an *n* dimensional space by taking the product of the window feature vector and the *n* primary eigenvectors, where *n* is determined using a stopping rule. Writeprints uses the Kaiser-Guttman stopping rule where all eigenvectors with an eigenvalue greater than one are selected (Jackson 1993), or a user-defined number of eigenvectors. Figure 6 illustrates the key steps in the writeprints process for a sample two-dimensional projection. The pro-



(a) Text Annotation View

Feature occurrences are highlighted in blue. The selected bag-of-words feature is highlighed in red (CounselEnron).

al CounselEnronWholesaleServices1400Smith Street -
EB3892HoustonTexas77008(713853-7459(713646-3490(fax I
thought of another item to add to the list:Cruising World 2002
wall calendar -- available on Amazon.com for 11.95Mark
TaylorVicePresidentand GeneralCounselEnronWholesale
Services1400Smith Street - EB3892Houston, Texas
77008(713853-7459(713646-3490(fax Hey Mike -- here are my de
Taylor 781322 Rutland St.Houston, TX 77008(713863-7190
(homemark.taylor@enron.comLet me know how I can help.Mark
TaylorVicePresidentand GeneralCounselEnronWholesale
Services1400Smith Street - EB3892Houston, Texas
77008(713853-7459(713646-3490(fax If you need to reach me this
will be at my parent s home in Cincinnati: (513541-8649.Mark
TaylorVicePresidentand GeneralCounselEnronWholesale
Services1400Smith Street - EB3892Houston, Texas
77008(713853-7459(713646-3490(fax In my view an assignment

(b) Ink Blots View

Colored circles (blots) superimposed onto feature occurrence locations in text. Blot size and color indicates feature importance and usage. Selected feature's blots are highlighted with black circles.

al Counsel Enron Wholesale Services1400 Smith Street -
EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax I
thought of another item to add to the list:Cruising World 2002
wall calendar -- available on Amazon.com for 11.95Mark
TaylorVice President and General Counsel Enron Wholesale
Services1400 Smith Street - EB3892Houston, Texas
77008(713853-7459(713646-3490 (fax Hey Mike -- here are my de
Taylor 781322 Rutland St.Houston, TX 77008(713863-7190
(homemark.taylor@enron.comLet me know how I can help.Mark
TaylorVice President and General Counsel Enron Wholesale
Services1400 Smith Street - EB3892Houston, Texas
77008(713853-7459(713646-3490 (fax If you need to reach me thi
will be at my parent s home in Cincinnati: (513541-8649.Mark
TaylorVice President and General Counsel Enron Wholesale
Services1400 Smith Street - EB3892Houston, Texas
77008(713853-7459(713646-3490 (fax In my view an assignment

Figure 4. Text Overlay Views in CyberGate

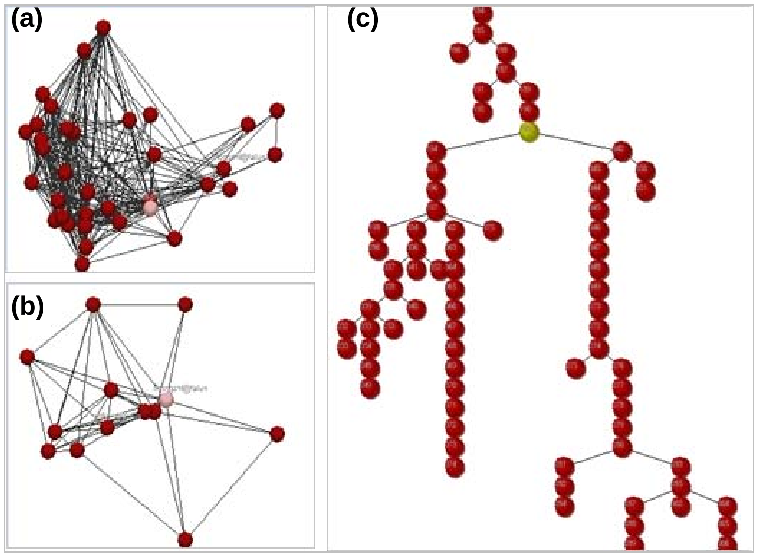


Figure 5. Interaction Views in CyberGate for Representing Interpersonal Information

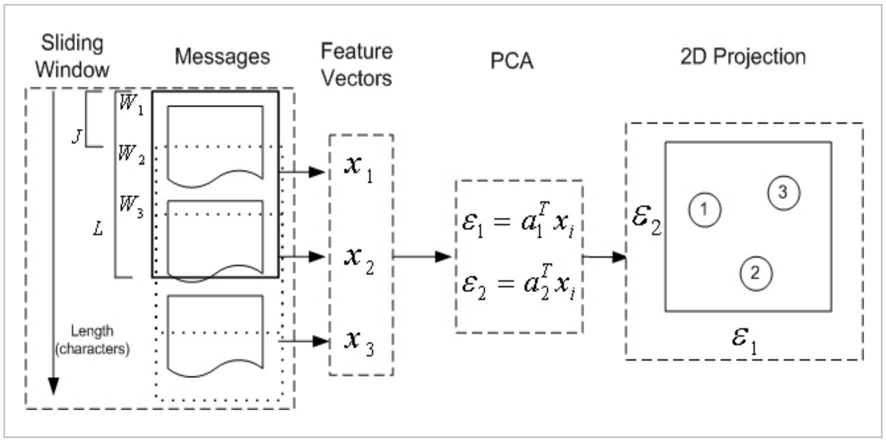


Figure 6. Writeprints Process Illustration on Two Dimensions

duct of the window feature vector and the first eigenvector is used to get the x-axis coordinate (ε_1) while the product of the feature vector and the second eigenvector produces the y-axis coordinate (ε_2). Writeprints is geared toward showing occurrence variation patterns. These patterns can be used for text categorization of stylistic information or analysis of information types serving the ideational and textual meta-functions.

Ink Blots

The steps for the ink blots technique are

(1) Separate input text into two classes (one for class of interest, one class containing all remaining texts).
(2) Extract feature vectors for messages.
(3) Input vectors into DTM as binary class problem
(4) For each feature in computed decision tree, determine blot size and color based on DTM weight and feature usage.
(5) Overlay feature blots onto their respective occurrences in text.
(6) Repeat steps 1 through 5 for each class.

The ink blots process is shown in Figure 7. The ink blots technique identifies the most important features for a given class using a binary class decision tree model (DTM). DTM efficiently considers feature interactions, unlike methods such as information gain and log likelihood (Forman 2003). A class can refer to an author, opinion, emotion, topic etc. The class of interest is input into the DTM along with a second class containing text from all other classes. The DTM determines the key features that differentiate the class of interest from other classes, weighted by their level of entropy reduction. For each selected feature, the weights determined by the DTM are used to determine the attributes' blot size (higher weight = larger blot size). Blot colors are determined based on feature usage. Red is assigned to features for which the class has the highest usage, while blue is for features never occurring in the class's text. All other features are assigned yellow. Let us assume we have 10 topics of interest for which we would like to identify the key blot features. For each topic, a DTM is generated comparing that topic against all others (to determine the topic's key features). These features are assigned weights and colors based on their DTM rankings and occurrence frequencies, respectively. The process is repeated for each topic. Finally, text overlay is performed by superimposing a topic's blot features at every location where the features occur. Once each class's key features have been extracted and assigned weights and colors, they can be used for categorization and analysis. For categorization, superimposing a class's blots onto an unclassified text can provide insight into whether the text belongs to that particular class. Correct class–text matches should result in patterns featuring high levels of red and yellow (features that occur frequently in this class's texts) and a minimal amount of blue (features rarely or never occurring in this class's texts). Ink blots can also be used to analyze how key class features occur and interact within a certain piece of text.

A CMC Text Analysis Example Using CyberGate: The Enron Case

We present an application example from the Enron e-mail corpus to illustrate how CyberGate can be used for data characterization of CMC text. The example utilizes writeprints

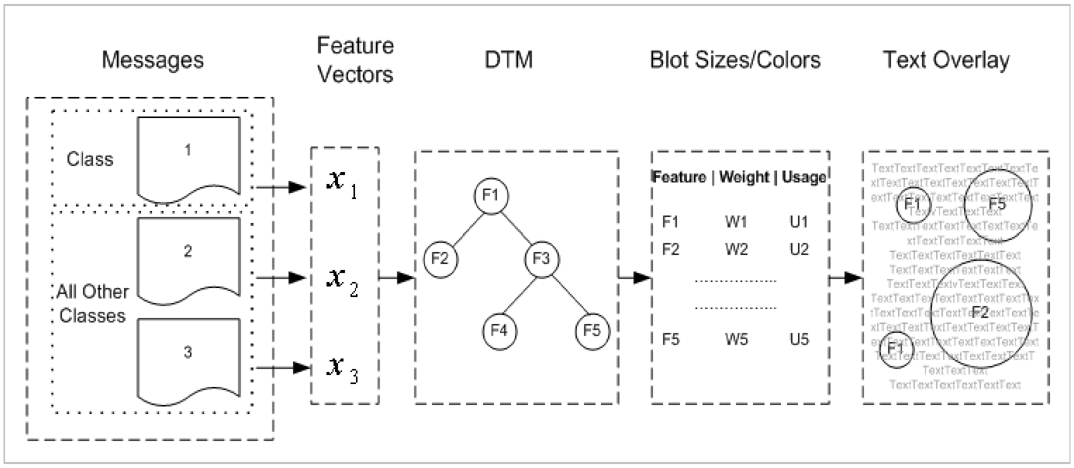


Figure 7. Ink Blots Process Illustration

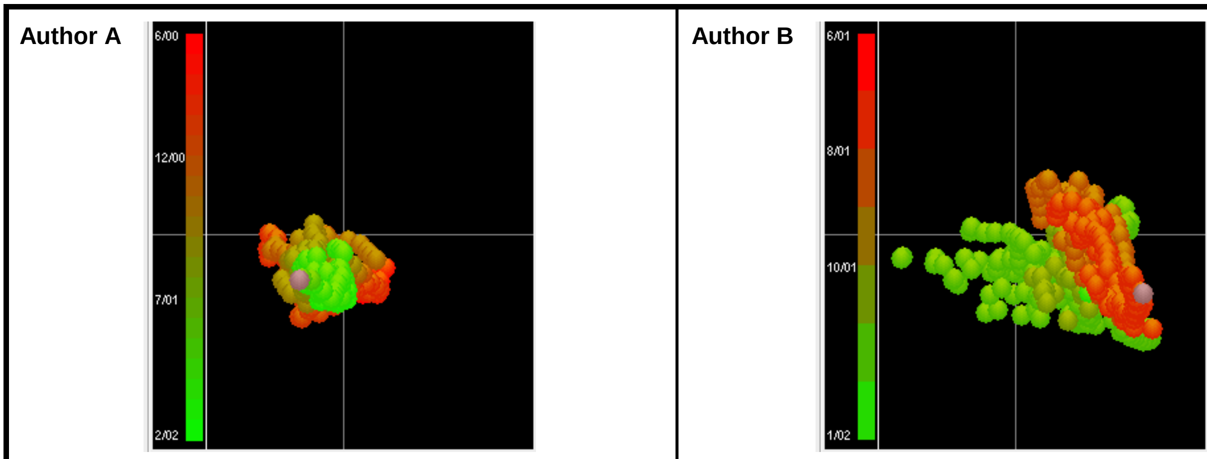
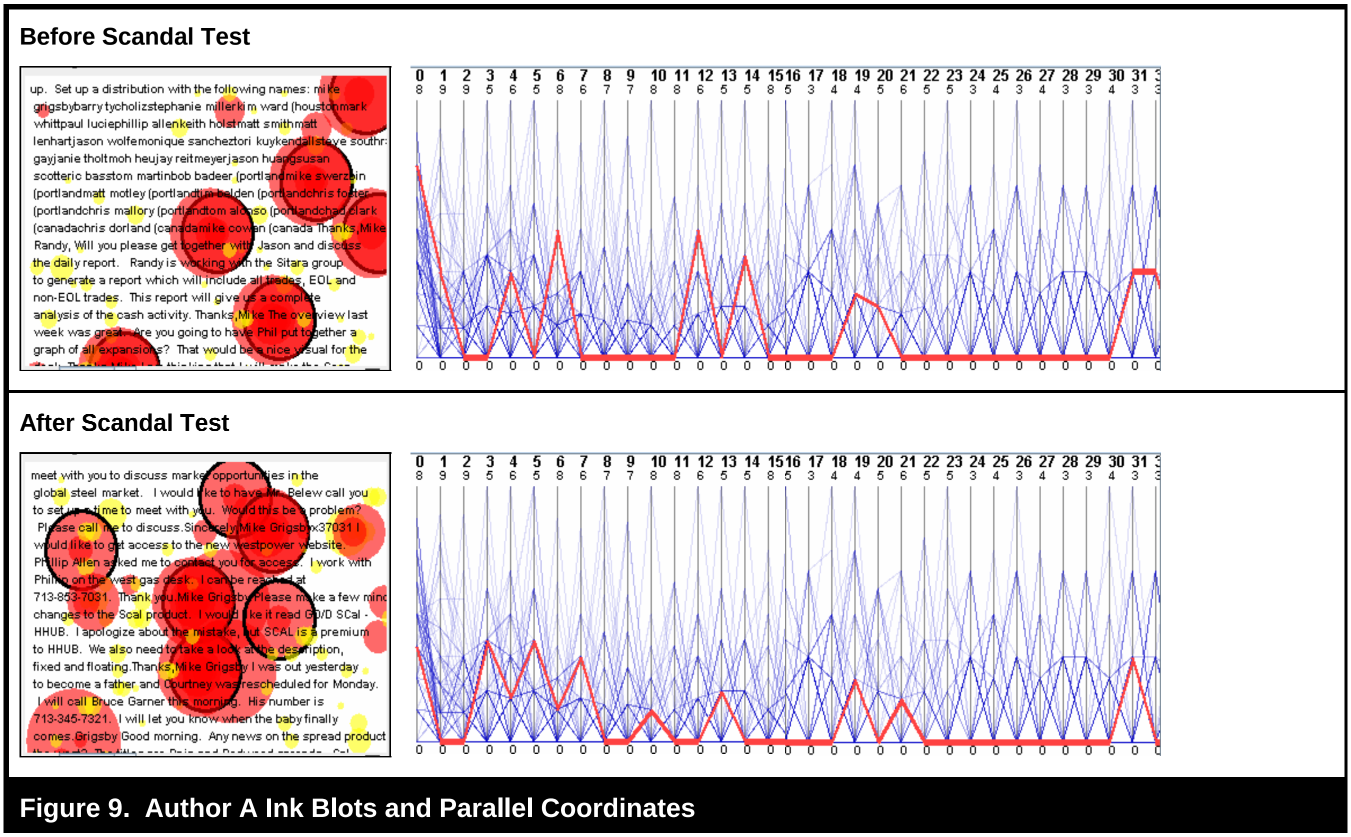


Figure 8. Writeprints for Two Enron Employees

and ink blots as well as additional CyberGate views such as parallel coordinates and MDS plots. The example relates to two Enron employees, neither of whom were directly involved in the scandal. Author A worked in the sales division while Author B was in the company's legal department. Figure 8 shows a temporal view of the two authors' writeprints taken across all features (lexical, syntactic, structural, semantic, n-grams, etc.). Each circle denotes a text window that is colored according to the point in time at which it occurred. The bright green points represent text windows from e-mails written after the scandal had broken out while the red points represent text windows from e-mails written before the scandal. Looking at the two patterns, we can see that Author B has greater overall feature variation as well as a distinct difference in the spatial location of points prior to the scandal (located more toward the right) as opposed to afterward (drifting toward the left). In contrast, Author A has no such difference, with his newer (green) text points placed directly on top of his older (red) ones. This suggests that Author B has had a profound change with respect to the text in his e-mails, while Author A exhibits no such changes. In order to further investigate this, we sampled points from the green and red regions for both authors and analyzed them using ink blots and parallel coordinates.

Figure 9 shows the ink blots and parallel coordinate views for sample points taken from Author A for text windows prior to and following the scandal. The ink blots show the author's



Before Scandal Test

up. Set up a distribution with the following names: mike grigsbybarry tycholizstephanie millerkim ward (houstonmark whittpaul luciephillip allenkeith holstmatt smithmatt lenhartjason wolfemonique sancheztori kuykendallsteve southr: gayjanie tholtmoh heujay reitmeyerjason huangsusan scotteric basstom martinbob badeer (portlandmike swerzbin (portlandmatt motley (portlandtim belden (portlandchris foster (portlandchris mallory (portlandtom alonso (portlandchad clark (canadachris dorland (canadamike cowan (canada Thanks,Mike Randy, Will you please get together with Jason and discuss the daily report. Randy is working with the Sitara group to generate a report which will include all trades, EOL and non-EOL trades. This report will give us a complete analysis of the cash activity. Thanks,Mike The overview last week was great. Are you going to have Phil put together a graph of all expansions? That would be a nice visual for the

After Scandal Test

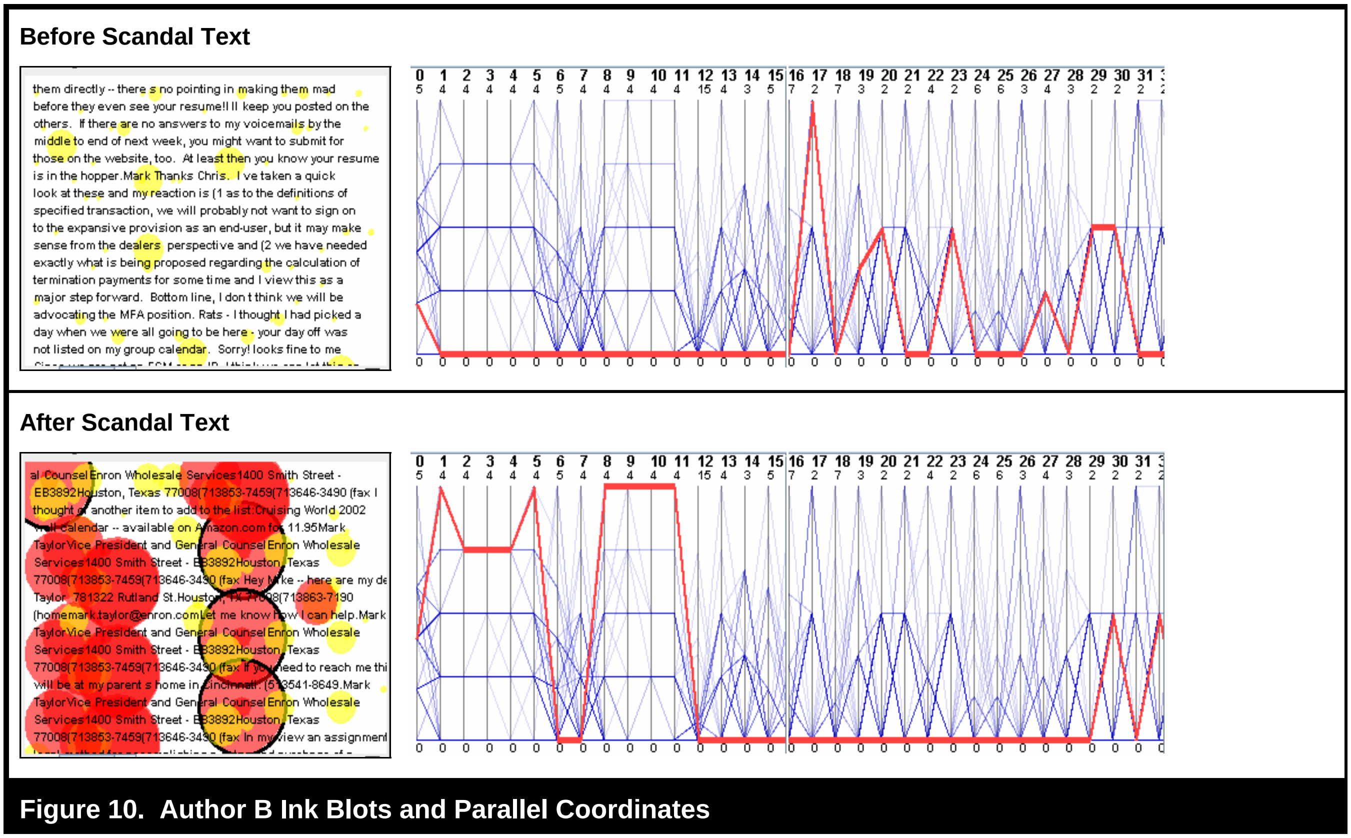
meet with you to discuss market opportunities in the global steel market. I would like to have Mr. Belew call you to set up a time to meet with you. Would this be a problem? Please call me to discuss.Sincerely,Mike Grigsbyx37031 I would like to get access to the new westpower website. Phillip Allen asked me to contact you for access. I work with Phillip on the west gas desk. I can be reached at 713-853-7031. Thank you.Mike Grigsby Please make a few minc changes to the Scal product. I would like it read GD/D SCal - HHUB. I apologize about the mistake, but SCAL is a premium to HHUB. We also need to take a look at the description, fixed and floating.Thanks,Mike Grigsby I was out yesterday to become a father and Courtney was rescheduled for Monday. I will call Bruce Garner this morning. His number is 713-345-7321. I will let you know when the baby finally comes.Grigsby Good morning. Any news on the spread product

Figure 9. Author A Ink Blots and Parallel Coordinates

key features superimposed onto the text. The usage of these features before as compared to after the scandal seems similar. The parallel coordinates show the author's 32 most important bag-of-words, including sales and negotiation related terms. These features signify the major topical content of the author's text. Again, the before and after coordinate patterns seem fairly similar, suggesting little text content deviation attributable to the scandal.

Figure 10 shows the ink blots and parallel coordinate views for sample text windows taken from Author B before and after the scandal. The ink blots view for the after-scandal text has considerably greater occurrence of key blot features. While the e-mails before the scandal focus on legal aspects of business deals with terms such as *counterparties* and *negotiations*, the discourse after the scandal mostly revolves around Author B providing advice and legal counsel to fellow employees. The post-scandal e-mails are more formal, containing greater usage of e-mail signatures with the author's job title and contact information. The bag-of-words features for these signature terms (e.g., title, address, phone number) correspond to the first 12 features shown in the parallel coordinates view. The terms relating to business legalities mentioned above correspond to the latter features (e.g., 15–30) in the parallel coordinates view. The parallel coordinates view exemplifies the stark contrast in Author B's e-mails as a result of the scandal. Clearly this dramatic alteration is attributable to a change in Author B's job functions.

Yates and Orlikowski (2002) stated that "the purpose of a genre is not an individual's private motive for communicating, but a purpose socially constructed and recognized by the relevant organizational community" (p. 15). Important characteristics of a genre form are structural and linguistic features including the level of formality and text formatting. For Author B, the post scandal e-mails signify a shift in genres. The author's job function changes from working on business contracts to providing advice and counsel to fellow employees. Figure 11a shows the key bag-of-words terms clustered based on occurrence similarity using MDS plots. The large cluster represents the business legality related terms (features 15–30 in parallel coordinates shown in Figure 10) while the two smaller clusters near the bottom contain the author's contact information and job title related terms, respectively. Author B shifts from usage of terms in the large cluster to the smaller ones. Similarly, the number of employees interacting with Author B increases considerably after the scandal as the author advises fellow workers (Figure 11b and 11c).



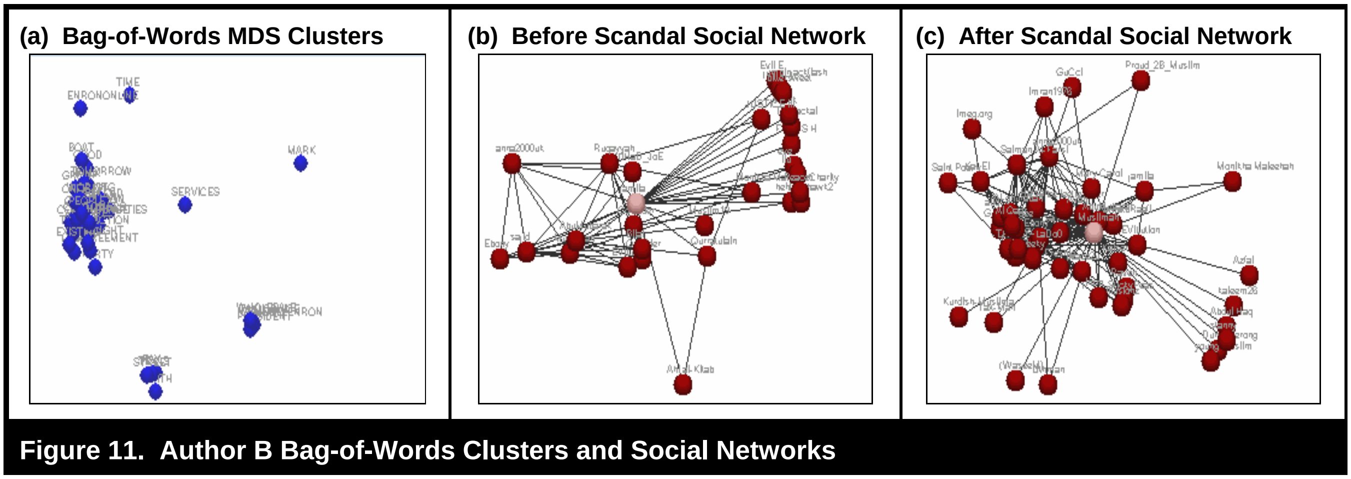
Before Scandal Text

them directly -- there s no pointing in making them mad before they even see your resume!! I ll keep you posted on the others. If there are no answers to my voicemails by the middle to end of next week, you might want to submit for those on the website, too. At least then you know your resume is in the hopper.Mark Thanks Chris. I ve taken a quick look at these and my reaction is (1 as to the definitions of specified transaction, we will probably not want to sign on to the expansive provision as an end-user, but it may make sense from the dealers perspective and (2 we have needed exactly what is being proposed regarding the calculation of termination payments for some time and I view this as a major step forward. Bottom line, I don t think we will be advocating the MFA position. Rats - I thought I had picked a day when we were all going to be here - your day off was not listed on my group calendar. Sorry! looks fine to me

After Scandal Text

al Counsel Enron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax I thought of another item to add to the list:Cruising World 2002 wall calendar -- available on Amazon.com for 11.95Mark TaylorVice President and General Counsel Enron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax Hey Mike -- here are my de Taylor 781322 Rutland St.Houston, TX 77008(713863-7190 (homemark.taylor@enron.comLet me know how I can help.Mark TaylorVice President and General Counsel Enron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax If you need to reach me thi will be at my parent s home in Cincinnati: (513541-8649.Mark TaylorVice President and General Counsel Enron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax In my view an assignment

Figure 10. Author B Ink Blots and Parallel Coordinates



(a) Bag-of-Words MDS Clusters

(b) Before Scandal Social Network

(c) After Scandal Social Network

Figure 11. Author B Bag-of-Words Clusters and Social Networks

This example illustrates how CyberGate and the proposed underlying framework's meta-design can be used for data characterization based representation of the meta-functions outlined in the meta-requirements. The example utilized a rich set of features: lexical, syntactic, structural, semantic, and various lexicons. A variety of ranking and projection based feature selection methods were incorporated (e.g., DTM, PCA, MDS). Multidimensional, text overlay, and interaction visualization techniques were also employed (writeprints, MDS plots, parallel coordinates, ink blots, social network graphs). The meta-design was used to represent information types (e.g., topics, genres, style, and interaction) associated with the ideational, textual, and interpersonal meta-functions.

Experimental Evaluation: Text Categorization Using CyberGate

Text categorization experiments were conducted using Cyber Gate. The experiments were intended to test the meta-design's effectiveness for discriminating information types used to represent the ideational, textual, and interpersonal meta-functions. Experiments evaluating the representation of the ideational and textual meta-function assessed CyberGate's features and selection/visualization techniques against comparison features and techniques. Evaluation of information types representing the interpersonal meta-function compared CyberGate's features against those used in other systems.

The representation of ideational meta-functions was evaluated by categorizing topics and opinions. For the textual meta-function, style and genres were tested. In these experiments, the writeprints or ink blots technique was compared against support vector machine (SVM). As previously alluded to, writeprints and ink blots both support text categorization. Writeprints are effective at capturing occurrence variation, which can be useful for categorizing style. Ink blots are geared toward occurrence frequency, which can be beneficial for genre, topic, and opinion categorization. SVM was incorporated since it has been a powerful machine learning algorithm for categorization of various information types including topics (Dumais et al. 1998), style (Zheng et al. 2006), and opinions (Pang et al. 2002). SVM was run using a linear kernel. In all experiments, a subset of the Cyber Gate's feature set was used based on the information type being evaluated. These feature subsets were composed of attributes commonly used for categorization of their respective information types. The same set of features was used for SVM and the CyberGate technique being evaluated in an experiment. In addition, a baseline configuration was included in all experiments, comprised of SVM run with bag-of-word (BOW) features (referred to as *baseline* from here on). BOWs have been used as the sole feature representation in virtually all text systems evaluated in prior research (Mladenic 1999; Tan 1999). BOWs are a fairly generalizable processing resource previously used for categorization of topics, style, opinions, and genres (Dumais et al. 1998; Pang et al. 2002). However, we do not believe BOW features are sufficient to effectively capture the various information types inherent in CMC text. Thus, while the SVM versus ink blots/writeprints comparison was intended to demonstrate the efficacy of these *techniques*, the comparison with the baseline was intended to illustrate the effectiveness of the CyberGate *features* over those included in standard text systems.

For representation of the interpersonal meta-function, the ability to accurately construct CMC interaction patterns was evaluated. As previously described, prior CMC systems rely solely on structural features derived from message headers (Sack 2000), while CyberGate uses structural and body text features for constructing interaction patterns in CMC. This is beneficial in CMC modes where headers are unavailable and interaction patterns less obvious. We evaluated the effectiveness of CyberGate's features against a baseline set comprised of only structural features. The experiments entailed evaluating the feature sets' ability to correctly assign user interactions (i.e., to which message or user a given message is responding).

Research Hypotheses

Table 7 presents hypotheses regarding CyberGate's ability to categorize information types representing the ideational, textual, and interpersonal meta-functions. For all experiments, pair-wise t-tests were used to evaluate the hypotheses. In order to enable easier summarization of the results, the p-values from the five ensuing experiments are included here.

Information Types Representing the Ideational Meta-Function

Experiment 1: Topic Categorization

We extracted e-mails pertaining to 10 topics from the Enron e-mail corpus. The messages were extracted and tagged by an independent coder who read each message before deciding upon a topic tag. Only messages that were tagged with a single topic were included. Example topics include energy, shares, and litigation. For each topic, 100 e-mail messages were used, resulting in a test bed of 1,000 e-mail messages. In order to gauge the effectiveness of the coding, a second coder tagged 100 messages from the test bed. The kappa statistic was computed between the two coders, with a value of 0.83 (which is considered reliable). Our feature set consisted of bag-of-words and noun phrases. Both feature representations have been effectively used for topic categorization (Chen et al. 2003; Dumais et al. 1998). A minimum frequency threshold of three was used to determine the number of bag-of-words and noun phrases to include (Joachims 1998).

Two experimental settings were run, one using 5 topics and the other using all 10 topics. The experiments featured ink blots in comparison with SVM and the baseline (SVM with only BOWs). All techniques were run using 10-fold cross validation. For ink blots, this meant that the DTM and occurrence analysis used to assign each topic class its blot sizes and colors was run on 90 percent of the data each fold, while the other 10 percent was used for evaluation. The class with the highest ratio of red to blue blot area was assigned the anonymous message.

Table 7. Hypotheses Testing Results for Text Categorization Experiments

Hypotheses	P-Values	
Representation of the Ideational Meta-Function	Setting 1	Setting 2
H1a: Techniques using CyberGate's features will outperform the baseline features for the categorization of *topics*.	< 0.001*	< 0.001*
H1b: CyberGate techniques will outperform SVM for the categorization of *topics*.	< 0.001+	< 0.001+
H2a: Techniques using CyberGate's features will outperform the baseline features for the categorization of *opinions*.	< 0.001*	< 0.001*
H2b: CyberGate techniques will outperform SVM for the categorization of *opinions*.	0.086	0.062
Representation of the Textual Meta-Function	Setting 1	Setting 2
H3a: Techniques using CyberGate's features will outperform the baseline features for the categorization of *style*.	< 0.001*	< 0.001*
H3b: CyberGate techniques will outperform SVM for the categorization of *style*.	< 0.001*	< 0.001*
H4a: Techniques using CyberGate's features will outperform the baseline features for the categorization of *genres*.	< 0.001*	< 0.001*
H4b: CyberGate techniques will outperform SVM for the categorization of *genres*.	0.127	0.103
Representation of the Interpersonal Meta-Function	Test Bed 1	Test Bed 2
H5: CyberGate's features will outperform the baseline features for categorization of *interaction* patterns.	< 0.001*	< 0.001*

*P-value significant at alpha = 0.01

Table 8. Topic Categorization Results (Accuracy)

Topic Setting	Techniques		
	SVM	Ink Blots	Baseline
5 Topics	**95.70**	92.25	88.75
10 topics	**93.25**	90.10	86.55

Table 8 shows the topic categorization results. Both techniques using the richer feature representation achieved over 90 percent accuracy, significantly outperforming the baseline (p-values < 0.001). However, SVM significantly outperformed the ink blot technique for the 5 and 10 topic experiment settings (p-values < 0.001). Error analysis on ink blots' misclassified messages revealed that the higher performance of SVM was likely attributable to its ability to better classify the small percentage of messages that were in the gray area between topics (e.g., messages primarily talking about energy, but also mentioning litigation). A second coder tagged these 60 misclassified messages, with a kappa statistic of only 0.65 with the original coding. The considerably lower inter-coder reliability of these messages as compared to the 0.83 overall kappa value supports our conclusion.

Experiment 2: Opinion Classification

The objective of the opinion classification experiment was to test the effectiveness of the CyberGate's features and techniques for capturing sentiment polarities. The test bed consisted of 2,000 digital camera reviews from www.epinions.com. The 2,000 reviews were composed of 1,000 positive (4–5 star) and 1,000 negative (1–2 star), with 500 reviews for each star level (i.e., 1, 2, 4, 5). Two problem scenarios were tested: (1) classifying 1 star versus 5 star reviews (extreme polarity) and (2) classifying 1+2 star versus 4+5 star reviews (milder polarity). The feature set encompassed a lexicon of 3,000 positive or negatively oriented adjectives (Turney and Littman 2003) and word n-grams (Pang et al. 2002). Once again SVM, ink blots, and the base-

Table 9. Opinion Classification Results

Sentiment Setting	Techniques		
	SVM	Ink Blots	Baseline
Extreme Polarity	**93.00**	92.20	83.00
Mild Polarity	**89.40**	86.80	77.10

line were run using 10-fold cross validation for each experiment setting (mild and extreme polarity).

The experimental results are presented in Table 9. SVM marginally outperformed ink blots. However the enhanced performance was not statistically significant (p-values on pair wise t-tests > 0.05). SVM and ink blots both significantly outperformed the baseline (p-values on pair wise t-tests < 0.001) by a margin of over 10 percent, highlighting the importance of a representational richness for opinion categorization. The overall accuracies for both SVM and ink blots were consistent with previous work which has been in the 85 to 90 percent range (e.g., Pang et al. 2002). Once again the improved performance of SVM was attributable to its ability to better detect messages containing sentiments with less polarity. In many cases, it was more difficult for the ink blots technique's to detect the overall orientation of these messages. This is evidenced by the fact that the ink blots' accuracy dropped more when switching from extreme to mild polarity as compared to SVM.

Information Types Representing the Textual Meta-Function

Experiment 3: Style Classification

We conducted authorship classification experiments to test the effectiveness of our features and techniques for capturing style. The objective of the experiments was to correctly categorize individuals based on their writing style. Our test bed consisted of authors from the Enron e-mail corpus. The experiments involved an entity resolution classification task in which half of the messages were used for training (treated as the known entity) and half for testing (considered an anonymous entity). The objective in such a task is to match anonymous entities to the correct known entities based on stylistic tendencies. The experiments were run using 25 and 50 authors. Thus, in the 25 author setting, we had 25 known entities and 25 anonymous entities, with each set constructed using one half of the messages. The feature set consisted of lexical, syntactic, structural, and semantic features. Lexical features included word and character level measures (e.g., words per sentence, characters per word, etc.). The syntactic features used were function words, punctuation marks, and POS and word n-grams. The structural features encompassed the use of greetings, quoted content, hyperlinks, etc. Semantic features used included noun phrases and named entities. These feature categories are described in greater detail in Table 5. The effectiveness of these features for capturing style has previously been demonstrated (Zheng et al. 2006). The writeprints technique was used in comparison with SVM. The ability of writeprints to capture feature variation patterns is conducive to stylistic classification. For writeprints, the sliding window was run over each entity creating an n-dimensional pattern. The anonymous patterns were each compared against the known patterns, with the anonymous entity being assigned to the known entity with the most similar pattern. Similarity was determined based on the average n-dimensional Euclidean distance between the two patterns' points. For SVM, 100 text tiles were created for each known and anonymous entity. The feature vectors for these 100 tiles were used for the training (known entity) and testing data (anonymous entity). The anonymous entities were classified as the known entity assigned the highest number of tiles by SVM during the testing phase. The experimental results are shown in Table 10.

Writeprints outperformed SVM by 8 to 10 percent for both experimental settings. The enhanced performance was statistically significant for 25 and 50 authors. Furthermore, the accuracy of writeprints is an improvement over prior research (Zheng et al. 2006). Writeprints and SVM both also significantly outperformed the baseline by 20 to 30 percent. This is attributable to CyberGate's use of features that can effectively capture stylistic information usage and variation.

Experiment 4: Genre Classification

For genre classification, a test bed of 3,000 forum postings from the Sun Technology Forum (forum.java.sun.com) was used. Categorization of genres in such forums can be useful for studying knowledge transfer patterns in electronic networks of practice (Wasko and Faraj 2005). The genres cate-

Table 10. Style Classification Results

Author Setting	Techniques		
	SVM	Writeprints	Baseline
25 Authors	84.00	**92.00**	62.00
50 Authors	80.00	**90.00**	51.00

Table 11. Genre Classification Results

Genre Setting	Techniques		
	SVM	Ink Blots	Baseline
Questions vs. Non-questions	98.10	**98.55**	90.10
All Three Genres	96.40	**96.50**	86.00

gorized included questions, informative messages, and general messages (uninformative comments), with 1,000 messages used for each genre. Two experiment settings were run: (1) questions (1,000 messages) versus non-questions (500 informative, 500 comments) and (2) all three genres (1,000 messages each). The feature set consisted of lexical, syntactic, structural, semantic, and n-gram features. The ink blots technique was compared against SVM and the BOW baseline. Each technique was run using 10-fold cross validation (same settings as the topic and opinion categorization experiments).

The experimental results are presented in Table 11. Both SVM and ink blots significantly outperformed the baseline (p-values < 0.001). Ink blots outperformed SVM; however, the margin was not statistically significant (p-values > 0.05). The overall accuracies for both SVM and ink blots were consistent with prior results dealing with 2 or 3 genres (Santini 2004), validating the efficacy of the underlying features and techniques for genre categorization.

Information Types Representing the Interpersonal Meta-Function

Experiment 5: Interaction Classification

For interaction classification, we used two test beds: four conversation threads taken from the Sun Java Technology forum (1,00 messages posted by 120 users) and three threads taken from the LNSG social discussion forum (400 messages posted by 100 users). Two independent coders tagged the test beds for message interactions. The coders carefully read each message to determine which prior posting (if any) it referenced or responded to. The inter-coder reliability had kappa statistics of 0.88 and 0.81 for the two test beds, respectively.

CyberGate's feature set consisted of structural features (taken from the message headers) as well as function words, bag-of-words, noun phrases, and named entities derived from body text. These features are intended to represent various interaction cues, including direct address (reference to user names) and lexical relation (reference to keywords from prior postings). The baseline feature set consisted of only structural features, as used in prior systems (Donath et al. 1999; Smith and Fiore 2001). CyberGate assigns direct address relations when a user name is referenced (in which case a relation is assigned between the message author and the referenced user). Lexical relations are assigned using a modified vector space model (Fu et al. 2008). The structural interaction cues are matched by comparing message titles and quoted content with the title and content of prior postings. Consistent with prior research, the F-measure was used to evaluate the effectiveness of each feature set (Fu et al. 2008).

The experimental results are presented in Table 12. CyberGate's extended feature set significantly outperformed the baseline (p-values < 0.001). The performance difference was more pronounced on the LNSG forum. Users in this forum make less use of structural features when interacting with one another, instead preferring to rely on text-based interaction cues. The results illustrate the importance of using richer features for representing CMC interactions.

Table 12. Interaction Classification Results

Test Bed	Features	
	CyberGate	**Baseline**
Sun Java Forum	**86.00**	77.40
LNSG Forum	**77.11**	55.55

Results Discussion

CyberGate's feature set was better at representing information types associated with the three meta-functions as compared to baseline feature sets commonly used in prior systems. The extended feature set effectively represented topic, opinion, style, genre, and interaction information. It significantly outperformed the BOW and structural feature baselines. The CyberGate techniques also performed well, with accuracies generally over 90 percent. SVM appeared to perform better on information types supporting the ideational meta-function. Writeprints and ink blots outperformed SVM in experiments on information types representing the textual meta-function. For instance, SVM had significantly higher accuracy for topic classification, while writeprints and ink blots performed better on style and genre classification.

The objective of the experiments was to test the meta-design's effectiveness for discriminating information types associated with the ideational, textual, and interpersonal meta-functions. CyberGate's extended feature set enhanced representation of the three meta-functions. The writeprints and ink blot techniques were also successful in discriminating information types related to the ideational and textual meta-functions, although more so for the textual meta-function. The results suggest that an extended feature set (using language and processing resources) and complementary feature selection and visualization techniques can enhance data discrimination based representation of information types reflective of the three meta-functions.

In the previous section, an application example was used to illustrate the meta-design's ability to characterize information types associated with the meta-functions. This section presented text categorization experiments to assess the data discrimination capabilities of features, feature selection, and visualization techniques based on the meta-design. Collectively, the results lend validity to the meta-design satisfying meta-requirements. Systems using the proposed design framework may foster better analysis of CMC text by representing the ideational, textual, and interpersonal meta-functions (Sack 2000). CMC systems supporting only a subset of the language meta-functions are likely to lose the deeper understanding that arises from the synergy created by representing the three meta-functions in unison.

Conclusions

Our major research contributions are two-fold. First, using Walls et al.'s (1992) model, we developed a design framework for systems supporting CMC text analysis. The framework advocates the development of systems that support all three language meta-functions described by systemic functional linguistic theory. The framework also provides guidelines for the choice of appropriate features, feature selection, and visualization techniques necessary to effectively represent the ideational, textual, and interpersonal meta-functions. Second, we developed the CyberGate system based on our design framework. CyberGate includes an array of language and processing resources capable of representing various information types. It also incorporates a bevy of ranking and projection based feature selection methods and various complementary visualization formats, including the writeprints and ink blots techniques. Using CyberGate's features and techniques, text categorization experiments and an application example were used to validate the meta-design elements of the proposed design framework.

Our design framework and the resulting CyberGate system are not without their shortcomings. There are likely to be additional information types, and feature selection and visualization techniques that could have been considered but were omitted. Nevertheless, we believe that the proposed framework has important implications for practitioners, including CMC system users and developers. Our intention and hope is that future research will improve upon our design framework, resulting in CMC systems with enhanced text analysis capabilities.

Acknowledgments

This research has been supported in part by the following National Science Foundation grant #0646942: "Online Stylometric Authorship Identification: An Exploratory Study," September 2006–February 2008.

References

Abbasi, A., and Chen, H. 2005. "Identification and Comparison of Extremist-Group Web Forum Messages Uing Authorship Analysis," *IEEE Intelligent Systems* (20:5), pp. 67-75.

Abbasi, A., and Chen, H. 2006. "Visualizing Authorship for Identification," in *Proceedings of the 4th IEEE Conference on Intelligence and Security Informatics*, S. Mehrota, D. D. Zeng, H. Chen, B. M. Thurasingham, and F-Y Wang (eds.), San Diego, CA, May 23-24, pp. 60-71.

Allan, J., Carbonell, J, Doddington, G., Yamron, J., and Yang, Y 1998. "Topic Detection and Tracking Pilot Study: Final Report," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, pp. 194-218.

Allan, J., Leuski, A., Swan R. C., and Byrd, D. 2001. "Evaluating Combinations of Ranked Lists and Visualizations of Inter-Document Similarity," *Information Processing and Management* (37:3), pp. 435-458.

Andrienko, N., and Andrienko, G. 2003. "Informed Spatial Decisions through Coordinated Views," *Information Visualization* (2:4), pp. 270-285.

Argamon, S., Whitelaw, C., Chase, P., Raj Hota, S., Garg, N., and Levitan, S. 2007. "Stylistic Text Classification Using Functional Lexical Features," *Journal of the American Society for Information Science and Technology* (58:6), pp. 802-822.

Barua, A., Konana, P., and Whinston, A. 2004. "An Empirical Investigation of Net-Enabled Business Value," *MIS Quarterly* (28:4), pp. 585-620.

Butler, B. S. 2001. "Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures," *Information Systems Research* (12:4), pp. 346-362.

Chen, H. 2001. *Knowledge Management Systems: A Text Mining Perspective*, Tucson, AZ: Knowledge Computing Corporation.

Chen, H., Lally, A. M., Zhu, B., and Chau, M. 2003. "HelpfulMed: Intelligent Searching for Medical Information over the Internet," *Journal of the American Society for Information Science and Technology* (54:7), pp. 683-694.

Chia, R. 2000. "Discourse Analysis as Organizational Analysis," *Organization* (7:3), pp. 513-518.

Cothrel, J. P. 2000. "Measuring the Success of an Online Community," *Strategy and Leadership* (20:2), pp. 17-21.

Cunningham, H. 2002. "GATE: A General Architecture for Text Engineering," *Computers and the Humanities* (36), pp. 223-254.

Daft, R. L., and Lengel, R. H. 1986. "Organizational Information Requirements, Media Richness and Structural Design," *Management Science* (32:5), pp. 554-571.

Donath, J. 2002. "Semantic Approach to Visualizing Online Conversations," *Communications of the ACM* (45:4), pp. 45-49.

Donath, J., Karahalio, K., and Viegas, F. 1999. "Visualizing Conversation," *Journal of Computer-Mediated Communication* (4:4) (http://jcmc.indiana.edu/vol4/issue4/donath.html).

Dorre, J., Gerstl, P., and Seiffert, R. 1999. "Text Mining: Finding Nuggets in Mountains of Textual Data," in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Mining*, U. Fayyad, S. Chaudhuri, and D. Madigan (eds.), San Diego, CA, August 15-18, pp. 398-401.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. 1998. "Inductive Learning Algorithms and Representations for Text Categorization," in *Proceedings of the 7th of ACM Conference on Information and Knowledge Management*, G. Gardarin, J. C. French, N. Pissinou, K. Makki, and L. Bouganim (eds.), Bethesda, MD, November 3-7, pp. 148-155.

Erickson, T., and Kellogg, W. A. 2000. "Social Translucence: An Approach to Designing Systems that Support Social Processes," *ACM Transactions on Computer-Human Interaction* (7:1), pp. 59-83.

Fairclough, N. 2003. *Analyzing Discourse: Textual Analysis for Social Research*, New York: Routledge.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press.

Fiore, A. T., and Smith, M. A. 2000. "TreeMap Visualizations of News Groups," poster presentation at the IEEE Symposium on Information Visualization, Boston, MA, October 9-10.

Fjermestad, J., and Hiltz, S. R. 1999. "An Assessment of Group Support Systems Experimental Research: Methodology and Results," *Journal of Management Information Systems* (15:3), pp. 7-49.

Forman, G. 2003. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," *The Journal of Machine Learning Research* (3), pp. 1289-1305.

Fu, T., Abbasi, A., and Chen, H. 2008. "A Hybrid Approach to Web Forum Interactional Coherence Analysis," *Journal of the American Society for Information Science and Technology* (59:8), pp. 1195-1209.

Guyon, I., and Elisseef, A. 2003. "An Introduction to Variable and Feature Selection," *The Journal of Machine Learning Research* (3), pp. 1157-1182.

Halliday, M. A. K. 2004. *An Introduction to Functional Grammar* (3rd ed., revised by C. Matthiessen), London: Hodder Arnold.

Han, J., and Kamber, M. 2001. *Data Mining: Concepts and Techniques*, San Francisco: Academic Press.

Hara, N., Bonk, C. J., and Angeli, C. 2000. "Content Analysis of Online Discussion in an Applied Educational Psychology Course," *Instructional Science* (28), pp. 115-152.

Havre, S., Hetzler, E., Whitney, P. and Nowell, L. 2002. "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *IEEE Transactions on Visualization and Computer Graphics* (8:1), pp. 9-20.

Hearst, M. A. 1999. "Untangling Text Data Mining," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, R. Dale and K. Church (eds.), College Park, MD, June 20-26, pp. 3-10.

Henri, F. 1992. "Computer Conferencing and Content Analysis," in *Collaborative Learning through Computer Conferencing: The Najaden Papers*, A. R. Kaye (ed), Berkeley, CA: Springer, pp. 115-136.

Herring, S. C. 2002. "Computer-Mediated Communication on the Internet," *Annual Review of Information Science and Technology* (36:1), pp. 109-168.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.

Huang, S., Ward, M. O., and Rundensteiner, E. A. 2005. "Exploration of Dimensionality Reduction For Text Visualization, in *Proceedings of The Third International Conference on Coordinated and Multiple Views in Exploratory Visualization*, J. C. Roberts (ed.), London, July 5, pp. 63-74.

Jackson, D. 1993. "Stopping Rules in Principal Component Analysis: A Comparison of Heuristical and Statistical Approaches," *Ecology* (74:8), pp. 2204-2214.

Joachims, T. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proceedings of the 10th European Conference on Machine Learning*, C. Nedellec and C. Rouveirol (eds.), Chemnitz, Germany, April 21-24, pp. 137-142.

Keim, D. A. 2002. "Information Visualization and Visual Data Mining," *IEEE Transactions on Visualization and Computer Graphics* (7:1), pp. 100-107.

Koppel, M., and Schler, J. 2003. "Exploiting Stylistic Idiosyncrasies for Authorship Attribution," in *Proceedings of IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis,* A. G. Cohn (ed.), Acapulco, Mexico, pp. 69-72..

Lee, A. S. 1994. "Electronic Mail as a Medium of Rich Communication: An Empirical Investigation Using Hermeneutic Interpretation," *MIS Quarterly* (18:2), pp. 143-157.

Losiewicz, P., Oard, D., and Kostoff, R. N. 2000. "Textual Data Mining to Support Science and Technology Management," *Journal of Intelligent Information Systems* (15), pp. 99-119.

March, S. T., and Smith, G. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 251-266.

Markus, M. L., Majchrzak, A., and Gasser, L. 2002. "A Design Theory for Systems that Support Emergent Knowledge Processes," *MIS Quarterly* (26:3), pp. 179-212.

Miller, N. E., Wong, P.C., Brewster, M., and Foote, H. 1998. "Topic Islands: A Wavelet-based Text Visualization System," in *Proceedings of the 9th IEEE Conference on Visualization*, T-M. Rhyne and R. Moorhead (eds.), Research Triangle Park, NC, October 18-23, pp. 189-196.

Mladenic, D. 1999. "Text-Learning and Related Intelligent Agents: A Survey," *IEEE Intelligent Systems* (14:4), pp. 44-54.

Montoya-Weiss, M., Massey, A. P., and Song, M. 2001. "Getting it Together: Temporal Coordination and Conflict Management in Global Virtual Teams," *Academy of Management Journal* (44:6), pp. 1251-1262.

Nasukawa, T., and Nagano, T. 2001. "Text Analysis and Knowledge Mining System, *IBM Systems Journal* (40:4), pp. 967-984.

Nigam, K., and Hurst, M. 2004. "Towards a Robust Metric of Opinion," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Menlo Park, CA: The AAAI Press, pp. 598-603.

Paccagnella, L. 1997. "Getting the Seats of Your Pants Dirty: Strategies for Ethnographic Research on Virtual Communities, *Journal of Computer-Mediated Communication* (3:1) (http://jcmc.indiana.edu/vol3/issue1/paccagnella.html).

Pang, B., Lee, L., and Vaithyanathain, S. 2002. "Thumbs Up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, J. Hajic and Y. Matsumoto (eds), Philadelphia, PA, July 6-7, pp. 79-86.

Panteli, N. 2002. "Richness, Power Cues and Email Text," *Information and Management* (40:2), pp. 75-86.

Picard, R. W. 1997. *Affective Computing*, Cambridge, MA: MIT Press.

Rudman, J. 1997. "The State of Authorship Attribution Studies: Some Problems and Solutions," *Computers and the Humanities* (31), pp. 351-365.

Sack, W. 2000. "Conversation Map: An Interface for Very Large-Scale Conversations," *Journal of Management Information Systems* (17:3), pp. 73-92.

Santini, M. 2004. "A Shallow Approach to Syntactic Feature Extraction for Genre Classification," in *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, Birmingham, UK, January 6-7.

Seo, J., and Shneiderman, B. 2005. "Rank-by-Feature Framework for the Interactive Exploration of Multidimensional Data," *Information Visualization* (4), pp. 99-113.

Simon, H. A. 1996. *The Sciences of the Artificial* (3rd ed.), Cambridge, MA: MIT Press.

Smith, M. A. 2002. "Tools for Navigating Large Social Cyberspaces," *Communications of ACM* (45:4), pp. 51-55.

Smith, M. A., and Fiore, A. T. 2001. "Visualization Components for Persistent Conversations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* J. Jacko and A. Spears, Seattle, WA, March 31-April 5, pp. 136-143.

Subasic, P., and Huettner, A. 2001. "Affect Analysis of Text Using Fuzzy Semantic Typing," *IEEE Transactions on Fuzzy Systems* (9:4), pp. 483-496.

Tan, A. 1999. "Text Mining: The State of the Art and the Challenges," in *Proceedings of the PAKDD Workshop on Knowledge Discovery and Data Mining*, N. Zhong and L. Zhou, Beijing, China, pp. 65-70.

Turney, P. D., and Littman, M. L. 2003. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Transactions on Information Systems* (21:4), pp. 315-346.

Viegas, F. B., and Smith, M. "Newsgroup Crowds and Author Lines: Visualizing the Activity of Individuals in Conversational Cyberspaces," in *Proceedings of the 37th Hawaii International Conference on System Sciences*, Los Alamitos, CA: IEEE Computer Society Press, pp. 10-18.

Viegas, F., Boyd, D., Nguyen, D. H., Potter, J., and Donath, J. 2004. "Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments," in *Proceedings of the 37th Hawaii International Conference on System Sciences*, Los Alamitos, CA: IEEE Computer Society Press.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an Information System Design Theory for Vigilant EIS," *Information Systems Research* (3:1), pp. 36-59.

Wasko, M. M., and Faraj, S. 2005. "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice, *MIS Quarterly* (29:1), pp. 35-57.

Wellman, B. 2001. "Computers Networks as Social Networks," *Science* (293), pp. 2031-2034.

Wenger, E. C., and Snyder, W. M. 2000. "Communities of Practice: The Organizational Frontier," *Harvard Business Review* (78:1), pp. 139-145.

Wilson, S. M., and Peterson, L. C. 2002. "The Anthropology of Online Communities," *Annual Review of Anthropology* (31), pp. 449-467.

Wise, J. A. 1999. "The Ecological Approach to Text Visualization," *Journal of the American Society for Information Science and Technology* (50:13), pp. 1224-1233.

Xiong, R., and Donath, J. 1999. "PeopleGarden: Creating Data Portraits for Users," in *Proceedings of 12th Annual ACM Symposium on User Interface Software and Technology*, B. Vander Zanden and J. Marks (eds.), Asheville, NC, November 7-10, pp. 37-44.

Yates, J., and Orlikowski, W. J. 2002. "Genre Systems: Structuring Interaction through Communicative Norms," *The Journal of Business Communication* (39:1), pp. 13-35.

Zheng, R., Li, J., Chen, H., and Huang, Z. 2006. "Framework for Authorship Analysis of Online Messages: Writing-Style Features and Techniques," *Journal of the American Society for Information Science and Technology* (57:3), pp.378-393.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., and Twichell, D. 2004. "Automating Lingusitics-Based Cues for Deception Detection in Text-Based Asynchronous Computer-Mediated Communication," *Group Decision and Negotiation* (13:1), pp. 81-106.

Zhu, B., and Chen H. 2002. "Visualizing the Archive of a Computer Mediated Communication Process," in *Proceedings of the 2nd ACM/IEEE Joint Conference on Digital Libraries*, Portland, OR, July 14-18, p. 385.

About the Author

Ahmed Abbasi is an assistant professor of Management Information Systems in the Sheldon B. Lubar School of Business at the University of Wisconsin–Milwaukee. He received a B.S. and MBA in Information Technology from Virginia Tech and a Ph.D. in Management Information Systems from the University of Arizona. His research interests include the application of text mining and information visualization techniques to computer-mediated communication and electronic markets. His research has appeared in various journals, including *Journal of Management Information Systems*, *ACM Transactions on Information Systems*, *IEEE Transactions on Knowledge and Data Engineering*, and *IEEE Intelligent Systems*. Ahmed is a member of AIS, IEEE, and DSI.

Hsinchun Chen is the McClelland Professor of Management Information Systems at the University of Arizona. He received a B.S. degree from the National Chiao-Tung University in Taiwan, an MBA from SUNY Buffalo, and a Ph.D. in Information Systems from New York University. Hsinchun is a Fellow of IEEE and AAAS. He received the IEEE Computer Society 2006 Technical Achievement Award. He is author/editor of 13 books, 17 book chapters, and more than 140 journal articles covering digital library, intelligence analysis, biomedical informatics, data/text/web mining, knowledge management, and web computing. He serves on 10 editorial boards and has served as a scientific counselor/advisor of the National Library of Medicine. Hsinchun's work has been funded by various organizations, including the National Science Foundation, Department of Justice, National Library of Medicine, Department of Defense, and Department of Homeland Security.